HEALTH SURVEY RESEARCH METHODS

Conference Proceedings

nchs

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control and Prevention
National Center for Health Statistics

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Library of Congress Cataloging-in-Publication Data

Conference on Health Survey Research Methods (6th : 1995 : Breckenridge, Colo,)
Health survey research methods conference proceedings/ edited by Richard Warnecke.
p. cm. — (DHHS publication ; no. (PHS) 96-1013)
"April 1996."
Includes bibliographical references.
1. Health surveys-Congresses. I Warnecke, Richard B. II. Title, III. Series.
RA408.5.C58 1995
614.4'2--dc20 **96-13418**
CIP

HEALTH SURVEY RESEARCH METHODS

Conference Proceedings

Edited by
Richard Warnecke, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
April 1996

DHHS Publication No. (PHS) 96-1013

National Center for Health Statistics

Jack R. Anderson, *Acting Director*

Jennifer H. Madans, Ph.D., *Acting Deputy Director*

Jacob J. Feldman, Ph.D., *Associate Director for Analysis, Epidemiology, and Health Promotion*

Gail F. Fisher, Ph.D., *Associate Director for Planning and Extramural Programs*

Jack R. Anderson, *Acting Associate Director for International Statistics*

Stephen E. Nieberding, *Associate Director for Management*

Charles J. Rothwell, *Associate Director for Data Processing and Services*

Monroe G. Sirken, Ph.D., *Associate Director for Research and Methodology*

Foreword

These are the proceedings of the Sixth Conference on Health Survey Research Methods, held in Breckenridge, Colorado, June 24–26, 1995. As in the past, this conference moved ahead on the continuum of methodological improvements, adding to what is known and addressing new areas that had been featured briefly at previous conferences. The long-range objectives of this conference, as of its predecessors, were to improve the quality of health survey data and to enhance their value and use by policy makers responsible for shaping health practice, policy, and programs. The immediate aims of this, the Sixth Conference on Health Survey Research Methods, followed those of predecessor conferences and were to

1. invite formal presentation on the state of the art in defined areas of survey methodology as it relates to the quality of data on the nation's health;
2. promote discussion of these presentations using a format combining formal discussion of each presentation with open discussion and comments from all conference participants;
3. prepare a written summary of the discussion compiled from the formal papers, invited discussion, and open discussion in the form of conference proceedings; and
4. publish and disseminate the proceedings as a timely statement about the current understanding of the sources of nonsampling survey error, new knowledge of ways to reduce this error, and required research relevant to health surveys.

In all, 65 persons attended this conference. Twenty-nine papers were selected from 150 submissions. The selected papers were divided into five sessions: "Measuring Medical Care and Health Status" (six papers), "Research on Survey Questions" (six papers), "Sampling and Cooperation" (six papers), "Special Populations and Sensitive Issues" (six papers), and "Integrating Surveys and Other Data" (five papers). A chairperson and rapporteur were assigned to each session, and there were two formal discussants persession.

Background and History of the Health Survey Research Methods Conferences

In 1975, led by Leo Reeder, a group of over 50 methodologists and substantive researchers with common interests in health survey research methods convened in conference at Airlie House in Airlie, Virginia. The agenda was a discussion of how best to provide a venue for the discussion of the results of methodological research in health surveys from which they could be communicated to the large body of researchers engaged in the broad areas of health services research and epidemiology. The specific goals of that first meeting were to

1. identify critical methodological issues or problem areas for health survey research and the state of the art or knowledge with respect to these problems;
2. define the types of research problems that needed priority funding;
3. identify policy issues that could be addressed by survey data; and
4. communicate the results, recommendations, and implications to (a) the broader community of health researchers who use survey methods, (b) relevant government agencies and policy makers, and (c) other potential users of survey data.

The conclusions of this initial conference reemphasized the need for ongoing discussion about health survey research methods—in particular, what defined the state of the art and what did not work. To stress the need for continuity over time, Leo Reeder concluded the introduction by stating that the first proceedings was "tentatively planned as Volume 1," in the hope that "conferences and reports such as this [would] occur on a biennial or triennial basis." These first proceedings stressed the importance of these conferences for facilitating communication among researchers who use, develop, or evaluate survey methods but who are dispersed both geographically and in their work settings. They pointed to the fact that there was no other specific venue for convening those with interests in health survey research for discussion of common interests and problems encountered in their work in the specific area of application, health survey research methods. Finally, they stressed the need for working conferences that would allow for discussion of work in progress, negative results, and other topics that were not ordinarily addressed in the usual journals or meetings.

The second and third conferences were held at 2-year intervals in 1977 and 1979. At each conference, the agenda and presentation format became more specific. By the 1979

conference, papers describing specific projects with responses by discussants had supplanted the general thematic discussion that comprised the programs at the first two conferences. However, the key aspects of these conferences did not change: They remained by invitation only, the papers were meant to guide general discussion, and the important role of the rapporteur at each conference in capturing and recording the general discussion that followed the formal presentations was retained. It is a unique aspect of these proceedings that the general discussion is always written up and included as part of the publication. Moreover, the summary of the discussion for each session is generally considered to be as important as the formal presentations.

The fourth conference was held in 1982, 3 years after the third conference. Seven years elapsed between the fourth conference and the fifth conference, which took place in 1989. The present conference occurred 6 years after the fifth. The longer time intervals between the last three conferences as compared with the first three reflect the growing difficulty in finding adequate support to fund the conferences. Until the present conference, support came almost entirely from the National Center for Health Services Research (later the National Center for Health Services Research and Health Care Technology Assessment and now the Agency for Health Care Policy and Research [AHCPR]) and the National Center for Health Statistics (NCHS). Over these years, the conferences received some ancillary support from Veteran's Affairs, the Milbank Memorial Fund, and the Commonwealth Fund. However, by this sixth conference, a broad base of support from a variety of federal agencies and private foundations that make use of survey data was needed to ensure the conference would be held. In addition to AHCPR, which supported the conference through a grant, and NCHS, which supported the conference and is publishing the proceedings, funding for this conference came from the Robert Wood Johnson Foundation, National Institute on Alcohol Abuse and Alcoholism, Substance Abuse and Mental Health Services Administration, National Cancer Institute, and Health Resources Services Administration. The fact that we were able to obtain this broad base of support reflects the growing interest in and need for valid survey data among many of the major federal and private users of such data. The broadened base of support required a conscientious effort on the part of the organizers of the sixth conference to recruit investigators and policy makers from many agencies that are concerned with health and are reliant upon health data. Their participation will broaden dissemination of the results and enhance the influence of the conference findings on how survey data are collected and used in the planning and implementation of health services.

Conference Themes: The Past as Prologue to the Present

The focus of the conferences has always been on survey methods, particularly on nonsampling error, although the specific themes have varied from conference to conference. For example, all the conferences but the present one had a specific session addressing the issue of total survey design. All six conferences have addressed issues of validity of survey data both in independent sessions and through consideration of questionnaire design, respondent recall and burden, and validation of survey responses through records. Most have had a session on sample design and locating rare populations or hard-to-contact respondents. At each conference, there has also been a session on mode of data collection.

How these general themes were addressed at each conference was affected by the major policy issues for which the data were needed at the time of each conference. In 1979, the government was beginning to collect data on access to and costs of various elements of health care. The major themes in the Third Conference on Health Survey Research Methods reflected those policy objectives. Most of the paper sessions and several special sessions in the program addressed the design and implementation of major surveys designed to assess the costs of health care and access by the population to health care services. General sessions contained papers on the design of the National Ambulatory Medical Care Survey and the National Medical Care Expenditure Survey (NMCES). There were also special sessions during which the recommendations of the Technical Consultant Panel for the redesign of the National Health Interview Survey and selected methodological features of the 1980 Census were presented. The agenda for the fourth conference explored further methodological issues related to the NMCES and health surveys in other countries.

By the fifth conference, there had been a major shift of focus from access and costs of health care to the epidemiology of AIDS and how to measure seroprevalence. Although there were at least two papers addressed to issues of access and costs, the theme of the fifth conference was the effects of total survey design on surveys related to homelessness, AIDS, and measuring seroprevalence. Once again, the major new survey strategies and designs being planned to address the major health problems were the major conference themes. Those attending heard a conference address called "Designing a Household Survey to Estimate HIV Prevalence: An Interim Report on the Feasibility Study of the National Household Seroprevalence Survey." An entire session was devoted to topics related to the National Survey of Health and Sexual Behavior. Also given was a conference address called "Obligations Attending Gaining Information: A Moral Question for Health Survey Researchers," which focused attention on the issues surrounding collecting data on highly sensitive topics, such as sexual practices, and updated themes of dealing with confidential data that had appeared in the first, second, and fourth conferences.

Another major policy issue related to health concerned estimates of the homeless population, and the conference also heard an address titled "1990 Census: Counting Selected Components of the Homeless Population." Finally, a full session of this conference was devoted to the method-

ological issues associated with the major surveys of older adults and nursing home populations.

Discussion of the major surveys of the elderly and on AIDS continued in the present conference, but no new major surveys were introduced by the federal survey agencies. Instead, as indicated by a session titled "Integrating Surveys and Other Data," the sixth conference heard more about a growing interest by federal programs in using existing data sets in creative ways for program evaluation and obtaining policy relevant data. Presentations in this session described creative strategies being used to assess various kinds of outcomes through the use of add-on items to existing surveys, administrative data, and other strategies. The presentations focused on the opportunities and challenges that such strategies provide.

Through the years, discussions of sampling have seemed to focus on identifying rare or hard-to-locate populations. The third and fourth conferences heard discussions of multiplicity and network sampling as tools to identify and sample cancer patients and other relatively rare populations. By the fifth conference, there was still a strong focus on identifying and sampling rare and hard-to-reach populations, but the sampling universes of interest had shifted to homeless populations, prostitutes, and people with AIDS or HIV infection, and issues surrounding the external validity of generalizations based on these frames were also considered. Another emerging theme related to access was sampling of elderly populations. Major surveys of elderly populations had been introduced during the periods covered by the fifth and sixth conferences, and discussion of the sampling and interviewing problems associated with these surveys, some of which were panels, was a recurring theme of these conferences.

During the present conference, there were updates on sampling issues concerned with sampling the aged, populations of street prostitutes, homeless populations, and populations at high risk for HIV infection due to drug use or sexual practices. However, a new issue—the problem of defining appropriate frames for sampling patients for studies assessing satisfaction with medical care—was also introduced in the present conference as part of an overall policy theme associated with measuring patient status and treatment outcomes. These issues of measuring patient status and satisfaction were also considered in the fourth and fifth conferences, but the urgency and scope of discussion in the present conference signaled the growing policy concerns in this area.

From the earliest conferences, the issues of respondent burden and the validity of the data provided by respondents have been major themes. When the conference format shifted from thematic discussions to papers describing specific projects, several specific themes emerged, particularly in the third and fourth conferences. In both of these conferences, sessions addressed how to obtain reliable data using memory aids and diaries. These were strategies appropriate for collecting data from providers and patients in face-to-face or mail format in which use of computer-based feedback programs for panel respondents to assist them in updating their reports of contacts with the health care system, diaries, and other visual aids were appropriate. During the fifth and sixth conferences, the themes shifted from visual aids to strategies of questionnaire design and administration based on cognitive theory. These cognitive-based strategies were introduced briefly during the fifth conference and by the sixth have been the focus of two sessions, a specific session on questionnaire design and one on measuring patient status and satisfaction.

These shifts in emphasis reflect in part the ongoing change in mode of interviewing toward telephone and away from face-to-face strategies. They also reflect the growing influence of theories of cognition on questionnaire design. With the introduction of cognitive theory, there has been a more general resurgence of interest in the whole issue of how the way the questions are asked affects the validity of the results. Introduced in the fifth conference, the issues of pretesting, monitoring interviewer-respondent interaction, and cognitive testing of questions have become major themes in conferences on survey measurement.

A session on mode effects again focused on technological changes in the way interviewing takes place. As in preceding conferences, various forms of computer-based data collection were considered. In particular, new strategies for collecting data from youth were discussed. These focused on both computerized data collection and use of tape recorders. Also presented were strategies using computerized survey formats to gain access to very reticent populations. These topics all represent important innovations that have emerged over the past 10 years as the technology has developed.

Richard B. Warnecke
Director, Survey Research Laboratory
University of Illinois at Chicago

Previous Health Survey Research Methods Conference Proceedings

NCHSR research proceedings series: Advances in health survey research methods (DHEW Publication No. [HRA] 77-3154). Rockville, MD: National Center for Health Services Research.

Reeder, Leo G. (Ed.). NCHSR research proceedings series: Health survey research methods (DHEW Publication No. [PHS] 79-3207). Hyattsville, MD: National Center for Health Services Research.

Sudman, Seymour (Ed.) NCHSR research proceedings series: Health survey research methods (DHHS Publication No. [PHS] 81-3268). Hyattsville, MD: National Center for Health Services Research.

Cannell, C. F., & Groves, R. M. (Eds.). (1984). NCHSR research proceedings series: Health survey research methods (DHHS Publication No. [PHS] 84-3346). Hyattsville, MD: National Center for Health Services Research.

Fowler, Floyd J. Jr. (Ed.). (1989). Conference proceedings: Health survey research methods (DHHS Publication No. [PHS] 89-3447). Rockville, MD: National Center for Health Services Research.

Acknowledgments

Specific thanks are due to Elinor Walker (Agency for Health Care Policy and Research), who was the program officer for the conference grant from that agency and whose creative assistance made it possible for many of the other sponsors of the conference to participate. She was assisted by Ralph Sloat, the grants manager at the Agency for Health Care Policy and Research. Marcie Cynamon and Owen Thornberry represented the National Center for Health Statistics, and through their efforts, we were able to obtain the commitment of the National Center for Health Statistics to publish these proceedings. In addition, Marcie was extremely helpful in putting together the coalition of funders listed above. Without her efforts, the resources needed for this conference would not have been located. Representatives from the funding agencies served as active members of the conference planning committee and participated in the program. Special recognition should be made of the contributions of Brenda Edwards (National Cancer Institute), Michael Hilton (National Institute on Alcohol Abuse and Alcoholism), James Knickman (Robert Wood Johnson Foundation), Katherine Marconi (Health Resources Services Administration), and Deborah Trunzo (Substance Abuse and Mental Health Services Administration). Academic members of the planning committee included Lu Ann Aday (University of Texas at Houston), Richard Campbell (University of Illinois at Chicago), James Chromy (Research Triangle Institute), Floyd J. Fowler Jr. (University of Massachusetts–Boston), Richard Kulka (Research Triangle Institute), and James Lepkowski (University of Michigan).

A successful conference is always a combination of interesting and important presentations, good discussions, and carefully compiled proceedings. However, a conference such as this is heavily dependent upon organization to succeed. Attendance was by invitation. Participants included representatives from academia, whose travel was supported by the funding from the conference grants, and government representatives, who had to pay their own expenses. The location of a conference where participants are asked to stay for 2 to 3 days and work very hard is always an important element in its ultimate success. Thus, special recognition and thanks are due Diane O'Rourke of the Survey Research Laboratory at the University of Illinois, who handled all the arrangements, including site selection and negotiation, travel, and special needs of all participants. She did a wonderful job. Diane was assisted by Bernita Rusk and Jennifer Parsons, also of the Survey Research Laboratory, at the conference site. Marya Ryan, the Survey Research Laboratory's coordinator of survey research information services, edited these proceedings with clerical support from Bernita Rusk.

Finally, as chair of the conference planning committee, I want to thank all the participants who gave up a weekend in June to participate in this conference. In the final analysis, the success of these endeavors is always dependent upon the willingness of those who are doing the cutting edge methodological research to share their work and, more importantly, their ideas in this venue. Over the years, we have been very successful in recruiting these key participants, and this conference was no exception. Their willing cooperation and the help of the planning committee and Survey Research Laboratory staff made chairing this conference a real pleasure. I appreciate the honor of being asked to do it.

Richard B. Warnecke

List of Acronyms

ABMS	American Board of Medical Specialties
ACASI	audio computer-assisted self-interviewing
ACS	American Cancer Society
ADL	activity of daily living
AFOMS	Automated Field Office Management System
AHA	American Hospital Association
AHCPR	Agency for Health Care Policy and Research
AHEAD	Asset and Health Dynamics of the Oldest Old
AIDS	acquired immunodeficiency syndrome
AIM	Aging in Manitoba Longitudinal Studies
AMA	American Medical Association
ARF	Area Resource Files
ASSIST	American Stop Smoking Intervention Study
BLSA	Baltimore Longitudinal Study of Aging
CAI	computer-assisted interviewing
CAPI	computer-assisted personal interviewing
CARE	Comprehensive Assessment and Referral Evaluation Depression Scale
CASI	computer-assisted self-interviewing
CATI	computer-assisted telephone interviewing
CDC	Centers for Disease Control and Prevention
CES-D	Center for Epidemiologic Studies Depression scale
CHAMPUS	Civilian Health and Medical Program for the Uniformed Services
CHAMPVA	Civilian Health and Medical Program, Veteran's Affairs
CIDI	Composite International Diagnostic Interview
COOP	Primary Care Cooperative Information Project
CPCI	Components of Primary Care Instrument
CPS	Current Population Survey
CPT	Common Procedure Terminology
CSFII	Continuing Survey of Food Intakes by Individuals
DC*MADS	the Washington, DC, Metropolitan Area Drug Study
DHHS	Department of Health and Human Services
DRG	Diagnostic Related Group
DSM	Diagnostic and Statistical Manual of Mental Disorders (Roman numerals designate edition; "R" designates revision)
ECA	Epidemiological Catchment Area Study
EDB	Enrollment Data Base (of Medicare beneficiaries)
ELISA	enzyme-linked immunosorbent assays
ESRD	end stage renal disease
GMS	Geriatric Mental State Interview
GoM	Grade of Membership model
HALS	Health and Activity Limitation Survey
HCFA	Health Care Financing Administration
HICN	health insurance claim number
HIV	human immunodeficiency virus
HLM	hierarchical linear modeling
HMO	health maintenance organization
HRS	Health and Retirement Survey
HRSA	Health Resources Services Administration
IADL	instrumental activity of daily living
ICD	International Classification of Diseases (numeral designates revision)
IDU	injection drug user
IOM	Institute of Medicine
IRB	institutional review board
IV	intravenous
LOS	length of stay
LPN	licensed practical nurse
LSOA	Longitudinal Study of Aging
MACS	Multi-City AIDS Cohort Study
MCBS	Medicare Current Beneficiary Survey
MCI	mild cognitive impairment
MD	doctor of medicine
MEC	mobile examination center
MEDPAR	Medicare Provider Analysis and Review
MOS	Medical Outcomes Study
MPS	Medical Provider Survey
MSA	metropolitan statistical area
NCH	National Claims History database
NCHS	National Center for Health Statistics
NCHSR	National Center for Health Services Research
NCI	National Cancer Institute
NDI	National Death Index
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NHSDA	National Household Survey on Drug Abuse
NIDA	National Institute on Drug Abuse
NMCES	National Medical Care Expenditure Survey

NMES	National Medical Expenditure Survey
NSAM	National Survey of Adolescent Males
NSFG	National Survey of Family Growth
PACTAPH	portable audiocassette tape player with headphones
PANAS	Positive Affect/Negative Affect Scale
PAPI	paper-and-pencil interviewing
PORT	Patient Outcomes Research Team
PPS	probability proportional to size
PSU	primary sampling unit
PTSD	post-traumatic stress disorder
RDD	random-digit dialing
RN	registered nurse
RTI	Research Triangle Institute
SAMHSA	Substance Abuse and Mental Health Services Administration
SAQ	self-administered questionnaire
SCLD	State Cancer Legislative Database
SEER	Surveillance, Epidemiology, and End Result system
SHPPS	School Health Policies and Programs Study
SOPH	subjectively rated overall physical health
SPMSQ	Pfeiffer Short Portable Mental Status Questionnaire
SSA	Social Security Administration
SSI	Supplemental Security Income
SSU	secondary selection unit
STD	sexually transmitted disease
TKR	total knee replacement
TUS	Tobacco Use Supplement (to the CPS)
UPC	usual provider of care index
USDA	U.S. Department of Agriculture
VA	Veteran's Affairs
WAIS	Wechsler Adult Intelligence Scale
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index
YRBS	Youth Risk Behavior Survey

Contents

Measuring Medical Care and Health Status

This session focuses on two principal themes. Six feature papers and two discussion papers comprise the session. The first three address issues of measurement and focus on themes such as the reliability and validity of measures of disability and components of primary care. These papers by Verbrugge, Merrill and Liu; Ofstedal, Lentzner and Weeks; and Flocke, Stange, and Zyzanski raise important psychometric issues regarding the measurement process. The second three papers in this session by Schwarz, Mathiowetz, and Belli; Fowler and Bin; and Harris, Tierney, and Weinberger address the related issue of response effects on validity and reliability. These papers point to the subjective effects of response and how they are affected by question context and by patterns of nonresponse.

Disability Parsimony

Lois M. Verbrugge, Susan Merrill, and Xian Liu

Introduction

Disability is a multifaceted phenomenon. Health-related limitations can occur in numerous roles and activities, such as job performance, personal care, household management, socializing with friends, active recreation, and passive leisure. There are also various dimensions of those limitations, such as degree of difficulty, use of equipment or personal assistance, pain while engaged in the activity, and satisfaction with performance. Faced with such diversity, when designing surveys, researchers select certain activities and dimensions that seem most germane for the age-gender groups studied or for public policy. The most common choices are questions about difficulty or assistance in performing personal care (activity of daily living [ADL]), household management (instrumental activity of daily living [IADL]), and job activities. Even with this restricted scope, the number of disability questions in surveys has become large, posing burdens for interviewers, respondents, survey analysts, and disability statistics users.

In distinct contrast to this situation, there has been movement toward parsimony in measuring morbidity. Although health status is also multifaceted (involving the presence/absence of specific conditions, severity, duration, etc.), a global item to summarize it is routinely included in surveys: self-rated health. Its value for prediction of dire outcomes, such as institutionalization and death, is equal to or better than arrays of detailed morbidity items. The item is brief to administer and has good colloquial merit (it makes sense to respondents). In short, one question about health happens to be realistic, comprehensive, and prescient.

The sharp difference between survey approaches to disability and morbidity is the underlying motivation for this paper. For all the interest in and wide use of a global morbidity item, there has been little work to develop and use a global disability item. It may indeed be possible to find one that has strong analytic value and also captures the real-world experience of disability well. If so, it should be regularly included in health surveys either as a companion to detailed disability questions (in surveys with extensive focus on chronic morbidity and functioning) or by itself (in surveys with sharp time limits or brief coverage of morbidity/functioning). Alternatively, it may be possible to reduce the number of detailed items about disability by dropping some activities or dimensions with little loss of analytic value. Either approach—a global item or reduced items—will achieve disability parsimony, but the global item is certainly most economical.

This paper has four parts: First, we provide some background for thinking about disability measurement and the issue of parsimony. Second, we present empirical results from two projects—one relates chronic conditions to a great number of detailed disability items and the other relates a global disability item to detailed ones and to self-rated health. Third, from the results, we draw conclusions about (a) how detailed disability items can be culled and (b) whether global disability is worth adding to surveys. Lastly, recommendations are offered about work that can be done with existing surveys or in small-scale, laboratory-based studies to promote compact questioning of disability in health surveys.

Background on Global and Detailed Disability Items

"Disability" refers to the impacts health problems have on people's social functioning, that is, their ability to perform roles and activities (Pope & Tarlov, 1991; Verbrugge & Jette, 1994). "Social functioning" includes the whole range of typical and personally desired activities an individual does, ranging from the most basic and universal (such as eating and dressing) to the most discretionary and distinctive (such as a person's favorite hobby or recreation). Disability can be short-term or long-term, and it can be due to acute or chronic conditions. Because research and policy interests are typically on long-term dysfunctions associated with chronic conditions, that is our focus here.

The aim of a global disability item is to measure overall social functioning briefly but well. It must refer to protracted, health-related difficulties in a large span of activities. The question format can be one single question, a branch-and-stem item (main question plus probes about duration and health relatedness), or a small set of questions (short ones that are combined into a single variable during

Lois M. Verbrugge, Distinguished Research Scientist; Susan Merrill, Postdoctoral Fellow; and Xian Liu, Assistant Research Scientist, are at the Institute of Gerontology, University of Michigan, Ann Arbor.

analysis). These formats are compact in the questionnaire itself and thus brief to administer and easy to analyze.[1]

There are some examples of global disability items in contemporary U.S. and Canadian surveys (Verbrugge, 1994). Most have been created through a mixture of judgment and consultation, copying from prior surveys, and pretesting. Ideally, choices should be based also on empirical evidence about content (what aspects of disability the item covers) and analytic value (relationships to predictors or outcomes). Little evidence of that sort exists, so good craftsmanship is the mainstay for designing items.

What sorts of methodological work can help in evaluating global disability items? There are two basic approaches: cognitive and statistical. Cognitive approaches are well suited to studying the processes that respondents use to think about questions and come up with answers. These studies are usually small-scale and often laboratory based. Statistical approaches are used on moderate- to large-scale data sets to study multivariate item structure, reliability, and concurrent and predictive validity. Examples of analyses that can inform us about global disability are analyses of (a) relationships between a global item and specific disability items to determine the global item's included and excluded content, (b) relationships between global disability and global morbidity to see if they are nonredundant, (c) models relating chronic conditions and global disability to assess its health relatedness, and (d) the prediction ability of global disability by itself (apart from self-rated health) on subsequent outcomes. In general, the evidence compares and contrasts global disability with detailed disability items and global morbidity. Ideally, one wants a global disability item to have good coverage of detailed disabilities (high correlations with them) and be distinct from self-rated health (moderate to low correlation and strong net relationship to outcomes).

The notion of parsimony is also relevant for surveys that contain detailed disability items. The questions are usually about a rather narrow set of activities (ADLs, IADLs, job activities; also, physical and sensory limitations) with several dimensions for each (difficulty, equipment assistance, personal assistance). Parsimony could be achieved by reducing the number of detailed items. Statistical approaches can inform us on this issue; for example, in addition to (a) through (d) in the previous paragraph, analyses of (e) relationships between chronic conditions and specific disabilities to assess whether disabilities have similar morbidity precursors, (f) how detailed items predict prospective dire outcomes, and (g) clustering and hierarchy of items assessing if any given detailed question actually represents a whole disability profile. Items with low health relatedness or low prediction can be considered for elimination. If scaling analyses show strong hierarchy, then an economical approach to asking about disability can be considered (items are ordered according to the scale, questioning begins somewhere in the middle, and it proceeds up or down the scale until a "yes" for disability occurs).

We now present results of two projects motivated by our interest in parsimony. One studies the health relatedness of numerous detailed disability items ([e] above). The other studies relationships of a global disability item to detailed disabilities ([a]) and self-rated health ([b]) and also the global item's health relatedness ([c]). We draw conclusions about winnowing detailed items and about the merits of a global item.

Morbidity Precursors of Detailed Disabilities

Are chronic conditions strongly related to presence and degree of disability or only weakly so? Are the links between morbidity and disability distinctive (different chronic conditions are implicated for each disability) or nondistinctive (the same conditions come into play for virtually all disabilities)? The answers will indicate the health relatedness of dysfunctions and similarities in morbidity-disability relationships.

We utilized data from the Asset and Health Dynamics of the Oldest Old (AHEAD) Survey Wave 1 (Merrill & Verbrugge, 1995). AHEAD is a population-based sample of U.S. community-dwelling persons aged 70 and older at Wave 1 (1993–94; N = 8,224). The questionnaire has information on the presence/absence of 25 chronic conditions, 22 specific disabilities (ADLs, IADLs, and physical limitations,[2] with various dimensions: degree of difficulty, use of assistance, need for assistance, pain when doing activity, tiredness when doing it, long time to do it), and 9 productive activities. The disability items were used as is and also in aggregated forms (such as "any ADLs" and "sum/ of ADLs"). The full set of chronic conditions (X) were related to each disability outcome (Y) by logistic and linear regressions, controlling for age and gender.

Descriptive statistics for variables and tables with results are in Merrill and Verbrugge (1995). Three tables most pertinent to this article are available on request: One illustrates results for detailed disability, the next illustrates results for aggregated disability, and the last lists chronic conditions that always/almost always have significant relationships with disability items.

The most striking result is that the same eight chronic condition routinely have statistically significant associations

[1]We distinguish them from two other formats: (a) An aggregated item adds up the number of specific disabilities. This is analytically compact, but not compact in the questionnaire itself. (b) A short-form instrument covers multiple, diverse concepts about health and functioning with about 5 to 20 questions total. By contrast, a global indicator covers just one concept.

[2]Conceptually, physical limitations are aspects of functional limitation, not disability (Verbrugge & Jette, 1994). For the sake of economy, this is not emphasized in the paper.

with the many disability items (detailed and aggregated). They are stroke, diabetes, arthritis, hip fracture, urinary incontinence, poor vision, frequent pain, and the residual "other conditions."[3] The other chronic conditions are related to certain disabilities or disability domains, but not consistently across the board.

R^2s are generally .10 to .20 for specific ADLs and IADLs and .20 to .30 for specific physical limitations. Aggregated variables (such as any ADLs and sum of ADLs) produce higher R^2s than their detailed source items, the increase being about .10. Moreover, more chronic conditions have significant relationships with these aggregate items than with the detailed ones.

The results lead to two conclusions. First, there is plenty of redundancy in the health relatedness of disability items. Thus, if a survey needs to include the topic of disability but does not really need disability details, then any four to five items will serve that purpose adequately. The most sensible choice is asking about one dimension (such as difficulty) for several diverse activities (spanning ADLs, IADLs, and physical limitations). Second, the association between morbidity and specific disabilities (R^2) is modest but increases notably for aggregated disability variables. Thus, for analytic parsimony, one should use the aggregates and skip the detailed items. But there is no fieldwork parsimony in this approach, since aggregate variables depend on having asked the plethora of detailed items! In short, the AHEAD analyses suggest how to use detailed items with parsimony in two ways: by reducing the number of detailed items placed in a questionnaire or, if that doesn't happen, by reducing the number of disability variables analyzed.

Distinctive Features of Global Disability

Is global disability related to all specific disabilities or to some far more strongly than to others? Is global disability closely related to global morbidity (self-rated health) or weakly so? How health related is global disability? The answers will indicate how well a global item compasses activity domains, whether it is really something different from global morbidity, and how well it reflects underlying health problems.

We utilized data from the Health and Retirement Survey (HRS) Wave 1.[4] HRS is a population-based sample of U.S. community dwellers aged 51 through 61 in 1992 plus their spouses (N = 12,654). The questionnaire has information on disability (limitation in job performance, housework, or other activity, five ADLs and difficulty doing each, five physical limitations and any difficulty doing each; no IADLs asked) and health status (presence/absence of 19 specific conditions, self-rated health).

We used the limitation in job/housework/other items to create a global disability variable, as follows: All persons were asked if they have an impairment or health problem that limits the kind or amount of paid work they can do; the subset saying "no" were asked about health-related housework limitations; and lastly, the further subset saying "no" to the housework question were asked about health-related "limitation in any way in activities." Our global variable is dichotomous, scored 1 for "yes" to any of the three items and 0 otherwise.[5] The percentages are 29.5% disabled and 70.5% nondisabled. (The building block percentages from the three items are 21.5% for job limitation, an additional 3.5% for housework limitation, and an additional 4.5% for other limitation.) The other disability items were used as is and also in aggregated forms (such as "any ADLs" and "sum of ADLs").

For the first part of the analysis, detailed disabilities (X) were related to global disability (Y) in logistic regressions. For the second part, aggregate and global disability (X) were related to self-rated health (Y) in logistic regressions. Lastly, we were able to look again at the health relatedness of disability, this time with a genuine global disability item. Age, gender, and education were controlled in all regressions.

Descriptive statistics for variables are available on request. Three tables are included here.

How are detailed disabilities associated with the global indicator (see Table 1)? Three models were estimated: with both physical limitations and ADLs as predictors (Model I), just ADLs (Model II), and just physical limitations (Model III). On their own, physical limitations are strongly linked with global disability (Model III). Each item ("walk several blocks," "climb several stairs," "pull/push large objects," "lift/carry 10 pounds," "pick up a dime") has a statistically significant coefficient for the total sample and each gender. Similarly, difficulties in ADLs ("walk across room," "bathe," "transfer in/out of bed," "dress," but not "eat") are also associated with global disability (Model II). But in models with combined predictors (Model I), the ADL coefficients fade in size while those for physical limitations remain essentially as large as before. Log likelihood values show this relative importance as well; ADLs add almost nothing to the prediction strength of physical limitations (comparing Models III and I).

The relationship of disability to self-rated health was studied next (see Table 2). Self-rated health is scored in two ways, as a dichotomous variable of "poor" versus other responses ("fair," "good," "very good," "excellent") and a dichotomous variable of "poor"/"fair" versus "good"/"very

[3]Questionnaire items for determining presence/absence vary for the conditions (e.g., physician diagnosis of condition, own statement about presence of condition, symptoms in past year). Details can be found in AHEAD documents (it is a public use data set) or the manuscript cited.

[4]The HRS analyses were conducted by authors Verbrugge and Liu.

[5]Because of the sequential questioning, the two items on housework limitations and other activity limitations cannot be analyzed on a whole sample basis, and prevalence rates for them cannot be estimated from the HRS.

Table 1. Effects of specific disability items on global indicator (logistic regression)

Explanatory variables	Model I (n = 11,972)	Model II (n = 12,429)	Model III (n = 12,035)
ADL items			
Walk across room	0.685	2.343	
Bathe	0.944	2.281	
Transfer in/out of bed	0.697	1.807	
Eat	–0.918	0.384	
Dress	0.794	1.488	
Physical limitation items			
Walk several blocks	0.964		1.045
Climb several stairs	0.704		0.869
Pull/push large objects	1.214		1.277
Lift/carry 10 pounds	0.945		1.024
Pick up a dime	0.398		0.569
Intercept	–2.829	–2.353	–2.762
Log Likelihood	4,058.03	2,063.80	3,985.42

Table 2. Association between global health indicator and global disability indicator

Explanatory variables	Model I (n = 12,443)	Model II (n = 12,053)	Model III (n = 12,598)
Poor self-rated health			
ADL indicator	2.626		
Physical limitation		2.911	
Global disability			3.455
Intercept	–1.936	–2.623	–2.589
Log likelihood	1,646.52	1,434.72	2,250.01
Poor/fair self-rated health			
ADL indicator	2.318		
Physical limitation		2.026	
Global disability			2.418
Intercept	–0.791	–0.861	–0.707
Log likelihood	2,350.69	2,542.08	3,565.42

good"/"excellent". Only 8.0% of this middle-aged sample reported poor health; 14.3% reported fair health. Results show that our models do a better job predicting "poor"/"fair" health than "poor" health—not surprising, given the rarity and thus unusual circumstances underlying poor health at these ages. Of the three disability variables, global disability has the strongest relationship to self-rated health; this is seen both in coefficients and log likelihood values.

Lastly, we studied relationships of chronic conditions to detailed, aggregated, and global disability. The results are surprising and welcome: Chronic conditions are excellent predictors of global disability, more so than for aggregated disability (see Table 3) and much more so than for detailed items (table available from authors on request). This extends and replicates the AHEAD results (which showed that aggregated items were better than detailed ones). On this basis, we can state that there is a hierarchy of health relatedness for disability variables, with global disability

Table 3. Effects of specific chronic conditions on three disability indicators

Explanatory variables	ADL indicator (n = 12,091)	Physical limitation (n = 11,711)	Global indicator (n = 12,245)
Hypertension	0.113	0.310	0.211
Diabetes	0.681	0.644	0.568
Cancer	0.386	0.360	0.652
Lung disease	0.301	0.756	0.731
Heart disease	0.269	0.725	1.012
Stroke	1.102	0.878	1.533
Psychiatric problems	0.423	0.463	0.821
Arthritis	0.687	0.380	0.586
Kidney disease	0.394	0.284	0.492
Other diseases	0.561	0.483	0.625
Intercept	–3.604	–2.376	–2.919
Log likelihood	2,141.77	3,176.14	4,119.18

ranking best of all. Stroke, heart disease, and psychiatric problems are the strongest predictors of global disability. Stroke also proved consistently strong in the AHEAD results.

We arrive at three conclusions: First, physical limitations are the foundation for disability in midlife. ADL difficulties are not very common at these ages, but even so, their presence is much less predictive of global disability than are physical limitations. Stated another way, generic functional problems are more implicated in general disability status than any specific disabilities are. This might strike some readers as odd—the global item is more closely related to its precursors than its components. Whether this result holds up in older samples and in data sets with larger arrays of physical, mental, and social functioning items remains to be determined. Second, there is sizable overlap between global disability and global morbidity. This result was expected. Our analyses are very simple, and a firmer judgment of what "sizable" really means would come from models using both global disability and global morbidity as predictors (X) of concurrent or prospective outcomes (Y). Because global items are rare, we found no examples of such analysis in the literature; there are examples with multiple or aggregated disability items as predictors. Third, global disability is far more health related than are activity domains (ADLs, physical limitations) or detailed activities. This is a welcome result; its strength surprised us.

Conclusions

Integrating the analyses above, we come to the conclusions discussed below for detailed disability and global disability.

Plenty of detailed items are appropriate in surveys if every single one of them has a scientific or public policy rationale. Each one will be analyzed on its own at some point to fulfill those initial purposes. But for more general analyses of the data, aggregated variables such as "any ADLs" or "sum of ADLs" have better analytic yield. (In our AHEAD analyses, this specifically meant stronger health relatedness.)

In many population health surveys, there is no good rationale for having numerous detailed items on disability. Will just a few do, and if so, which ones? The AHEAD results suggest that the many specific ADLs, IADLs, and physical limitations have similar relationships to chronic morbidity, so choosing any small set of them will suffice to represent disability. This is acceptable for a cross-sectional survey setting. In a longitudinal setting, prediction ability as well as health relatedness must come into the winnowing decision. One needs to know from prior studies if detailed items have similar prospective prediction or not.

A global disability indicator reflects the disablement process very well. The HRS results show that it has stronger relationships to causal precursors (chronic conditions and physical limitations) than detailed disability or aggregated disability variables do. That is a plus in its favor. But global disability has strong overlap with global morbidity (self-rated health). The extent of overlap needs more explicit study by comparing the two items' strength as predictors of concurrent and prospective outcomes. At issue is the net effect of global disability, controlling for self-rated health.

The HRS indicator is oddly constructed and not ideal; the component items were not designed with their pooling into a global item in mind. Nevertheless, the indicator is analytically sturdy, showing distinctive and systematic results when compared with detailed disability, aggregated disability, and self-rated health. We have no doubt that overtly designed global items will do as well—and likely better.

Recommendations

Four recommendations for research and two for questionnaire design that spring from our work are offered below.

What research can be done, economically and soon, to further the goal of parsimonious questioning about disability? We make four recommendations for research.

First, existing data sets with numerous detailed items can be analyzed, closely studying item correlations and scaling characteristics. The motivation for the work is not just psychometric analysis but practical decision making about (a) items that can be dropped or (b) efficient questioning strategies in an ordered series of items.

Second, with imagination, global indicators can be generated from existing data sets. Many surveys now have series of activity limitation questions that can be pooled into a single variable (as the National Center for Health Statistics (NCHS) routinely does for the National Health Interview Survey [NHIS]). Or numerous detailed items can be pooled into an "any disability" variable. The analytic merits of these pooled variables can be compared with detailed items.

Third, surveys with genuinely global items are few and far between, but the search for them should be made and opportunities exploited. We are currently analyzing data from the Centers for Disease Control and Prevention Behavioral Risk Factor Surveillance System Survey, which included global morbidity and disability items for the first time in 1993. We also note the two Health and Activity Limitation Surveys conducted in Canada and the 1994–95 Disability Supplement for the NHIS (NHIS-Disability) in the United States; we leave their analytic potential to readers' scrutiny.

Fourth, global items can be crafted and then evaluated for colloquial sense and content in laboratory settings. A crucial aspect of this work is to determine the best place for the two essential qualifiers (disability is protracted and health related).[6] Nonverbal formats, such as the COOP

[6]The several options are to place these qualifiers in an initial preface (asking respondents to think about long-term health-related problems in the following questions), in each question, or in follow-up probes (checking about duration and health relatedness after respondents say "yes" to the disability question). Which approach achieves and maintains the desired focus without excessive verbiage?

Chart of Daily Activities (Beaufait et al., 1992), should be considered and tested in conjunction with verbal ones.

What can be done immediately, without additional research information, when designing surveys? We make two recommendations for survey design.

First, every survey that includes self-rated health should also include a global disability item. The briefest rationale is that functional status is just as important as health status. Items used in other surveys to date are shown in Verbrugge (1994), and good candidates are noted for consideration in future surveys. They appear here as Figure 1.

Second, if detailed items are needed, every one should have excellent rationale and conceptual integrity. Its analytic use should be known in advance (if it doesn't exist, neither should the item). The conceptual niche that each holds should be stated clearly. Further, overall coverage of the concept "disability" should be considered afresh when a survey is designed. This means resisting the pressures, which are very strong, to repeat items used in other surveys. For example, if n questions are desired, surveys can have better coverage of the disability experience by asking about more activity domains and just one dimension, in contrast to contemporary practices of asking about few domains and several dimensions.

Summing up, the goal is to measure disability in comprehensible, comprehensive, veridical, and useful ways in health surveys. We think it can be done with more parsimony than now exists.

References

Beaufait, D. W., Nelson, E. C., Landgraf, J. M., Hays, R. D., Kirk, J. W., Wasson, J. H., & Keller, A. (1992). COOP measures of functional status. In M. Stewart, F. Tudiver, M. J. Bass, E. V. Dunn, & P. G. Norton (Eds.), Tools for primary care: Research methods for primary care (pp. 151–167). Newbury Park, CA: Sage.

Merrill, S. S., & Verbrugge, L. M. (1995). Evaluating the many facets of disability. Manuscript submitted for publication.

Pope, A. M., & Tarlov, A. R. (Eds.). (1991). Disability in America: Toward a national agenda for prevention. Washington, DC: National Academy Press.

Verbrugge, L. M. (1994). A global disability indicator: Companion to self-rated health. In S. Schechter (Ed.), Proceedings of the 1993 NCHS Conference on the Cognitive Aspects of Self-reported Health Status (NCHS Cognitive Methods Staff Working Paper Series, No. 10, pp. 60–88). Hyattsville, MD: Office of Research and Methodology, National Center for Health Statistics.

Verbrugge, L. M., & Jette, A. M. (1994). The disablement process. Social Science and Medicine, 38, 1–14.

Figure 1. Candidates for a global disability indicator

1. National Population Health Survey, Canada[a]

"The next few questions deal with any health limitations which affect . . . 's daily activities. In these questions, 'long-term conditions' refer to conditions that have lasted or are expected to last 6 months or more."

"Because of a long-term physical or mental condition or a health problem, are you limited in the kind or amount of activity you can do:

At home?
At school?
At work?
In other activities such as local travel, sports or leisure?"

For each: yes, no.

2. New (developed by author)[b]

"Because of a physical, mental, or emotional condition, are you limited in doing your daily activities like personal hygiene, house or yard care, shopping, your work, or other things you need to do?" Yes, no.

If yes: "Has the limitation lasted for at least 6 months or is it expected to last that long?" Yes, no.

If yes to 6+ months: "Are you limited just a little, somewhat, or a great deal in your daily activities?" Just a little, somewhat, a great deal.

3. Modified from a pilot study on subjective health[c]

"Is there anything about your health that makes it hard for you to do your usual activities?" Yes, no.

If yes: "Has the difficulty with your activities lasted 6 months or more, or do you expect it to last that long?" Yes, no.

If yes to 6+ months: "What are the activities you have trouble doing because of health?" Interviewer records responses.

"Would you say your difficulty doing these activities is a little, some, or a lot?" A little, some, a lot.

4. Modified from NHIS-Disability, United States, 1994–95[d]

After specific questions about physical conditions, if yes to any: "During the past 12 months, did any of these problems seriously interfere with your ability to work or attend school or to manage your day-to-day activities?" Yes, no.

After specific questions about cognitive and emotional problems, if yes to any: (same question).

5. Modified from the Baltimore Longitudinal Study of Aging (BLSA) Follow-up 1[e]

"Would you describe your overall level of functioning in your home, work, and leisure activities as: excellent, very good, good, fair, poor, don't know?" Excellent, very good, good, fair, poor.

[a]Canada has used global disability items in its population census and in national surveys such as the 1986–87 and 1991 Health and Activity Limitation Surveys (HALSs) and the 1994–95 National Population Health Survey. The items have been very similar, with a little modification from one census/survey to the next. We show the contemporary item but alter the descriptor for other activities from "such as transportation to or from work or leisure time activities" to "such as local travel, sports or leisure," a close reprise of what appeared in some of the prior surveys.

[b]The item covers many domains, has a 6-month reference period for disability, and has severity gradations.

[c]The item is modified from a small-scale pilot study conducted by Charles Cannell and colleagues for NCHS in 1975. We simplify the lead question, add the reference period for disability, and use different severity gradations.

[d]The supplement accompanies the 1994–95 NHIS. It has two phases. Phase 1 occurs at the same time as the NHIS Core and has disability questions about all household members; Phase 2 is conducted several months later for persons who screen in from Phase 1 as having disabilities. Here, we use a Phase 1 item about emotional/cognitive problems, adding a parallel one about physical problems.

[e]BLSA is a lifelong study of adults conducted by the Gerontology Research Center, National Institute on Aging. Participants have medical exams and questionnaires every 2 years. The follow-up was conducted in 1989 on dropouts (people who had not returned for the biennial exam). We modify the item by adding the descriptor "in your home, work and leisure activities" and including the category "very good" (to match the five response categories—"excellent," "very good," "good," "fair," "poor"—now used for self-rated health items in the United States). This question must be asked in the context of health/functioning; without that context, "functioning" is vague.

Evaluating Alternative Ways to Measure Change in Functioning: Subjective Versus Objective Indicators in the Longitudinal Study Of Aging

Mary Beth Ofstedal, Harold Lentzner, and Julie Dawson Weeks

An individual's perception of his or her health has been shown to be an important predictor of subsequent decline in health status and, ultimately, death. Numerous studies have documented significant effects of subjective health ratings on mortality among both elderly and nonelderly individuals, independent of the effects of objective indicators of health (Singer, Garfinkel, Cohen, & Srole, 1976; Mossey & Shapiro, 1982; Kaplan & Camacho, 1983; Kaplan, Barell, & Lusky, 1988; Idler & Kasl, 1991; Wolinsky & Johnson, 1992). Given the importance of self-rated health measured at a single point in time, one might expect that an individual's perception of how his or her health has changed over a given period would also be a valuable indicator of health status transitions. As with self-rated health, it is plausible that a subject's perception of change in status may be a better predictor of future health and health transitions than a measure of change that is constructed by comparing the subject's status on a particular indicator across two or more points in time.

Despite the provocative nature of this hypothesis, little research has been conducted to compare the two types of measures with regard to their association with later health outcomes or even to evaluate the congruence between them. One related study by Singer (1977) examined associations between objective measures of current functioning and change in functioning and subjective evaluations of change in functioning among Parkinson's patients, with the purpose of assessing the validity of substituting subjective indicators of change for more objective estimates derived from a before-after design as predictors of some other outcome. The most significant finding was that current functioning was at least as powerful a predictor of subjective change as was actual change in functioning, and Singer cautions the reader against assuming that perceived change in health status provides an appropriate substitute for more objectively derived measures.

A more recent study compared alternative measures of change in post-traumatic stress disorder (PTSD) among war veterans (Spiro, Shalev, Solomon, & Kotler, 1989). As part of the evaluation, a battery of psychological tests was administered to participants at different times both before and after the program. Upon completion of the program, participants were asked to complete a questionnaire that included a retrospective assessment of change in daily functioning and symptomatology. The major finding of interest was the apparent contradiction between the retrospective change assessment and changes in scores on the psychological tests. In particular, retrospective self-assessments tended to reflect more favorably on the effectiveness of the program as far as reducing symptoms of PTSD and improving social functioning, compared with results from the psychological tests.

Both the study by Singer and that by Spiro et al. found some disparity between subjective and objective indicators of change in health status, suggesting that the two types of indicators may be measuring somewhat different dimensions of health status. The current study extends previous research by drawing on data provided in the Longitudinal Study of Aging (LSOA) to examine the level of agreement between subjective and objective indicators of change in physical functioning among the elderly and evaluate the relative power of each type of indicator as a predictor of subsequent health outcomes. Details of the survey design and content and an outline of the analyses presented in the paper are provided in the next section.

Data and Methods

The LSOA is a multiwave panel study that was conducted by the National Center for Health Statistics in collaboration with the National Institute on Aging. The survey is based on the Supplement on Aging to the 1984 National Health Interview Survey (NHIS) and follows a cohort of 7,527 noninstitutionalized persons aged 70 years and older in 1984. Data collection for the baseline survey was conducted via personal interviews and by telephone for the three follow-up waves in 1986, 1988, and 1990. Proxy respondents were used when subjects were unable to respond for themselves. In addition to the interview data, the LSOA survey records were matched to the National Death Index (NDI), multiple cause of death, and Medicare files. These linkages provide data on fact, date, and cause of death, as well as utilization of health care services.

A major focus of the LSOA was to examine transitions in physical functioning among the elderly. To this end, a

Mary Beth Ofstedal, Harold Lentzner, and Julie Dawson Weeks are in the Office of Analysis, Epidemiology, and Health Promotion at the National Center for Health Statistics, Hyattsville, Maryland.

series of questions on activities of daily living (ADLs) and instrumental activities of daily living (IADLs) was included in each wave of the survey. These questions asked respondents if they had any difficulty performing a variety of activities and, if so, how much difficulty they experienced ("some," "a lot," or "unable to perform activity"). By comparing these measures for a given activity or set of activities across survey waves, we can assess whether a subject's functioning improved, declined, or remained the same from one wave to the next, as indicated by changes in self-reports of functional status.

In addition to the repeated ADL and IADL measures, each of the follow-up interviews contained questions concerning respondents' perceptions of change that had occurred in their ability to perform a given activity since the time of the last interview. Specifically, respondents were asked whether they were experiencing "more" difficulty, the "same amount," or "less" difficulty performing each activity compared to what they were experiencing at the time of the previous interview. Respondents were asked their perceived change in status only if they indicated that they had some level of difficulty at the current interview. Respondents reporting having no difficulty with a given activity at the current interview were not asked whether their condition had changed. Thus, any analysis that incorporates the measures of perceived change in functioning is necessarily restricted to the sample of subjects who reported having at least some difficulty at one or more of the follow-up waves. In keeping with Singer's (1977) terminology, we refer to respondents' perceptions or evaluations of change in functioning across survey waves as subjective change and differences in ADL and IADL scores across survey waves as objective change.

This paper has two primary objectives: (a) to examine the level of discordance between subjective and objective indicators of change in functional status and (b) to determine which type of change indicator is a better predictor of later health outcomes, such as institutionalization and death. To examine the degree of discordance, we present results from a series of cross-tabulations that compare subjective change with the change implied by cross-wave comparisons of the ADL measures, or objective change. We also explore whether certain types of activities (e.g., basic functional activities, such as eating or dressing) elicit greater or lesser agreement than others. To address the second question, we conducted two sets of logistic regression analyses to examine the relative effects of subjective and objective change in functional ability on subsequent mortality and institutionalization. In this paper we focus on results pertaining to the six ADL measures obtained in the survey; however, parallel analyses will eventually be conducted to incorporate data on IADLs. With the exception of the data in Tables 1 and 2, the data presented here were weighted to represent national population totals, and estimates of standard errors were calculated using the SUDAAN software package to take account of the complex sample design of the NHIS.

Table 1. ADL scores for degree of difficulty bathing across survey waves: LSOA, 1984 and 1986

	1986			
1984	None	Some	A lot	Unable
None	3,124 (76.9)	242 (6.0)	75 (1.8)	191 (4.7)
Some	79 (1.9)	46 (1.1)	12 (0.3)	50 (1.2)
A lot	29 (0.7)	25 (0.6)	9 (0.2)	41 (1.0)
Unable	22 (0.5)	15 (0.4)	7 (0.2)	95 (2.3)

Table 2. Objective and subjective indicators of change in ability to bathe, 1984–1986

	Self-reported change in ability to bathe		
Change in ADL score for bathing	More difficulty	No change	Less difficulty
Increased difficulty	469 (58.6)	125 (15.6)	11 (1.4)
No change	99 (12.4)	48 (6.0)	2 (0.2)
Decreased difficulty	29 (3.6)	15 (1.9)	2 (0.2)

Defining Objective Change in Functional Status

The first two tables simply provide illustrative cross-tabulations of the measures used in this study for a single ADL, bathing. Table 1 shows a tabulation of the level of difficulty with bathing as reported in 1984 by the level reported in 1986. The first figure in each cell identifies the number of cases in the cell, and the figure in parentheses represents the cell percentage. Focusing first on the cell in the upper left-hand corner, the data show that 3,124 subjects reported having no difficulty bathing in 1984 and again in 1986, representing 77% of all subjects with nonmissing data on this item at both survey points (i.e., 3,124 out of 4,062). The cells on the diagonal identify subjects who reported the same level of difficulty in 1984 and 1986, those above the diagonal identify subjects whose level of functioning declined by one or more levels between the two interviews, and those below the diagonal identify subjects whose functioning improved between the two time points.

The main purpose for showing this table is to identify the group of subjects on which subsequent analyses are based. As noted earlier, questions on perceived change in status during the interval were only asked of subjects who reported some level of difficulty in performing the activity in question at the end of the interval. Thus, having reported no difficulty in 1986, all cases in the first column lack data on subjective change in ability to bathe between the 1984 and 1986 interviews. This includes the 3,124 subjects who reported no difficulty at both waves, as well as 130 subjects who reported having some or more difficulty bathing in 1984 and reported having no difficulty in 1986. As a result, it is only those subjects in columns 2 through 4 who

reported having some or more difficulty bathing in 1986 (n = 808) who are eligible for inclusion in the comparisons of subjective and objective change for bathing between 1984 and 1986. Although we don't present the breakdowns for all ADLs here, numbers of cases on which comparisons are based are provided in subsequent tables, and from these it is possible to derive the proportion of subjects who reported no difficulty versus some or more difficulty with a given activity at the end of the interval in question.

Defining Concordance and Discordance

The second table provides a tabulation of differences between the 1984 and 1986 ADL scores for bathing (i.e., the indicator of objective change), shown in the far left column, and subjects' perceptions of change in ability to bathe (across the top row) between the 1984 and 1986 interviews. (Note that this tabulation is based on the group of 808 subjects who were asked about perceived change in bathing, minus 8 subjects for whom data were missing on the variable for perceived change.) The numbers on the diagonal identify subjects whose responses on the ADL questions for bathing in 1984 and 1986 were concordant with their 1986 report of perceived change. For example, focusing on the middle cell, there were 48 subjects whose responses on the ADL questions implied no change in difficulty and who reported that their ability to bathe had not changed since the previous interview. Likewise, subjects in the upper left cell also gave concordant responses on the subjective and objective measures—with both measures suggesting a decline or deterioration in ability to bathe between 1984 and 1986. Finally, there were two subjects who perceived themselves as having less difficulty and whose ADL scores implied an improvement between 1984 and 1986. Taken together, a total of 519 out of 800 subjects (65%) gave responses to the two indicators of change in bathing ability between 1984 and 1986 that were concordant.

Cells off the diagonal correspond with what we have defined as discordant reports for subjective versus objective change. Subjects in these cells can be grouped into three main classes: (a) those who perceived themselves as having either more or less difficulty in 1986 but whose ADL scores showed no change in functioning, (b) those who perceived themselves as having experienced no change in ability but whose ADL scores implied either a decline or an improvement in functioning, and (c) those for whom the objective and subjective indicators implied change in opposite directions.

Based on this classification, the most common form of discordance in subjective and objective reports of change corresponds with subjects' reporting no change on the subjective indicator but showing either a decline or an improvement on the objective indicator. As shown in Table 2, a total of 125 subjects reported no change in difficulty for bathing but showed a decline based on ADL scores, whereas 15 subjects reported no change and showed an improvement based on ADL scores. Also common is the situation wherein subjects showed no change based on ADL scores but reported having more or less difficulty on the subjective indicator (99 subjects reported having more difficulty on the subjective indicator for bathing but showed no change in ADL scores, and 2 subjects reported having less difficulty but showed no change in ADL scores). Although the overall levels of discordance vary somewhat across different ADLs, as will be shown in the next table, these general patterns in the composition of discordance are quite consistent across activities.

Research Findings

Patterns of Discordance in ADLs

Based on the definition provided above, Table 3 identifies the percentage of subjects with discordant responses on the subjective and objective indicators of change. Percentages are shown separately for each of the six ADLs included in the survey and are further broken out according to survey interval. The number of cases on which each distribution is based is provided in parentheses.

In general, the level of discordance is fairly high, ranging from 22% for using the toilet in the 1984–86 interval to 50% for walking in the 1988–90 interval. Discordance tends to be higher for walking and bathing and lower for eating and using the toilet, although there is some variation in these rankings across intervals. These rankings tend to parallel difficulty rankings across activities, in that the proportions of subjects who report having difficulty walking and bathing are high relative to other activities, and the proportions reporting difficulty eating and using the toilet are quite low.

Another interesting finding pertains to the pattern across survey intervals, for which the percentage of subjects with discordant responses tends to increase at each subsequent interval. For example, the percentage of subjects with discordant responses for walking increased from 39% in 1984–86 to 50% in 1988–90. With the exception of dressing and eating, which changed very little across intervals, all other activities showed similar increases in the percentage of discordant responses.

Table 3. Percentage of subjects with discordant responses in each interval (base n in parentheses)

ADL measure	1984–86		1986–88		1988–90	
Walking	39.0	(1,348)	46.3	(1,108)	50.3	(1,324)
Bathing	36.2	(800)	37.3	(669)	41.4	(789)
Dressing	31.5	(508)	33.6	(460)	34.0	(548)
Eating	30.7	(227)	32.4	(203)	28.9	(242)
Transferring	30.6	(774)	40.0	(652)	43.2	(769)
Using toilet	22.2	(399)	24.7	(356)	35.6	(423)

Although the tabulations are not presented here, it should be noted that the level of discordance is highly associated with the degree of difficulty for a specific ADL both at baseline and follow-up. In particular, the higher the level of difficulty at either point, the more likely the respondent was to give discordant responses on the objective and subjective measures. This was most apparent among subjects who were unable to perform the activity at baseline (and were therefore unable to decline further on the objective measure) but who perceived themselves as experiencing more difficulty performing the activity two years later. This ceiling effect may account, at least in part, for both the variation in the level of discordance across activities (because subjects reported higher levels of difficulty bathing and walking compared with eating and using the toilet) and the increase in the level of discordance across intervals, and we intend to explore this problem further in future research.

Predicting Future Health Outcomes

A second major objective of the study was to evaluate the relative utility of the objective and subjective change indicators as predictors of subsequent health outcomes. In order to address this issue, we estimated several sets of logistic regression models to evaluate the separate and joint effects of subjective and objective change on subsequent mortality and institutionalization, controlling for demographic and baseline health status characteristics. In the analysis presented here, we focus on change in functioning that occurred during the 1984–86 interval and mortality and institutionalization as measured in 1988.

The objective and subjective indicators of change in functioning used in this analysis are composite measures based on the full range of ADLs. First, for the subjective change indicator, we started with a three-category variable for each activity, with the categories corresponding with subjects' reports of less difficulty (coded –1), no change (coded 0), and more difficulty (coded +1). The composite measure used in the analysis was then constructed simply by summing these activity specific variables across all activities. Likewise, for objective change, we started with the same three-category measure for each activity, with categories reflecting either improvement (coded –1), no change (coded 0), or decline in functioning (coded +1) based on differences in ADL scores between the 1984 and 1986 interviews. Again, these measures were simply summed to provide a composite measure of objective change. The resulting composite indicators of subjective and objective change reflect, in some sense, the net improvement or decline in functioning across all activities. The indicators range from –6 to +6 and are represented as continuous variables in the analysis. Other independent variables included in the model as controls were age, sex, baseline measures of self-rated health, number of IADL difficulties, and number of ADL difficulties.

The first set of models evaluates the effects of subjective and objective change on subsequent mortality and focuses on the subsample of respondents who were interviewed in 1984 and 1986 and for whom information on all of the independent variables was available. The dependent variable in these models is a dichotomous measure indicating whether the subject died at any point between the time of the 1986 interview and the 1988 follow-up contact (deceased = 1; not deceased = 0). The second set of models evaluates the effects of subjective and objective change on the log odds of institutionalization for subjects who were not deceased in 1988. Here the dependent variable is a dichotomous measure indicating whether the subject was living in an institution at the time of the 1988 follow-up contact (institutionalized = 1; noninstitutionalized = 0).

For each outcome, three models were estimated. Model 1 included all of the control variables plus the objective change indicator, Model 2 included the control variables plus the subjective change indicator, and Model 3 included both the objective and subjective indicators, as well as the controls. By comparing the magnitude of the coefficients associated with the subjective and objective change measures in addition to the model chi-square statistics across models we can gain some insight into the relative power of the subjective and objective change indicators for predicting death and institutionalization.

Table 4 presents the odds ratios (and 95% confidence intervals) associated with the subjective and objective indicators for each model, as well as the model chi-square

Table 4. Odds ratios and chi-square statistics for the effects of subjective and objective change on subsequent mortality and institutionalization

Outcome measure	Model 1	Model 2	Model 3
Deceased in 1988			
Objective change, OR (95% CI)	1.36 (1.26, 1.47)	—	1.23 (1.10, 1.39)
Subjective change, OR (95% CI)	—	1.30 (1.22, 1.38)	1.13 (1.02, 1.24)
Model X^2 (df)	66.7 (1)	57.8 (1)	73.4 (2)
Institutionalized in 1988			
Objective change, OR (95% CI)	1.52 (1.30, 1.78)	—	1.42 (1.17, 1.73)
Subjective change, OR (95% CI)	—	1.38 (1.15, 1.61)	1.09 (0.94, 1.28)
Model X^2 (df)	41.2 (1)	28.3 (1)	42.4 (2)

statistics for each of the two outcomes of interest, mortality and institutionalization.

Focusing first on Model 1 for mortality, the results suggest that the objective indicator of change in functioning is significantly associated with later mortality (OR = 1.36, CI = 1.26, 1.47). An odds ratio greater than one corresponds with an increased risk of mortality and suggests that the greater the net decline in a subject's functioning across ADLs (or the less the net improvement), the greater the likelihood of dying within a 2-year period. Results of Model 2 suggest that controlling for age, sex, and baseline health status, one's subjective evaluation of change in functioning is also significantly associated with mortality, in the same direction (and of roughly the same magnitude) as was observed for objective change. A comparison of the chi-square statistics for Models 1 and 2 suggests that the model that incorporates objective change provides a better fit to the data. Turning to Model 3, the odds ratios for both objective and subjective change are somewhat reduced when both indicators are included in the model; however, the effects of each remain statistically significant. Hence, it appears that in the case of mortality, a subject's perception of the change that has occurred in his or her functioning is a significant predictor of mortality above and beyond any actual change that has taken place.

With respect to institutionalization, the results are quite similar in that both objective and subjective indicators of change are statistically significant predictors of institutionalization in the absence of one another (Models 1 and 2); however, when both indicators are incorporated into the model (as in Model 3), the effect of subjective change is reduced substantially and is no longer statistically significant. This finding suggests that actual change in functioning (as measured by repeated ADL measurements) is the critical factor influencing institutionalization, and once that is taken into account, one's perception of change does not provide any additional information.

Conclusions

One of the major purposes of this study was to examine the degree of agreement between subjects' perceptions of change in functional ability and changes in their reports of ability over time and, to the extent that there is some disagreement, to identify factors that may be associated with inconsistent reporting. In regard to this objective, we found evidence of a fair level of discordance between the two indicators of change. The most common pattern of discordance occurred when respondents reported no change on one of the indicators and some level of decline in functioning on the other. The analyses conducted to date have provided little insight into the characteristics of persons giving discordant answers, however, and it may be necessary to take a more case-oriented approach in addressing this objective in future research. In the next stage of analysis, we plan to examine the consistency of concordance within subjects over time, as well as identify the direction of discordance (i.e., whether perceptions tend to be more optimistic than differences in ADL scores or vice versa) in analyses that examine the factors associated with discordant reporting.

A second objective was to compare the two types of measures with respect to their utility for forecasting future health outcomes. Results of preliminary analysis suggest that objective indicators of change in functioning are more strongly associated with both institutionalization and death than are subjective indicators of change. However, with respect to mortality, the subjective indicator remains significant in the full model. This finding suggests that at least for a major outcome like death, the two indicators may be tapping into different aspects of functional decline, and the finding supports the idea that individuals may have some subjective awareness regarding impending death that is not reflected in more objective measures of health status or change in health status. In future analysis, we plan to expand the model to include observations from all waves of data collection, as well as mortality data from the NDI for the latest date available (currently 1992).

A major limitation of the study is the sample restriction imposed by the question structure, which required a subject to report at least some difficulty on the activity in question in order to be asked the subjective question pertaining to change in ability. As a result, the analysis sample is highly selective of elderly individuals who have some level of functional impairment, and the results found in this study may not be generalizable to elderly persons with no impairment or lower levels of impairment. Proposed plans for a second LSOA, to be conducted during the 1990s, may allow us to correct this oversight by including subjective questions for all respondents on follow-up waves. Should this study go forward, availability of panel data from two different cohorts of elderly persons would open up a number of intriguing research possibilities.

References

Idler, E. L., & Kasl, S. (1991). Health perceptions and survival: Do global evaluations of health status really predict mortality? Journal of Gerontology, 46, S55–S65.

Kaplan, G., Barell, V., & Lusky A. (1988). Subjective state of health and survival in elderly adults. Journal of Gerontology, 43, S114–S120.

Kaplan, G. A., & Camacho, T. (1983). Perceived health and mortality: A nine-year follow-up of the Human Population Laboratory Cohort. American Journal of Epidemiology, 117, 292–304.

Mossey, J. M., & Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. American Journal of Public Health, 72, 800–808.

Singer, E. (1977). Subjective evaluations as indicators of change. Journal of Health and Social Behavior, 18, 84–90.

Singer, E., Garfinkel, R., Cohen, S. M., & Srole, L. (1976). Mortality and mental health: Evidence from the Midtown Manhattan Restudy. Social Science and Medicine, 10, 517–525.

Spiro, S. E., Shalev, A., Solomon, Z., & Kotler, M. (1989). Self-reported change versus changed self-report: Contradictory findings of an evaluation of a treatment program for war veterans suffering from post traumatic stress disorder. Evaluation Review, 13, 533–549.

Wolinsky, F. D., & Johnson, R. J. (1992). Perceived health status and mortality among older men and women. Journal of Gerontology, 47(6), S304–S312.

The Domains of Primary Care and Health Outcomes

Susan A. Flocke, Kurt C. Stange, and Stephen J. Zyzanski

Introduction

Improved delivery of primary care health services is increasingly seen as critical to efforts to improve health care access and quality while controlling costs. Primary care has been shown to be parsimonious in resource utilization (Greenfield et al., 1992) and generally effective in providing quality care (Franks, Clancy, & Nutting, 1992). However, the specific aspects of primary care associated with the delivery of important services have not been elucidated. Defining and measuring the specific domains of primary care is critical to efforts to determine which components affect important outcomes.

Building on its seminal work in 1978, the Institute of Medicine (IOM) has recently released an interim report defining primary care (1994). Driven by the trend toward greater complexity of health care delivery, the greater interdependence of health professionals (IOM, 1994), and an interest in increasing the proportion of practicing primary care clinicians (U.S. Department of Health and Human Services, 1991; "President's Health Care," 1993), the report describes a framework within which to view primary care in the current health care delivery realm.

The IOM's 1994 report defines primary care as "the provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community" (p. 1). The specific components and their definitions have not changed from the components proposed in the 1978 report: comprehensiveness, coordination, continuity, accessibility, and accountability. However, the interpretation has been broadened to include aspects of the family, the community, and the perspective of a health care system changing toward integrated care.

The advancement of research regarding the components of primary care has been limited by the lack of standardization of terminology in the field of primary care (Starfield, 1990). The IOM's recently proposed definition of primary care and its specific components is meant to assist researchers, practitioners, and policy makers in their quest for advancing the delivery of primary care, but the broad concepts need translation into measurable qualities (Starfield, 1990).

Several investigators have attempted to measure individual components of primary care (e.g., continuity and accessibility), and a few have attempted to measure the components comprehensively (Safran, Tarlov, & Rogers, 1994; Smith & Buesching, 1986). Previous attempts to measure the components of primary care have been limited by poor operationalization of the component for measurement or methods that overestimate (such as physician self-reports) or underestimate (such as patient encounter logs) the delivery of primary care.

The ability to evaluate the outcomes of various aspects of primary care is limited by a lack of well-validated comprehensive measures of the components of primary care. Using the IOM 1994 interim report and complementary work by Starfield (1990; 1992) as a starting point, an instrument was developed to measure several components of primary care from the perspective of the patient. The purpose of this study was to evaluate this instrument and to evaluate the association of the components of primary care with the outcomes of patient satisfaction and preventive service delivery.

Methods

Instrument Development

Instrument development involved several sequential steps, including developing the theoretical basis of what to measure; operationalizing the concepts; writing multiple items per concept; refining the items; negatively wording several of the items to avoid halo effects; choosing a response format; assessing the content validity of the items; assessing the reading level of the items and the acceptability of the items to the intended audience; and, finally, pilot testing and revising the instrument (Allen & Yen, 1979; DeVellis, 1991). In choosing which dimensions to measure, the dimensions proposed by the two IOM reports (1978; 1994) were evaluated, as were some suggested components by Starfield (1992). The dimensions selected for inclusion in the new instrument include comprehensiveness of care,

The authors are at Case Western Reserve University, Cleveland, Ohio. Susan A. Flocke is a Research Analyst in the Department of Family Medicine, Kurt C. Stange is an Assistant Professor of Family Medicine and is in the Departments of Epidemiology and Biostatistics and Sociology, and Stephen J. Zyzanski is a Professor and Director of Research in the Department of Family Medicine.

continuity, longitudinality, coordination, first contact, and two additional components: physician accumulated knowledge about the patient and interpersonal communication. The patient's belief in the importance of continuity of care and coordination were other concepts considered for inclusion.

The instrument is oriented toward the patient's perception of the patient-physician interaction. The patient may be in the best position to evaluate these dimensions and for some of the dimensions (e.g., visits to other physicians) is the sole source of the information. The operationalization of each of the dimensions follows. Comprehensiveness of care has traditionally been some measure of the range of services the physician provides and is typically assessed by chart or physician report. This is measured from the perspective of the patient and their belief in the physician's ability to address the majority of health problems they are likely to encounter. Three items were selected for inclusion in the instrument for this dimension.

The second dimension, continuity of care, has been previously measured with mathematical formulas that use systemwide data that are generally not available in the United States. Data are available for measuring continuity of care by assessing the usual provider of care index (UPC). UPC is a calculated score of the number of visits to the usual provider divided by the total number of physician visits in the past year. With the traditional measures of continuity (including the UPC), a low level of continuity could be due to a system or practice environment factor or a patient factor (e.g., continuity not being a priority). Therefore, in addition to the UPC measure, the patient's indication of the importance of continuity is also assessed. Three items were selected for inclusion in the instrument for this dimension.

The third dimension, longitudinality, is strictly the length of time the individual has been a patient of the observed physician. Coordination of care is the fourth dimension and has not been well measured in the primary care setting. Previous measures have included use of a single medical record for a patient within a health care system and knowledge of care by other physicians through referrals. Items were written to assess the patient's perception of the physician's active use of information from visits to specialists and coordination of care via following up on problems through subsequent visits or phone calls. Four items were selected for inclusion in the instrument for this dimension. A single item on the preference for one doctor to coordinate the patient's care was also included.

The fifth dimension, first contact, refers to the primary care provider's role as an individual's entrée into the health care system. This is a difficult concept to measure, as it very likely depends on the situational factors for which the individual is seeking care. A single generic item was included to represent this dimension.

The sixth dimension is depth and comprehensiveness of physician knowledge about the patient's medical history, family history, and medical needs. These items measure the physician's knowledge of the patient through the patient's eyes, that is, how well the physician knows the patient's needs, values, and relevant medical history. Four items were selected for this dimension.

The final dimension is interpersonal communication. This dimension overlaps with what some have called "interpersonal accountability" but focuses on the communication between the patient and the physician. The patient evaluates how well the physician listens and explains. Three items were selected to be included in the instrument to represent this dimension.

The goals of developing this instrument were to create a valid and brief measure. The instrument was designed to be completed by patients immediately following their visit. Several individuals provided input regarding the content validity of the items, the clarity of the wording, and the appropriateness of the response format. The response format for the majority of the items is a 5-point Likert-type scale anchored by "strongly agree" and "strongly disagree." A revised draft was pilot tested with 40 patients, and items were slightly modified based on their comments. The instrument was called the Components of Primary Care Instrument (CPCI).

Data Collection

This study was a cross-sectional design that enrolled members of the Northeast Ohio Research Association of Practicing Physicians—a community-based research network of 134 family physicians—and their patients. Research nurses collected data from consecutive patients visiting participating physicians' offices during a typical scheduled practice day. The content of the medical encounter was assessed with direct observation, medical records review, billing data, a physician questionnaire, and a patient exit questionnaire. Patients were asked to complete the exit questionnaire in the office or as soon as possible after the visit. The majority of data reported in this paper are derived from the patient exit questionnaire; measures included demographics, items about the type of payment system, a measure of patient satisfaction, a checklist of preventive service delivery, and the new CPCI.

Outcome Variables

Two main outcome variables were measured. Patient satisfaction with the visit was assessed using the nine-item Medical Outcomes Study (MOS) Visit Rating Form (Rubin, Gandek, Rogers, Kosinski, McHorney, & Ware, 1993). The rating form is a reliable self-administered survey (Ware & Hayes, 1988) that was completed by the patient as part of the patient exit questionnaire.

The second outcome variable is the delivery of preventive services to eligible patients. Patient eligibility for specific services was determined using an age- and gender-specific algorithm based on the U.S. Preventive Services Task Force (USPSTF) guideline (1989). The calculation of this outcome variable used data from the patient exit ques-

tionnaire, chart review, and the direct observation checklist. Patients who had evidence on either their chart review or patient exit questionnaire of having received a particular preventive service that is recommended for their age and gender were counted as having had that service at baseline. Patients who had no evidence of a recommended service were considered eligible for that service during the observed encounter. Patients were designated as up-to-date on a particular service (i.e., individuals had it at baseline or were eligible and service delivery was observed) or not up-to-date. Summary scores for recommended screening services and counseling services were calculated for each individual.

Analyses

For the analyses in this paper, a random sample of 500 respondents was selected, which provided ample power for stable estimates for the proposed analyses. Descriptive statistics for the study sample demographic variables are reported. An item analysis of the CPCI instrument was conducted to evaluate the items. Missing data for these items were assessed and recoded to the mean for those items for which missing data accounted for less than 5%.

An exploratory factor analysis was employed to cluster the items. A principal components solution with a varimax rotation was used. Only statistically significant (eigenvalues > 1) and internally consistent factors were interpreted. A second independent random sample of 500 respondents was selected to test replication of the factor solution. Scale scores were computed and the association among the factor scale scores assessed by correlation. The internal consistency of each of the scale scores was assessed by Cronbach's alpha.

Finally, the associations of each of the scale scores with the characteristics of the patient were assessed by multivariate analysis of variance. The associations with the delivery of preventive services and patient satisfaction were assessed by correlation and partial correlation when potential confounding variables were included in the model. All tests were evaluated at $p \leq .01$.

Results

Of the 2,213 patients who agreed to participate in the study to date, 1,431 completed the patient exit portion of the protocol, which represents a 65% response rate. Those individuals with excessive missing data on the CPCI (> 5 items) were not eligible for selection for the random sample. For the random sample of 500 respondents selected for analyses, the patient characteristics are as follows: The median patient age was 45, 65% are female, and 94% are white. Only 16% of those 18 years of age or older reported having less than a high school education, and 71% reported they were currently married. Twenty-five percent of the sample reported they were part of a prepaid system; 45% fee-for-service; and 30% Medicare, Medicaid, other, or uninsured.

The factor analysis resulted in four statistically significant factors (see Table 1). The items are clustered by their primary loading, which is presented in bold-faced type. Only loadings $\geq .30$ are displayed. Based on their content, the four factors were named continuity belief, interpersonal communication, in-depth knowledge of patient, and coordination. Although UPC shows an association with the continuity belief factor, it did not contribute to the internal consistency of that cluster of items, and we chose to score it separately. These four factors and the UPC measure represent the original components fairly well. The internal consistency of the four factors is reported at the bottom of the table and is acceptable for the small number of items per factor. Factor scale scores were created by simply adding the items contributing to the factor and dividing by the number of items summed so that each scale score has a maximum of five. For example, responses to the four items with a primary loading on the coordination factor were summed and divided by 4. The mean and standard deviation for each of the scale scores are also reported at the bottom of the table. It is apparent that the scores are quite skewed, with the majority of patients having high scores. A second independent random sample of 500 respondents was used to replicate the factor analysis. The components measured by single items (first contact and comprehensiveness) shifted primary loading to different factors, but the initial solution's four clusters of items—continuity belief, interpersonal communication, in-depth knowledge of patient, and coordination— remained clustered together.

The overlap of factor loadings indicates that the factors are likely to be associated to some degree, and indeed they are correlated. Table 2 displays the correlation matrix of the scale scores. UPC is associated to a lesser degree with the other factors, but all correlations are at $p \leq .001$.

The association of the scale scores with the patient characteristics was tested using multivariate analysis of variance and revealed that neither any of the scale scores nor the UPC measure is associated with gender, marital status, education, or type of payment system (prepaid vs. fee-for-service). Age, however, is significantly associated with the delivery of primary care scale scores but not UPC. Age was categorized into four groups (0–17, 18–39, 40–64, and 65 or older) and demonstrated an increasing linear association with each of the scale scores ($p < .001$). Therefore, subsequent comparisons involving the scale scores will be controlled for the affects of age.

Patient satisfaction as measured by the MOS nine-item Visit Rating Form is significantly correlated (assessed by partial correlation controlling for age) with each of the four scale scores (see Table 3). UPC, the continuity of care item, was associated to a lesser degree. The correlations are moderate, and there are likely to be ceiling effects, in that both the factor scores and the satisfaction scores are skewed.

The delivery of preventive care outcome variables was assessed next. We chose to exclude two groups from this

Table 1. Factor analysis solution

	Factor			
	1	2	3	4
Content				
Continuity belief	.68	.35	—	—
Coordination	.65	—	—	—
Continuity belief	.65	—	—	—
Comprehensiveness of care	.59	—	—	.34
Continuity belief	.58	—	—	—
First contact	.53	—	.30	—
UPC	.44	—	—	—
Interpersonal communication	—	.77	—	—
Interpersonal communication	—	.75	—	—
Interpersonal communication	—	.63	—	—
Interpersonal communication	.32	.57	.31	—
Longitudinality	—	—	.80	—
Been through a lot	—	—	.67	—
Knowledge of patient	—	—	.67	—
Knowledge of patient	—	.44	.55	—
Knowledge of patient	—	.37	.49	.42
Coordination	—	—	—	.84
Coordination	—	—	—	.79
Coordination	—	—	—	.66
Coordination	—	—	.39	.55
Eigenvalues	6.77	1.52	1.41	1.24
Cronbach's α	.73	.69	.74	.78
Factor score M	4.50	4.40	3.50	4.00
SD	.60	.70	.90	.90

Table 2. Correlation among the empirical factors

	Interpersonal communication	Knowledge of patient	Coordination	UPC
Continuity belief	.46	.48	.54	.32
Interpersonal communication	1.00	.43	.52	.17
Knowledge of patient	—	1.00	.52	.22
Coordination of care	—	—	1.00	.23

Table 3. Correlation of patient satisfaction with primary care scale scores, controlling for age

	Patient satisfaction
Continuity belief	.30***
Interpersonal communication	.41***
Knowledge of patient	.23***
Coordination of care	.42***
UPC	.13*

*p = .005; all others p ≤ .001.

analysis: (a) individuals less than 18 years of age and (b) individuals for whom this was their first visit. The exclusions were justified by the following reasons: A limited number of the preventive services chosen for this study apply to the first group, and the second group provides inadequate data for both the delivery of primary care measure and the delivery of preventive services. One hundred and eighteen individuals were excluded by these criteria.

A total of 28 services recommended by the USPSTF report (1989) were divided into two composite scores: screening services (e.g., breast exam) and counseling services

(e.g., on sodium levels in diet). Overall, the scores ranged from 0 (up-to-date on none of the recommended services) to 1 (up-to-date on all of the services for which the patient was eligible). Patients were much more likely to be up-to-date on all of the recommended screening services (54%) than on the counseling services (13%).

The association of the delivery of primary care scale scores with these two outcome variables was assessed by partial correlation controlling for age. As the correlation coefficients displayed in Table 4 indicate, there is no association between the delivery of primary care as assessed by the scale scores of the CPCI instrument and the delivery of recommended screening and counseling preventive services. The near-0 correlation coefficients lead us to investigate two possible explanations: (a) nonlinear associations and (b) interaction effects with the age variable. We suspected a threshold effect occurring with the preventive service delivery variables and categorized them into thirds. The actual cut points differed for screening services and counseling services, due to the different distributions. A two-way multivariate analysis of variance was used to investigate the possibility of a nonlinear association and an interaction effect with age. The significance levels (actual p value) of the main effects and interaction term of each nine-cell analysis are reported in Tables 5 and 6. While some terms appear to be significant, when Bonferroni multiple testing corrections are applied to the p values, all are determined to be chance findings.

Table 4. Age-adjusted correlations of delivery of recommended screening services and counseling services with primary care scale scores

Scale score	Screening services	Counseling services
Continuity belief	−.007	.025
Interpersonal communication	.012	.084
Knowledge of patient	.059	.041
Coordination of care	.042	.088
UPC	−.108	−.066

Table 5. Probabilities of effects of associations of primary care scale scores with delivery of preventive screening services and age

	Effects		
Scale score	Screening services	Age	Screening services by age
Continuity belief	.717	.001	.406
Interpersonal communication	.604	.268	.016
Knowledge of patient	.450	.001	.241
Coordination of care	.692	.001	.738
UPC	.200	.871	.633

Table 6. Probabilities of effects of associations of primary care scale scores with delivery of preventive counseling services and age

	Effects		
Scale score	Counseling services	Age	Counseling services by age
Continuity belief	.774	.001	.839
Interpersonal communication	.171	.005	.762
Knowledge of patient	.360	.001	.223
Coordination of care	.667	.001	.826
UPC	.432	.762	.138

Discussion

The new instrument appears to measure five aspects of the delivery of primary care. The scales have good internal consistency and are each associated with a measure of patient satisfaction. The lack of an association of the scales with the delivery of preventive services lead us to several possible conclusions. It may be true that the theoretical basis upon which this association was tested is wrong: Perhaps the components of primary care are not associated with the delivery of preventive services. On the other hand, if the theory is correct, we have failed to adequately measure the dimensions of primary care and/or preventive services, are limited by the lack of variability in this patient population, or have failed to control for confounding variables that mask the effect.

It is possible that we have not adequately measured some of the primary care dimensions, but the core content on four of the dimensions is covered such that the addition of items to each scale is not likely to change the results. The measure of preventive service delivery derived from the evidence-based clinical guideline is the most sophisticated and comprehensive measure of its kind to date, and we have great confidence in the accuracy of this measure.

The skewed primary care scale scores were anticipated, given that this was a community-based cross section of patients, and those who were dissatisfied were likely to seek another primary care physician. Several of the items were very skewed, and thus variability was limited.

Additional potential confounding variables will be investigated once data regarding the physician characteristics and the practice environment become available. Since patients are clustered within physicians, physician characteristics and style may play a very important role in both measures. The next set of analyses will address the nested design. Further directions in analyzing this data may involve investigating specific preventive services or different subgroups of patients. The value of the new instrument

will be determined as it is tested in additional settings and with other indicators of quality of care and health outcomes.

A word on generalizability: The respondents in this sample represent a cross section of the patient pool of the physicians in the research network, and the generalizability of these findings to other geographical regions or other primary care disciplines is limited.

Present trends toward managed care and the increased role of primary care make it necessary to evaluate the quality of the delivery of primary care. Further development of measures of primary care and examination of the association of these measures with important outcomes will be critical to guide efforts to improve the organization and quality of health care.

References

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Wadsworth.

DeVellis, R. (1991). Scale development: Theory and applications. Newbury Park, NJ: Sage.

Franks, P., Clancy, C., & Nutting, P. (1992). Gatekeeping revisited: Protecting patients from overtreatment. New England Journal of Medicine, 327, 424–429.

Greenfield, S., Nelson, E., Zubkoff, M., Manning, W., Rogers, W., Kravitz, R., Keller, A., Tarlov, A., & Ware, J. (1992). Variations in resource utilization among medical specialties and systems of care: Results from the Medical Outcomes Study. Journal of the American Medical Association, 267, 1624–1630.

Institute of Medicine. (1978). Report of a study: A manpower policy for primary health care. Washington, DC: National Academy of Sciences.

Institute of Medicine. (1994). Defining primary care: An interim report. Washington, DC: National Academy Press.

President's health care reform plan: American Health Security Act of 1993. (1993). Washington, DC: Bureau of National Affairs.

Rubin, H., Gandek, B., Rogers, W., Kosinski, M., McHorney, C., & Ware, J. (1993). Patient's ratings of outpatient visits in different practice settings: Results from the Medical Outcomes Study. Journal of the American Medical Association, 270, 835–840.

Safran, D. G., Tarlov, A., & Rogers, W. H. (1994). Primary care performance in fee-for-service and prepaid health care systems: Results from the Medical Outcomes Study. Journal of the American Medical Association, 271, 1579–1586.

Smith, W. G., & Buesching, D. (1986). Measures of primary medical care and patient characteristics. Journal of Ambulatory Care Management, 9(1), 49–57.

Starfield, B. (1990). Commonalties in primary care research: A view from pediatrics. In Proceedings from AHCPR, primary care research: An agenda for the 90's (DHHS Publication No. PHS 90-3460). Washington, DC: U.S. Government Printing Office.

Starfield, B. (1992). Primary care: Concept, evaluation, and policy. New York: Oxford University Press.

U.S. Department of Health and Human Services. (1991). Healthy people 2000: National health promotion and disease prevention objectives (DHHS Publication No. PHS 91-50212). Washington, DC: U.S. Government Printing Office.

U.S. Preventive Services Task Force. (1989). Guide to clinical preventive services: An assessment of the effectiveness of 169 interventions. Baltimore, MD: William and Wilkins.

Ware, J. E., & Hayes, R. D. (1988). Methods for measuring patient satisfaction with specific medical encounters. Medical Care, 26, 393–402.

Assessing Satisfaction With Health and Health Care: Cognitive and Communicative Processes

Norbert Schwarz, Nancy Mathiowetz, and Robert Belli

Most health-related surveys include some measures of respondents' satisfaction with their health and/or aspects of their health care. How respondents arrive at these satisfaction judgments, however, has received little attention. In the present paper, we draw on an extensive research program that explored the cognitive and communicative processes underlying satisfaction judgments and highlight some implications for the assessment of health and health care satisfaction (see Schwarz & Strack, 1991, for a comprehensive review). To render our discussion relevant to issues of questionnaire construction, we emphasize processes that are affected by question wording and question order at the expense of processes that are outside of the researcher's control, such as respondents' mood at the time of judgment.

Mental Construal and the Emergence of Context Effects

Like any other evaluative judgment, satisfaction judgments require a mental representation of the target of judgment (e.g., "my health") and of a standard of comparison against which the target is evaluated. In constructing these representations, people rarely draw on the multitude of pieces of information that may potentially bear on the task. Rather, they truncate the search process as soon as enough information has come to mind to form the respective representations with sufficient subjective certainty. As a result, representations of the target of judgment as well as the standard are partially context dependent. They include information that is chronically accessible (and lends some stability to the judgments) as well as information that is only temporarily accessible (e.g., because it is brought to mind by the questionnaire). Whether the information that comes to mind results in assimilation or in contrast effects on evaluative judgments depends on how it is used. Information that is included in the representation formed of the target ("my health") results in assimilation effects, whereas information that is used in constructing a standard results in contrast effects (see Schwarz & Bless, 1992, and in press).

A variety of questionnaire variables determine the use of highly accessible information. Below, we first demonstrate how features of the questionnaire provide a frame of reference that respondents use in forming a judgment. Next, we address how the same event in a respondent's life may result in assimilation as well as contrast effects, depending on subtle features of question wording. Finally, we review issues related to general and specific question sequences.

Response Alternatives as Frames of Reference

Frequency Scales: The Case of Physical Symptoms

In many health studies, respondents are asked to report the frequency of physical symptoms by checking the appropriate categories in a set of numeric response alternatives provided to them. For example, the most widely used German symptoms checklist offers five response alternatives, ranging from "never" to "nearly daily" (Fahrenberg, 1975). What is often overlooked is that respondents extract relevant information from the specific values presented on these scales. They assume that the researcher constructs a scale that reflects his or her knowledge about the distribution of the symptoms and that the typical or average frequency is reflected in the middle range of the scale, whereas the extremes of the scale correspond to the extremes of the distribution. These assumptions affect respondents' frequency reports as well as subsequent evaluative judgments.

For example, Schwarz and Scheuring (1992) asked 60 patients of a German psychosomatic clinic to report the frequency of 17 symptoms on one of the following scales:

Low-frequency scale	High-frequency scale
never	twice a month or less
about once a year	once a week
about twice a year	twice a week
twice a month	daily
more than twice a month	several times a day

Reports along these scales can be compared by assessing the percentage of respondents who report having a given

Norbert Schwarz and Robert Belli are with the Institute for Social Research, University of Michigan, Ann Arbor; Nancy Mathiowetz is with the Agency for Health Care Policy and Research, Rockville, Maryland Address correspondence to Norbert Schwarz at the address on page 237.

symptom more than twice a month. As expected, across 17 symptoms, 62% of the respondents reported average frequencies of more than twice a month when presented with the high-frequency scale, whereas only 39% did so when presented with the low-frequency scale. The obtained differences ranged from 75% versus 21% for the ill-defined symptom "reactions to climatic changes" to 50% versus 42% for the better-defined symptom "excessive perspiration." These findings reflect that respondents used the range of the response alternatives as a frame of reference in estimating their own symptom frequency, a finding that has been obtained across a wide range of behaviors (see Schwarz, 1990, for a review).

More germane to our present concern, the range of the response alternatives also influenced respondents' evaluative judgments. The specific effect, however, depends on the specific nature of the question asked. Suppose respondents are asked a comparative question, such as, "How satisfied are you with your health compared to other people your age?" In this case, they can extract relevant comparison information from the distribution suggested by the scale of the frequency question. Checking "twice a month," for example, places a respondent at the low end of the high-frequency scale, suggesting that his or her symptom frequency is below average. Conversely, checking the same frequency on the low-frequency scale places a respondent at the high end of this scale, suggesting that his or her symptom frequency is above average. As a result, respondents reported higher health satisfaction when they had given their frequency reports along the high- (M = 8.3 on an 11-point scale) rather than the low- (M = 7.2) frequency scale, despite the fact that the same respondents had just reported higher absolute symptom frequencies. This finding, which is reliably replicable across a wide range of issues (see Schwarz, 1990), reflects that respondents extract comparison information from their placement on frequency response scales, which affects their assessment of health status.

Suppose, however, that the evaluative question does not require a comparison with others but pertains to how much the symptoms bother one. In this case, respondents are likely to turn to the perceived absolute frequency of their symptoms, which they estimate to be higher when presented with the high- rather than low-frequency scale. Accordingly, 34 respondents to a German follow-up study reported being more bothered by their symptoms when the high-frequency scale induced them to estimate a higher frequency (M = 9.3 on an 11-point scale) than when the low-frequency scale induced them to estimate a lower frequency (M = 6.7), thus reversing the direction of the effect obtained on comparative measures (Schwarz, unpublished data).

In combination, these findings illustrate that numeric frequency scales provide a relevant source of information that respondents use in computing behavioral reports and related judgments. Relying on the response alternatives as a frame of reference, respondents estimate higher behavioral frequencies when presented with a high- rather than low-frequency scale. Moreover, they use the absolute value of their estimate in making noncomparative judgments and the relative placement of their estimate in making comparative judgments, resulting in pronounced differences in satisfaction judgments. Given this impact of numeric frequency alternatives, researchers are well advised to assess frequency reports in an open-response format (see Schwarz & Hippler, 1991, for a more detailed discussion).

Satisfaction With Health Insurance

Whereas the preceding experiments pertain to frequency scales, similar frame-of-reference effects can be expected from other forms of response alternatives. Suppose that respondents are asked which services are covered by their health insurance using one of the following sets of response alternatives:

Response alternative set A	Response alternative set B
hospital care	hospital care
well child care	preventative dental care
emergency room visits	dental—orthodontics
office visits: physicians	well child care
prescribed medicines	emergency room visits
inpatient or outpatient surgery	general check-ups for adults
	prescribed medicines
	mental health services
	vision exams
	inpatient or outpatient surgery

Whereas the first set of response alternatives is restricted to services covered by most health insurance plans, the second set includes services that are less likely to be covered. If respondents use the response alternatives as a frame of reference in evaluating their own insurance plan, we may expect that they would report higher satisfaction when presented with the first rather than the second set. Preliminary data from an experiment with 172 respondents to a National Medical Expenditure Survey pretest confirm these predictions. Respondents exposed to list A reported an overall higher level of satisfaction with their health insurance coverage than respondents exposed to list B (M = 8.1 vs. 7.7, with 10 = "very satisfied").

Recalling a Specific Episode: Conditions of Assimilation and Contrast

Many health surveys include questions about specific episodes of illness or hospitalization. Answering these questions increases the likelihood that the episodes will come to mind when respondents are subsequently asked to evaluate their health. How accessible episodes influence the health satisfaction judgment, however, depends on how respondents use them. For example, respondents may

include an episode of hospitalization in the mental representation of the target "my health," resulting in reports of low health satisfaction (assimilation effect). On the other hand, respondents may use the episode in forming a standard of comparison, relative to which their current health seems very favorable (contrast effect).

How a specific episode is used depends on a number of variables, one of which is the time that has elapsed since the event. Not surprisingly, recent episodes are included in the representation of the target, resulting in assimilation effects. More distant episodes, however, are unlikely to be included in the representation formed of one's current health and serve as a standard of comparison relative to which current health seems rather good (see Strack, Schwarz, & Gschneidinger, 1985, and Schwarz & Strack, 1991). Whereas these influences of the temporal distance are not surprising, similar effects may arise from apparently minor variations in question wording.

For example, Schwarz and Hippler (unpublished data) asked first-year students to report a positive or a negative event that happened to them "two years ago." Subsequently, respondents reported their current life satisfaction. Under this condition, they reported higher current life satisfaction after recalling a positive (M = 8.7 on an 11-point scale) rather than a negative event (M = 7.4). This assimilation effect reflects that respondents included the recalled events in the mental representation of the current period of their lives (Schwarz & Bless, 1992). For other respondents, however, we changed the question wording by introducing a temporal landmark and asked them to report an event "that happened two years ago, that is, before you entered the university." In this case, the pattern reversed, and respondents reported higher life satisfaction after recalling a negative (M = 8.2) rather than positive (M = 6.2) event. Apparently, reminding respondents of a major role transition, namely entering university, induced them to "chunk" their lives into a previous high school episode and a current university episode. As a result, any event that happened 2 years before pertained to a previous episode in their lives. It could therefore not be included in the representation of the current episode, but served as a standard of comparison, resulting in contrast effects.

Thus, the same episode may elicit assimilation as well as contrast effects on satisfaction judgments, depending on its use in constructing mental representations of the target or a relevant standard. Moreover, the respective use of a highly accessible episode may vary as a function of apparently minor variations in question wording. Because calendar dates are usually not well represented in autobiographical memory (see Schwarz, 1990, for a review), researchers are often advised to anchor the time period they are interested in with a salient event (e.g., Loftus & Marburger, 1983). As the present experiment illustrates, however, introducing such landmarks may affect how respondents "chunk" the stream of life into discrete units, thereby strongly affecting subsequent evaluative judgments.

General and Specific Judgments

As a final example of context effects in health care satisfaction measurement, consider the order in which general and specific questions are asked. For example, Bachman and Alcser (1993) asked respondents to report their satisfaction "with the current U.S. health care system" and their satisfaction with their own health insurance plan. Most respondents who had health insurance reported high satisfaction with their own insurance plan, independently of the order in which both questions were presented (77.8% chose "very" or "somewhat" satisfied when the question was asked first and 76.4% when it was asked second). Their reported satisfaction with the U.S. health care system in general, however, showed a pronounced order effect. When this question was asked first, 39.6% of the respondents reported being "very" or "somewhat" satisfied, whereas only 26.4% did so when this question was preceded by the question about their own insurance plan.

These findings reflect an impact of conversational norms, as previously observed in other domains (Schwarz, Strack, & Mai, 1991). Specifically, conversational norms ask speakers to avoid redundancy and to provide new information in response to a question rather than to reiterate information that has already been given (see Schwarz, 1994, for a theoretical discussion). In the present case, respondents who had just reported on their own health insurance plan presumably interpreted the general question as a request for new information, much as if it were worded, "Aside from your own insurance, how satisfied are you with health insurance in the U.S. in general?" As a result, they excluded their own insurance, with which most were satisfied, from consideration when they were asked to evaluate the U.S. health care system in general, resulting in reports of lower satisfaction. Based on previous research in other domains (see Schwarz, 1994), we may expect that respondents would not have proceeded in this manner had the two questions been separated by several buffer items, thus clouding the conversational relatedness of the two questions asked. In fact, reminding them of their own insurance would most likely have increased their general satisfaction reports in this case (see Schwarz, Strack, & Mai, 1991), again illustrating that we need to consider respondents' use of information in addition to the information's accessibility per se.

Conclusions

As the selected examples illustrate, reports of health and health care satisfaction are highly context dependent. The movement toward "report card" standards for managed care practices and as part of health care reform, by which beneficiaries of health care plans would rate their satisfaction with the respective plans, suggests that the reporting of health care events and satisfaction with the delivery of care can be standardized. However, the findings presented here

underline the need for understanding the ramifications of context effects, especially with respect to reporting satisfaction. For example, health care reform legislation may mandate that health care providers achieve minimum levels of satisfaction or may require publication of satisfaction levels for comparison shopping by consumers. Understanding how respondents arrive at these judgments and how these judgments are shaped by the research instrument could determine whether specific standards should be followed in the design of health policy surveys. Although the basic cognitive and communicative processes underlying satisfaction judgments are relatively well understood (Schwarz & Strack, 1991; Schwarz, Wänke, & Bless, 1994), the application of these theoretical principles to health policy issues clearly requires further research.

References

Bachman, G., & Alcser, K. (1993). Limitations of conventional polling methods in studying attitudes about health care in America: An example of question sequence effects. Unpublished manuscript, University of Michigan.

Fahrenberg, J. (1975). Die Freiburger Beschwerdenliste FBL. Zeitschrift für Klinische Psychologie, 4, 79–100.

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Memory and Cognition, 11, 114–120.

Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. In C. Hendrick & M. S. Clark (Eds.), Review of personality and social psychology: Vol. 11. Research methods in personality and social psychology (pp. 98–119). Beverly Hills, CA: Sage.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. Zanna (Ed.), Advances in experimental social psychology (Vol. 26, pp. 123–162). San Diego: Academic Press.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: Assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), The construction of social judgment (pp. 217–245). Hillsdale, NJ: Erlbaum.

Schwarz, N., & Bless, H. (in press). Mental construal processes and the emergence of context effects in attitude measurement. Bulletin de Methodologie Sociologique.

Schwarz, N., & Hippler, H.-J. (1991). Response alternatives: The impact of their choice and ordering. In P. Biemer, R. Groves, N. Mathiowetz, & S. Sudman (Eds.), Measurement error in surveys (pp. 41–56). Chichester, UK: Wiley.

Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Anwortvorgaben des Fragebogens lernen. (Frequency-reports of psychosomatic symptoms: What respondents learn from response alternatives.) Zeitschrift für Klinische Psychologie, 22, 197–208.

Schwarz, N., & Strack, F. (1991). Evaluating one's life: A judgment model of subjective well-being. In F. Strack, M. Argyle, & N. Schwarz (Eds.), Subjective well-being: An interdisciplinary perspective (pp. 27–47). Oxford, UK: Pergamon.

Schwarz, N., Strack, F., & Mai, H. P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. Public Opinion Quarterly, 55, 3–23.

Schwarz, N., Wänke, M., & Bless, H. (1994). Subjective assessments and evaluations of change: Some lessons from social cognition research. In W. Stroebe & M. Hewstone (Eds.), European review of social psychology (Vol. 5, pp. 181–210). Chichester, UK: Wiley.

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, mood, and mode of thinking. Journal of Personality and Social Psychology, 49, 1460–1469.

Comparing Survey Measures of Quality of Medical Care

Floyd J. Fowler Jr. and Lin Bin

Introduction

There is great interest in using survey methods to evaluate the medical care that people receive. When studying treatment outcomes, researchers want to ask patients about their perceptions of their treatment and its results. It has been suggested that managed care plans can be evaluated by surveys of patients. Third, in general population surveys, it would be desirable to be able to evaluate the quality of medical care people receive or have access to by asking them questions.

A critical issue in implementing such proposals is to decide what questions to ask; a corollary is that we need to know what the answers to alternative questions mean. There are at least five different approaches one could take to assessing the quality of care people receive, based on the following aspects of care:

1. The process of care: If there are aspects of how care is delivered that people can agree are valuable, patients can be asked about their perceptions of those processes.
2. Specific results of care: If treatment has a clear goal, patients can be asked whether or not the goal was achieved.
3. Complications due to care: If treatments received produce unwanted effects, patients can be asked to describe those.
4. General health status: Patients can be asked to describe their state of functionality and well-being after treatment. Good health status after treatment could be an indicator of good medical care.
5. Ratings of care: Patients could be asked directly for their ratings of the medical care they receive or the results of treatment.

Surveys of patients treated for prostate cancer with either low-beam radiation or radical prostatectomy included examples of all of these kinds of measures. These data thus provide an opportunity to compare and contrast the results that would be derived from using these various kinds of measures to assess medical care. The specific goal of the analysis presented in this paper is to assess the way in which patient ratings of medical care or the results are associated with other, less subjective assessments of the quality of medical care received.

Floyd J. Fowler Jr. and Lin Bin are at the Center for Survey Research, University of Massachusetts–Boston.

This work was supported by grants HS06336 and 08397 from the Agency for Health Care Policy and Research.

Methods

The analysis presented in this paper melds the results from two parallel surveys. For the survey of prostate surgery patients, the 5% file of Medicare beneficiaries was sampled to identify those who had had radical prostatectomy to treat prostate cancer during a 3-year period (1988–90). A sample of 420 such patients was drawn, equally distributed across the 3 years of data collection. The data reported here were collected by mail with repeated follow-ups, followed by telephone interviews for nonrespondents. Out of 402 eligible patients, a total of 367 responded, yielding an overall rate of response of 91%. In addition, we collected data using the same methodology from 162 cases out of an oversampled Massachusetts group. These cases were weighted and added to the initial group of 367, yielding a total effective sample of 373 cases.

The sample of patients treated with radiation for cancer was drawn using the Surveillance, Epidemiology, and End Result (SEER) system in three states. In those states, all cases of cancer are to be referred to the registry. In a manner parallel to that used for the surgical cohort, samples of patients initially diagnosed from 1989 through 1991 who were covered by Medicare were selected, again with equal numbers chosen across years. These data were collected in 1994. Data collection methods were parallel to those for the other cohort, with a combination of mail and telephone data collection procedures being used. The response rate was 83%; a total of 621 cases are available for analysis.

The critical elements of this analysis are the measures included in the surveys. Specific question wording is included in the Appendix.

Process of Care

Published data do not establish that surgery, radiation, or no treatment at all offers a clear survival benefit,

particularly for Medicare patients (Fleming, Wasson, Albertsen, Barry, & Wennberg, 1993; Wasson et al., 1993). However, the treatment options have widely different immediate effects on patient quality of life (Fowler et al., 1993; Shipley et al., 1993). Hence, one potentially important aspect of the process of medical care is thoroughly discussing the available options with a patient. Our measure of process was simply the patient report of whether or not a physician discussed any alternatives to the treatment that the patient actually received.

Appropriateness of Care

If a patient does not have least 10 years of life expectancy, there is limited evidence that either surgery or radiation has survival value (Brendler & Walsh, 1992). While life expectancy cannot be predicted by age alone, those over 75 are less likely than younger men to derive any benefit from aggressive therapy.

Results of Care

The rationale for either surgical or radiologic treatment of prostate cancer is to cure the cancer. Approximately 29% of the patients in the survey reported that their cancer had recurred or that they had had additional treatment for recurrence of cancer since their primary treatment. For such patients, one could say the initial treatment was not effective. Our measure of recurrence was whether or not patients reported that they had follow-up treatment for recurrence of cancer or reported that they thought they still had cancer.

Complications Due to Care

The treatments for prostate cancer themselves can have important effects on patients. Prostate surgery and radiation can result in sexual impotence and mild to moderate incontinence (requiring patients to wear pads) (Fowler et al., 1993; Litwin et al., 1995; Shipley et al., 1993). Patients were asked about the rates at which they had experienced these effects of treatment. For sexual dysfunction, patients were classified by whether or not they reported having had any erections since their treatment. With respect to incontinence, patients were classified by whether or not they reported currently wearing pads or clamps to deal with wetness.

In addition to questions related to these two descriptive measures of the most common complications of treatment, patients were asked a series of questions about how problematic they considered each of these areas to be, as well as questions on how problematic they found bowel problems (another possible effect of radiation in particular) and worry about cancer. This index enabled patients to put their own stamp of values on the various things that had happened to them as a result of the diagnosis and treatment of cancer. Each potential problem was rated on a 5-point categorical scale from "big problem" to "no problem at all." The sum of the ratings with respect to seven potential problems constituted a problem score.

General Health Status

The measure of health status used in this analysis is the standard National Health Interview Survey self-rating of health, in which the person rates health from "excellent" to "poor."

Ratings of Care

Patients rated the quality of medical care they received for the treatment of prostate cancer on a 5-point scale from "excellent" to "poor." In addition, they were asked how they felt about the way that their treatment had worked out. The response task was to fill out a seven-category response scale ranging from "delighted" to "terrible," derived from the work of Andrews and Withey (1976).

The principal focus of the analysis was to relate our measures of the process of care and how things turned out as a result of care to patient ratings of the quality of medical care and how patients felt about the results of care. In short, we were looking at how patient assessments were related to more objective measures of care. We carried out the analysis using a series of cross-tabulations and regression analyses.

Results

The most important difference between the two treatment groups is that radiation patients are older than the surgery patients. In addition, the radiation group includes slightly more nonwhites than the surgery group. There is no difference between the two groups in terms of education, work, and marital status. Effects on sexual function and incontinence are higher for surgery patients; bowel effects are more common among radiation patients. Ratings of medical care and results do not differ by treatment.

Tables 1 and 2 summarize the results of cross-tabulations of the results of care with patient ratings of the quality of medical care and the results.

One striking finding presented in Tables 1 and 2 is that most of the associations were statistically significant at a probability of less than .0001. The exceptions were that neither of the ratings was related to patient age or to our measure of process of care, the extent to which physicians were said to have discussed alternatives to the radiation or surgical therapies that the patients underwent. Those who said that no alternative treatments were discussed were no more negative about their medical care or how treatment turned out than others. In addition, those who reported that they had no sexual function since treatment did not rate the

Table 1. Patient rating of medical care by various measures of quality of care

Measures related to quality of care	Medical care rating (percentages)				p[a]	n
	Excellent	Very good	Good	Fair or poor		
Process: Alternative treatments mentioned					ns	
Yes	60	24	11	5		335
No	57	27	10	6		629
Age at treatment					ns	
< 70	61	24	9	6		352
70–74	57	27	10	6		368
75+	57	28	12	3		232
Cancer recurred					p < .05	
Yes	51	32	10	7		258
No	61	24	10	5		667
Wear pads for wetness					p < .001	
Yes	49	26	14	11		147
No	60	27	9	4		773
Impotent after treatment					ns	
Yes	56	26	12	6		307
No	62	24	8	6		429
Problem index					p < .0001	
Top quartile	79	17	4	0		190
Middle two quartiles	61	27	9	3		441
Bottom quartile (most problems)	37	33	16	14		194
Health rating					p < .0001	
Good or better	65	25	7	3		683
Fair or poor	41	27	18	14		260

[a]"ns" means not significant at .05 level.

Table 2. Reported feelings about how treatment worked out by various measures of quality of care

Measures related to quality of care	Feelings about how treatment worked out (percentages)				p[a]	n
	Delighted/ pleased	Mostly satisfied	Mixed	Mostly dissatisfied/ unhappy/terrible		
Process: Alternative treatments mentioned					ns	
Yes	49	24	23	4		326
No	50	27	16	7		622
Age at treatment					ns	
< 70	55	25	14	6		343
70–74	46	28	20	6		362
75+	49	25	22	4		230
Cancer recurred					p < .0001	
Yes	35	28	26	11		256
No	56	25	15	4		653
Wear pads for wetness					p < .001	
Yes	34	31	24	11		143
No	53	25	17	5		762
Impotent after treatment					p = .05	
Yes	42	29	21	8		306
No	52	24	19	5		418
Problem index					p < .0001	
Top quartile	82	15	2	1		188
Middle two quartiles	49	31	17	3		431
Bottom quartile (most problems)	20	31	34	15		188
Health rating					p < .0001	
Good or better	58	24	15	3		674
Fair or poor	29	32	26	13		252

[a]"ns" means not significant at .05 level.

quality of their medical care lower than others, though they were less positive about how treatment turned out. A second observation about the tables is that while the associations are significant, at a face validity level, many patients gave what would appear to be positive ratings to their care and the results of treatment, even when by some standards, one might think the results were not very good.

From some perspectives, the worst outcome is that cancer recurred, since the goal of treatment for these patients was to treat cancer. For patients whose cancer had recurred, the majority (63%) rated their response to treatment "mostly satisfied" or better; 83% rated their medical care as "good," "very good," or "excellent."

The problem index, which summed patients' ratings of problems with urination, sexual and bowel function, and worry about cancer, was the variable most closely related to patient ratings of quality of care and results of treatment. However, even with respect to this measure, only 14% of the bottom quartile on this measure (the 25% who rated the problems resulting from treatment to be most problematic) rated their medical care as "fair" or "poor." Indeed, less than a majority (49%) of that group rated their feelings about the results of treatment to be "mixed" or negative.

We carried out two regression analyses, using the variables in the left-hand column of the tables to predict the rating of medical care and feelings about the results of treatment. A priori, we reasoned that three variables ought to make the most difference: the nonrecurrence of cancer, having few problems associated with treatment, and being healthy after treatment. The most parsimonious set of variables, then, was cancer recurrence, the problem index, and the single item self-rating of health.

When these three variables (in slightly expanded form) were used as predictors, they explained 27% of the variance in satisfaction with the results of treatment (multiple R = .52) and 16% of the rating of the quality of medical care received (multiple R = .40). When all the variables in Table 1 were used in a regression analysis, the result remained virtually the same.

Discussion

Evaluating medical care is not easy for patients in some respects. Based on the measures analyzed in this study, there were only two groups of patients who could clearly be said to have had suboptimal care: those who did not have treatment options presented to them and those men over 75. Treatment of prostate cancer is a condition for which it is easy to argue the treatment options must be presented to patients so they can participate in making choices. Given the lack of evidence of increased survival effects of radiation over surgery, or vice versa, and given the markedly different effects of treatments on patients, having only one treatment option presented without serious discussion of alternatives is arguably poor medical care and will lead to suboptimal treatment decisions.

In a similar way, as noted, if a patient does not have at least 10 years of life expectancy, there is little basis for recommending major intervention, with all its side effects. Hence, most 75 year olds probably have a legitimate complaint. However, neither patient age nor the extent to which options were discussed were related to patient ratings of the quality of the medical care they received. These are examples of standards of care that depend on patient reports (only patients can report what options they are given) but that are not reflected in patient ratings of care.

It could be argued that the various measures of how things turned out—complications such as incontinence and impotence, measures of patient health status, and whether or not cancer recurred—are not measures of the quality of medical care. These complications occur at high rates after these treatments and are not considered to be indications of lack of good technique or execution of the procedures. Most physicians would say they are simply part of the price that patients have to pay in order to undergo these therapies for their prostate cancer. While such measures are essential to outcome studies, in which researchers try to develop a profile of where patients are likely to end up after treatment, there is a reasonable case to be made that these are not good measures of the quality of medical care. Recurrence of cancer—the fact that 29% of the patients had follow-up treatment for recurrence within 3 years of primary therapy—could be taken as an indicator of the quality of care. Treating people surgically or with radiation when therapy is not likely to retard the growth of cancer arguably is not good medicine. However, one would have to have much more information about the patients, including what kinds of tests were done prior to treatment, in order to make an informed decision about whether recurrent treatment rates are actually a meaningful measure of the quality of medical care.

Overall, patients give high marks to the quality of medical care they receive, even when it leaves them impotent, incontinent, and with the cancer they were hoping to eliminate. Moreover, how decisions were made and patient age, both of which are key to the quality of medical decision making in this area, have no significant relationship at all to how patients rate their medical care.

Understandably, when patients rate how they feel about how things turned out, their posttreatment health status and complications of treatment play a much bigger role in those ratings. Almost 40% of the variance in how people rate the results of treatment can be explained. Such questions may be one reasonable way to assess the effects of medical care. However, because the rates and types of complications and things that can go wrong are so dependent on the particular condition and the kinds of treatment received, such measures can probably only be interpreted meaningfully on a condition-by-condition basis. Condition-by-condition analyses are the norm in clinical studies, but they are hard or complicated to do in general population surveys, due to the rarity of clinically similar people.

The calibration of patient ratings may be important in how they are used. While the words may sound all right, "very good" and "mostly satisfied" may not be very good ratings. Those are the ratings that are common when people are treated for cancer and end up wearing pads or with recurrence of cancer.

Many researchers are interested in how best to use surveys of patients to assess the quality of medical care to which people are exposed. There can be little doubt that patient reports have a critical contribution to make to the assessment of medical care. Moreover, especially ratings of results of treatment have strong, predictable relationships to some descriptive measures of the effects of care. However, these analyses illuminate how complicated it can be to use such reports. The occurrence of what appear to be adverse events do, on average, affect the ratings patients give, but it is clear there is wide variation in the response to the same events. Being impotent or wearing pads for wetness was a "terrible" outcome for some and an acceptable one for others. If a result does not have a negative effect from a patient's point of view, is it fair to count it as an adverse outcome? At the same time, there are examples in these analyses of patients who likely did not receive very good medical care whose ratings do not reflect those effects.

The results are consistent with the conclusions of those who argue that one of the most important ways to use patient reports is to describe aspects of the process of care that are established on other grounds to be indicators of the quality of care. Using that approach with these data, one would look for the rates at which patients said they did not have discussions about treatment options and look for patients who received surgery who were unlikely to have 10 years of life expectancy. Those measures, possibly more than patient ratings, might be the best indicators of substandard medical care. At the same time, patient ratings of how they responded to the treatments and the effects clearly also are needed in order to have a comprehensive picture of what patients have to say. Only patients can reliably report treatment effects such as impotence, incontinence, and worry about cancer.

Prostate cancer may be a particularly hard condition for which to assess the quality of medical care. There is significant controversy about what is the best treatment; the expected benefits of treatment—extended survival—are not observable in the short term. Treatments designed to reduce observable symptoms and improve functionality may be somewhat easier to evaluate. Nonetheless, for virtually all medical care, success is multifaceted. Effects on the treated condition, complications of treatment, posttreatment health status, and patient satisfaction are all aspects of quality of care. They sometimes are correlated, but they are far from identical. Producing a single score is likely to mask important variation. These data certainly suggest that simple solutions are likely to be poor solutions to the problem of how to measure quality of care in surveys.

Appendix: Questions Used in the Construction of the Measures of Quality of Care

1. Process of Care

(1) Before you began treatment, did any doctor discuss the possibility of not having any treatment at all?

(2) Did you think that doctor thought having no treatment at all was something you should seriously consider, or not?

(3) Before you began treatment, did any doctor discuss having prostate (surgery/radiation) instead of (radiation/surgery)?

(4) Did you think that doctor thought having (surgery/radiation) instead of (radiation/surgery), was something you should seriously consider, or not?

2. Results of Care

(1) As far as you know, do you have cancer anywhere now?

(2) Since your first (radiation/surgery) treatment, have you had:

  (a) any pills or injections of estrogens or hormones for prostate cancer?

  (b) surgery to remove the testicles?

(3) Have you had any (additional) radiation treatment?

3. Complications Due to Care

(1) Sexual Dysfunction

  (a) Since your (radiation/surgery) treatment, have you had any full erections at all?

  (b) Have you been able to have any partial erections?

(2) Incontinence

  (a) After (radiation/surgery) treatment, some men find they have a problem with dripping or leaking urine. Did you have that problem to any degree either right after (radiation/surgery) treatment or at anytime since?

  (b) Do you still have any problem at all with dripping or leaking urine?

  (c) Some men wear pads, rubber pants, adult diapers, or a clamp to help with wetness. Do you use anything like that now?

(3) Problems

Over the past month, how much have each of the following been a problem for you? [Response Categories: No Problem, Very Small Problem, Small Problem, Medium Problem, Big Problem]

  (a) Dripping or leaking urine?

  (b) Frequent urination?

  (c) Having to urinate without much warning?

  (d) Concern about sexual functioning?

(e) Having pain or discomfort with bowel movement?

(f) Having frequent bowel movement?

(g) Worry about prostate cancer?

(h) Worry about any other kind of cancer?

4. Self-rated Health

Overall, how would you rate your health now: excellent, very good, good, fair, or poor?

5. Ratings of Care

How would you rate the medical care you received for prostate cancer: excellent, very good, good, fair, or poor?

6. How Feel About How Treatment Worked Out

If you were to spend the rest of your life feeling the way you feel now, how would you feel about that: delighted, pleased, mixed (about equally satisfied and dissatisfied), mostly dissatisfied, unhappy, or terrible?

References

Andrews, F. M., & Withey, S. B. (1976). Social indicators of well-being. New York: Plenum Press.

Brendler, C. B., & Walsh, P. C. (1992). The role of radical prostatectomy in the treatment of prostate cancer. CA: A Cancer Journal for Clinicians, 42, 212–222.

Fleming, C., Wasson, J. H., Albertsen, P.C., Barry, M. J., & Wennberg, J. E. (1993). A decision analysis of alternative treatment strategies for clinically localized prostate cancer. Journal of the American Medical Association, 269, 2650–2658.

Fowler, F. J. Jr., Barry, M. J., Lu-Yao, G., Roman, A., Wasson, J., & Wennberg, J. E. (1993). Patient-reported complications and follow-up treatment after radical prostatectomy. The national Medicare experience: 1988–1990. Urology, 42, 622–629.

Litwin, M. S., Hays, R. D., Fink, A., Ganz, P. A., Leeke, B., Leach, G. E., & Brook, R. H. (1995). Quality-of-life outcomes in men treated for localized prostate cancer. Journal of the American Medical Association, 273, 129–135.

Shipley, W. U., Zietman, A. L., Hanks, G. E., Coen, J. J., Caplan, R. J., Won, M., Zagars, G. K., & Asbell, S. O. (1993). Treatment-related sequelae following external beam radiation for prostate cancer: A review with an update in patients with stage T1 and T2 tumors. Massachusetts General Hospital, MA: Unpublished manuscript.

Wasson, J. H., Cushman, C. C., Bruskewitz, R. C., Littenberg, B., Mulley, A. G. Jr., & Wennberg, J. E. (1993). A structured literature review of treatment for localized prostate cancer. Archive of Family Medicine, 2, 487–493.

Obtaining Patient Reports and Evaluations of Care for Quality Improvement in an Urban Teaching Hospital

Lisa E. Harris, William M. Tierney, and Morris Weinberger

Introduction

In today's increasingly competitive health care environment, health care organizations are recognizing the need to be sensitive and responsive to patients' judgments regarding the quality of the care they receive. However, while "consumer satisfaction" surveys are gaining popularity among a variety of provider organizations, primarily as tools for marketing and competition, high satisfaction ratings do not necessarily imply high-quality care.

To empower patients to affect the quality of their care, efforts directed at eliciting information on patient satisfaction need to go beyond global ratings, asking, rather, for reports of specific care events and processes. For example, while a question related to patient satisfaction with preparation for returning home from the hospital might ask patients to rate the hospital staff in preparing them to care for themselves once home, an alternate approach would be to ask patients to report on specific care events or processes related to preparation for discharge (for example, "Did any of the hospital staff assist you in getting help that you could not get from family or friends, such as nursing or housekeeping assistance?"). These two strategies (global assessment of care and assessment of specific events and processes) capture vital and complementary dimensions of satisfaction. However, for measures of patient satisfaction to be useful for improving the quality of care, emphasis should be placed on the latter approach, which seeks patient reports regarding specific events and processes, using these reports from patients to (a) identify problems for intervention and (b) determine the effect of the resulting quality improvement efforts. For this strategy to be successful in closing the loop linking patient reports and evaluations of care to improvements in health care quality, it must be sensitive both to the context in which that care is delivered as well as to the population served.

Lisa E. Harris is in the Department of Medicine, Wishard Memorial Hospital and Indiana University School of Medicine and with the Regenstrief Institute for Health Care, Indianapolis. William M. Tierney is in the Department of Medicine, Wishard Memorial Hospital and Indiana University School of Medicine, and with the Regenstrief Institute for Health Care; he is also with the Richard L. Roudebush Veteran's Affairs Medical Center. Morris Weinberger is at Indiana University, the Regenstrief Institute for Health Care, and the Richard L. Roudebush Veteran's Affairs Medical Center.

This paper describes our efforts to implement this process in an urban teaching hospital. We have targeted a special population that, though not representative of health care consumers as a whole, does represent under all scenarios of health care reform the most vulnerable and, therefore, the most important population in terms of assessing and improving health care quality. This socioeconomically disadvantaged minority population also represents a group that has been typically underrepresented in consumer satisfaction surveys, largely because of the difficulties encountered in achieving desirable survey response rates (Acuff, Martin, & Andrulis, 1994; Cleary et al., 1991). We will briefly describe our efforts to adapt a previously developed instrument for use in our population; then discuss the results of our initial pilot test of the instrument; and, finally, present the results of a randomized trial in which we compared two protocols for surveying this population in terms of response rate, data quality, and cost.

Survey Development

In 1988, the Picker/Commonwealth Program for Patient-Centered Care conducted a nationwide survey to determine which aspects of inpatient care are most important to patients, seeking reports regarding specific care events and processes as well as evaluations of care (Cleary et al., 1991; Cleary, Edgman-Levitan, McMullen, & Delbanco, 1992). The survey, administered to a total of 6,455 recently hospitalized patients by telephone interview, focused on events selected to indicate quality of care as reported and evaluated by patients along the following dimensions: (a) respect for patients' values and preferences, (b) coordination of care and integration of services, (c) communication between patients and providers, (d) physical comfort, (e) emotional support, (f) involvement of family and friends, and (g) transition and continuity from one locus of care to another. Maintaining the dimensions of the Picker/Commonwealth Survey, we conducted focus groups with patients recently discharged from Wishard Memorial Hospital, a 250-bed teaching hospital serving the inner-city indigent of Indianapolis. We then modified the instrument to reflect the unique nature of the teaching hospital and the needs of our patient population by constructing questions addressing the issues emerging from these focus groups and

combining them with relevant questions from the Picker/ Commonwealth instrument. Finally, in a series of patient interviews, we tested the questions to determine whether they were understood and accepted by our patients and, indeed, whether or not patients' interpretation of the questions reflected our intent.

The modified instrument incorporates questions that ask patients to report on specific care events as well as those that ask patients to rate their satisfaction with these aspects of care. A total of 116 items covers the range of inpatient care processes from admission through discharge and outpatient follow-up. Included are specific questions related to discrete facets of a process or aspect of care as well as overall ratings of care and care providers. Also included are open-ended questions asking what could have been done to improve specific aspects of care. For example, a series of questions asking patients to report on specific events related to preparation for returning home is followed by a question asking them to rate the hospital staff in preparing them to care for themselves at home and, finally, by an open-ended question asking them how the staff could have done a better job. Rather than using "yes" or "no" response categories for those questions asking for reports of care events or processes, where appropriate, we used the response options "always," "usually," "sometimes," "rarely," and "never" to increase the sensitivity of the instrument for detecting change related to quality improvement interventions developed in response to the patient surveys.

Pilot Testing the Survey Instrument

For our pilot test, we contacted patients by telephone within two weeks of discharge, using telephone numbers provided upon admission to the hospital. Of the 678 patients discharged from the general internal medicine service over a 6-week period, 62 (9%) had no phone number listed, 108 (16%) were discharged to nursing homes or other institutions, and 34 (5%) were ineligible because they were prisoners. Of the remaining 460 patients who were candidates for the telephone survey, 133 (29%) were unreachable either because of disconnected phones, phone numbers changed to unpublished numbers, or because a friend's or relative's phone number had been given. An additional 39 (8%) had been readmitted to the hospital, 27 (6%) were poor candidates for a telephone interview because of impaired hearing, and 68 (15%) were unreachable after a minimum of five calls. Of the 193 patients contacted by phone, 138 (72%) completed the 30-minute interview. However, these 138 patients represented only 30% of those patients who were discharged to home from the hospital and were thus candidates for the survey.

Survey Protocol Testing

We were concerned about the failure of this telephone protocol to reach the majority of the eligible candidates. Further, since almost 30% were lost to disconnected, unpublished, or otherwise bad phone numbers, we were particularly concerned that we might be missing those vulnerable patients most likely to experience problems with care because they might also be those most reluctant to provide personal information on admission to the hospital (e.g., true phone numbers) or least likely to have a phone where they could reliably be contacted.

Therefore, we next conducted a simultaneous test of two protocols. First, we contacted patients while they were hospitalized, informed them of the postdischarge survey, and obtained their consent to participate. At that time, we asked for current phone numbers and addresses with the assurance that the information was to be used only for the purpose of contacting them for the survey. Patients were then randomized (using the even or odd status of a sequentially assigned hospital number) to either a mail or telephone protocol for survey administration. In this way, we were able to simultaneously (a) test the effect of informing patients of the survey while they are hospitalized (and at the same time obtain current phone numbers) on the rate of response to a telephone protocol (comparing the results with our prior survey) and (b) compare response rates and data quality of mail and telephone survey protocols.

There were 390 patients admitted to and discharged from the general internal medicine service during the 4-week study period, 52 (13%) of whom were discharged to nursing homes or transferred to other institutions, 31 (8%) of whom were prisoners, 30 (8%) of whom were determined by the research assistant to be incapable of participating because of chronically impaired mental status or being homeless (having no address or telephone access), and 20 (5%) of whom were discharged before they could be contacted by the research assistant. Of the 257 eligible patients, 252 (98%) agreed to participate, 122 of whom were randomized to the telephone-first protocol and 130 to the mail-first protocol. As shown in Table 1, there were no between-group differences in demographic characteristics or length of hospitalization; in only one of the five most common categories of discharge diagnoses (symptoms, signs, and ill-defined conditions) was there a statistically significant difference between the two groups.

Outcome Measures

We compared the two survey protocols along three dimensions. Response rates were calculated for each group as the ratio of completed surveys to the total number of patients randomized to each protocol. Missing data was defined as the number of questions for which no answer was given and was calculated separately for the survey as a whole, as well as for all items not involved in skip patterns, and finally, for open-ended questions. Data collection costs per completed mailed survey included supplies, printing (of questionnaires, business reply envelopes, and labels), and postage. For the telephone survey, we used the $15 per completed survey charged by our university-affiliated survey research laboratory. To each survey obtained by either method, we added the wages and

Table 1. Study patient demographic and clinical characteristics by protocol

	Protocol	
	Mail first (n = 130)	Telephone first (n = 122)
Age (± SD)	51.6 ± 18.1	51.1 ± 16.1
Sex (% male)	46	49
Race (% black)	55	48
Days hospitalized (± SD)	4.8 ± 3.3	4.8 ± 3.4
Most common discharge diagnoses[a]		
Respiratory disease	25	24
Circulatory disease	22	16
Symptoms, signs, and ill-defined conditions	7	16*
Endocrine, nutritional, metabolic, immunologic diseases	8	10
Gastrointestinal diseases	7	10

[a]Most common discharge diagnoses, given as percentage of enrolled patients.

*p < .05.

fringe benefits of the personnel who contacted patients in the hospital and entered and managed the data. We calculated an average cost for each method (telephone-first and mail-first) by dividing the total cost for each method by the number of completed, usable surveys obtained by that method.

Statistical Analysis

For all outcomes, we used corrected chi-square and t tests to compare the two groups' categorical and continuous variables, respectively. We also performed logistic regression analysis to identify characteristics independently associated with nonresponse, retaining in the final model all variables with a multivariable p value of less than .05. Independent variables that were candidates for analysis included demographic and clinical characteristics, along with study group assignment.

Results

Of those 130 patients randomized to the mail-first protocol, 28 (22%) completed a survey by mail and 37 (28%) completed a survey by phone for a response rate of 50%. Of the 122 patients randomized to the telephone-first protocol, 80 (66%) completed the survey by telephone and 9 (7%) by mail for an overall response rate of 73%. This difference was statistically significant (p < .001; see Table 2). There were no significant differences in race or sex between respondents and nonrespondents for either the telephone-first or the mail-first protocols. However, for the telephone-first protocol only, respondents were significantly older than nonrespondents (mean age was 54.9 ± 14.75 vs. 42.2 ± 15.7, p < .001).

We performed logistic regression analysis to identify independent predictors of those patients who responded to the survey. Independent variables included the demographic and clinical data shown in Table 1, along with an indicator of study group assignment. Respondents more often had discharge diagnoses of respiratory disease (multivariable OR = 3.4, 95% CI 1.6–7.4, p = .002) or ill-defined symptoms, signs, or conditions (OR = 3.0, 95% CI 1.1–8.3, p = .04). The strongest predictor of responding, however, was study group assignment, with those randomized to the telephone-first protocol having the higher response rate (OR = 2.7, 95% CI 1.5–4.7, p = .0008).

We next examined missing responses in surveys completed by the two protocols. Of the instrument's 116 items, there were significantly more missing responses on surveys obtained via the mail-first protocol (28.7 ± 15.4; 25%) than the telephone-first protocol (24.1 ± 11.0; 21%, p < .05). There was no statistically significant difference in the completion of the 65 closed-ended items not involved in skip patterns (3.2 ± 9.3 for mail first vs. 1.5 ± 4.8 for telephone first), but among the 10 open-ended items, responses were missing for 6.4 ± 1.9 items in the mail-first protocol compared with 5.1 ± 1.7 in the telephone-first protocol (p < .05; see Table 2).

The total cost per completed survey was $56.40 for those obtained via mail and $22.94 for those obtained via telephone interview. For the mail-first protocol, by which 28 surveys were completed by mail and 37 by phone, the total cost per completed survey was $37.35. This figure is 42% higher per returned usable survey compared with the telephone-first protocol, where the mean cost was $26.32 for the 80 surveys completed by telephone and 9 returned by mail (p < .001; see Table 2).

Finally, we compared all surveys returned in the mail with all surveys completed by telephone interview in terms of missing data. For all 116 items, 36 (31%) were missing

Table 2. Study outcome variables by protocol

	Protocol	
	Mail first (n = 130)	Telephone first (n = 122)
Questionnaire completion		
Mail return	28 22%	9 7%
Telephone interview	37 28%	80 66%
Total	65 50%	89 73%*
Missing responses[a]		
All 116 items	28.7 ± 15.4	24.1 ± 11.0*
65 nonskip items	3.2 ± 9.3	1.5 ± 4.8
10 open-ended items	6.4 ± 1.9	5.1 ± 1.7*
Surveying costs[a]	$37.35 ± $16.70	$26.32 ± $10.15*

[a]Results are M ± SD per completed survey.

*p < .05.

from those returned in the mail compared with 23 (20%) for those completed by phone interview. For the 65 closed-ended questions not involved in skip patterns, data were missing on average for 7 questions (10%) on the mail-returned surveys, compared with 0.7 (1%) in the telephone surveys, p < .001. For the 10 closed-ended questions, a mean of 8 went unanswered on the mail-returned surveys, compared with 5 on the surveys completed over the telephone (p < .001; see Table 3).

Discussion

This randomized trial of two protocols for postdischarge surveys of patients hospitalized at an urban teaching hospital demonstrates the clear superiority of the telephone-first protocol over the mail-first protocol in terms of response rates, data quality, and cost. We were surprised when our cost estimates strongly favored telephone interviews. Because our estimates were based on the cost per completed survey, the increased cost associated with the mail-first protocol is largely attributed to the poor response rate. Moreover, of the 65 surveys completed by the mail-first protocol, 37 (57%) were completed by the backup telephone interviews. Further, for those surveys that were completed by mail, higher rates of missing data and skip pattern errors, along with fewer textual responses to open-ended questions, made data collected by the mail surveys even less useful.

In addition, using a strategy in which we contacted patients while they were still in the hospital, informed them of our postdischarge survey, and obtained current telephone numbers and addresses, we more than doubled the 30% response rate to the telephone-first protocol that we obtained in the pilot survey using a similar telephone protocol without contacting patients in the hospital. Our overall response rate of 73% for all eligible patients for this telephone survey with mail follow-up of nonresponders was markedly higher than that reported by the National Public Health and Hospital Institute to a study that used the Picker/Commonwealth Survey to interview patients served by seven Public Hospital Task Force institutions (Acuff et al., 1994). Although they reported a response rate of 90% for eligible patients, their denominator did not include 55% of the patients to be surveyed who were only considered ineligible because they could not be reached by telephone.

Our increased overall response rate could be attributed to two factors: obtaining accurate telephone numbers and establishing personal contact with patients while they were still in the hospital. We can with confidence attribute only 10% of our increased response rate to the telephone protocol to obtaining phone numbers and addresses from

Table 3. All mail-returned surveys compared with all telephone interviews

Outcome variable	Mail-returned surveys (n = 37)	Telephone interviews (n = 117)
Missing responses		
All 116 items	35.8 ± 18.0	23.0 ± 9.5***
65 nonskip items	7.1 ± 12.3	0.7 ± 2.9***
10 open-ended items	7.9 ± 1.8	5.0 ± 1.4***

NOTE: Results are M ± SD.

***p < .001.

patients that were different from those provided on admission. However, the majority of the 30% of patients unreachable by phone in the original protocol were recorded as such when the research assistant made telephone contact and was told either that she had reached a relative's or friend's phone number and that the patient was unavailable for an interview or that this was not the patient's residence. It is likely that by establishing personal contact with patients, explaining to them the importance of the survey and how the results would be used, answering their questions about it, and eliciting their consent to participate, we increased the potential respondents' interest in the study and their likelihood of agreeing to an interview when called. Indeed, for both mail and telephone surveys, it has been well documented that the greater the respondent's interest in and/or personal commitment to the study, the greater the chances of their returning a questionnaire or completing an interview (Skipper & Ellison, 1966).

This study has several limitations. First, we have focused on the inpatient setting and do not know if the results would generalize to the outpatient setting. However, we specifically chose the inpatient setting because there are many opportunities for patients to have poor experiences because of the multiplicity of health care professionals and the highly charged and stressful events that occur in unusual surroundings for an extended period. In addition, the cost per interview charged by our university's public opinion laboratory might be inordinately low. However, even doubling the rate per completed interview to $30 would not change our conclusions. On the other hand, their standard charge for this service might even be higher than their true costs, making it possible that the $15 we were charged per completed survey is an overestimate. Finally, our disadvantaged population may have supplied more erroneous telephone numbers (in order to avoid inpatient bills) and had a higher prevalence of illiteracy than patients cared for in most U.S. hospitals. Yet our response rate was equal to or greater than that obtained in more wealthy settings. Moreover, these are the most vulnerable patients, and studies have shown that they have higher rates of dissatisfaction with their inpatient care (Cleary et al., 1991; Hall, Feldstein, Fretwell, Rowe, & Epstein, 1990). Thus, we purposely conducted this study in a tax-supported inner-city hospital in order to define the most appropriate protocols for assessing the problems encountered by such vulnerable patients.

We therefore conclude that when patient surveys are used to obtain the detailed information required to design quality improvement interventions, the resulting survey length and complexity require telephone interviews, at least when they are conducted with the socioeconomically disadvantaged patients typically served by urban teaching hospitals. In addition to yielding a greater amount of usable information, telephone interviewing was less costly than mailing surveys. Furthermore, response rates can be markedly improved by informing inpatients of the survey and obtaining telephone numbers and addresses in the hospital.

References

Acuff, K. L., Martin, V., & Andrulis, D. P. (1994). Focus on the patient: Public hospitals and patient-centered care. Washington, DC: The National Public Health and Hospital Institute.

Cleary, P. D., Edgman-Levitan, S., McMullen, W., & Delbanco, T. L., (1992). The relationship between reported problems and patient summary evaluations of hospital care. Quality Review Bulletin, 18, 53–59.

Cleary, P. D., Edgman-Levitan, S., Roberts, M., Moloney, T. W., McMullen, W., Walker, J. D., & Delbanco, T. L. (1991). Patients evaluate their hospital care: A national survey. Health Affairs, 10, 254–267.

Hall, J. A., Feldstein, M., Fretwell, M. D., Rowe, J. W., & Epstein, A. M. (1990). Older patients' health status and satisfaction with medical care in an HMO population. Medical Care, 28, 261–270.

Skipper, J. K., & Ellison, M. D. (1966). Personal contact as a technique for increasing questionnaire returns from hospitalized patients after discharge. Journal of Health and Human Behavior, 3, 211–214.

Reducing Bias in the Measurement of Health Care Satisfaction

Catharine W. Burt

Introduction

These papers look at different issues that affect accurate measurement of the quality of health care or satisfaction with health care from the viewpoint of the patient. The paper prepared by Schwarz, Mathiowetz, and Belli examines the context effects of questionnaire wording and placement and how it relates to differences in perceived health or satisfaction. A second paper by Fowler and Bin examines the relation between objective measures of outcome and subjective ratings of satisfaction and quality of care, and the paper by Harris, Tierney, and Weinberger documents issues related to measuring satisfaction of care in urban teaching hospitals when the population of interest tends to have a higher proportion of both minority patients and patients with lower socioeconomic status compared with other hospitals. These papers all examine some of the main aspects affecting survey error: response error and nonresponse error. The first two papers look at instrument and respondent bias, and the third examines ways to reduce nonresponse bias. All three papers apply their research to measuring the construct of patient satisfaction with health and/or health care.

Response Bias

Schwarz, Mathiowetz, and Belli review literature and unpublished data to discuss the difficulty associated with question wording, placement, and response options in measuring such health concepts as frequency of symptoms and satisfaction with care. They highlight the theoretical discussion on the cognitive process associated with making evaluative judgments. According to their theory, the respondent will make a mental representation of the target of judgment with a standard of comparison. If the target is more favorably compared to the standard, then the evaluation will be positive. If the target is less favorably compared to the standard, then the evaluation will be negative. Depending on the data provided to the respondent on the questionnaire, the perception of both the target and the standard can be affected, and therefore, the evaluative comparison will be impacted. The information that is brought to mind from the questionnaire can have either an assimilation or contrast effect on the judgment. If the information is used to help form the target, then the information leads to an assimilation effect. If the information helps form the standard, then it leads to a contrast effect. Such response errors can radically shape the nature of measures of health satisfaction. The research that Schwarz et al. review helps clarify this theoretical approach with empirical data. Response options can help create a picture of the standard. If you are high in relation to the options, you will have a different evaluation than if you are low. Even your absolute frequency will differ depending on the options presented.

Question wording and placement can also determine whether the information provided on the questionnaire will result in an assimilation or contrast effect. Questions that specifically ask the respondent to compare his or her health to other people's may cause the respondent to use the information provided in the response options of previous items to define the standard. Wording that allows the respondent to "chunk" an event into the past will also lead to a contrast effect, now versus then (Schwarz & Bless, 1992). Survey designers must be very clear about what they wish to measure and draft items that will elicit the correct kind of comparison or eliminate interfering factors. NSchwarz recommends using open response format for frequency reports for better measures of both absolute frequency and evaluative judgments. Subjective measures of satisfaction with health or health care would best be placed before specific items that may cause the respondent to alter the standard of comparison. If your perception of the standard is changed because you hear about all the horrible problems that could happen as a result of health care treatment, then you may tend to be more satisfied with your health or the medical treatment you received than you would otherwise have been.

The theory of assimilation versus contrast effect may help explain some of the results found in the study by Fowler and Bin. They examine the relationship between less subjective and more subjective measures of quality of medical care and satisfaction with health. They surveyed almost 1,000 men who had received either a radical

Catharine W. Burt is the Chief of the Ambulatory Care Statistics Branch in the Division of Health Care Statistics at the National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland.

The author would like to thank Dr. Douglas Herrmann, National Center for Health Statistics, for his review of and comments on the manuscript.

prostatectomy or radiation therapy for prostate cancer. The questionnaire contained seven measures of quality of medical care: process of care, appropriateness of care, results of care, complications due to care, self-rated health status, ratings of care, and attitude toward health. The last two measures were used as dependent variables in an analysis of the association of the previous less subjective measures with the more subjective measures. They found that almost all the men (94%) were very satisfied with both their care and their current health, even though a goodly number of the men received less than optimal care, had some pretty nasty complications, and rated their current health condition as "fair" or "poor." On the face of it, this sounds unlikely, yet on closer inspection, it seems predictable.

Because the general ratings of satisfaction were last in the battery of survey items, most respondents heard (or read) but did not have the enumerated negative aspects of the medical care and complications. As a group, they were lucky in light of all the possible things that might have gone wrong. They therefore, rather predictably, felt they were satisfied with their health care. Respondents compared their health now with their health either before or during treatment, "chunking" their mental representation of their health between now and then. This explanation uses the contrast theory that Schwarz presents. But this theory alone does not fully explain the manifestation of satisfactory ratings or good feelings about one's health.

It has been widely shown that global ratings of health and health care satisfaction tend to be more positive than more objective measures would indicate if one were using rational cognitive processes (Idler, 1993). A likely explanation of the observed discrepancies is that such ratings are not based solely on the objective facts derived from health measures, but also on interactions with the respondent's affect toward the topic. The role of affect on the influence of cognitive processes is well documented (Hoffman, 1986). Affect may result in selective processing of the information about one's health or health care. Research indicates that the processing of data may cease in order to avert painful emotional experiences. There is a human tendency to maintain a positive internal state (Isen, Shalker, Clark, & Karp, 1978). Respondents will tend to recall information associated with positive rather than negative affect. Thus, one explanation for the high frequency of positive ratings is a "Pollyanna" effect (Matlin & Stang, 1978) in which the respondent thinks, "On the whole, things are better now than they were before; after all, I'm not dead. The health care has helped me."

In order for the Pollyanna effect to be a tenable explanation, one would have to accept that affect plays a role in attitudes toward health. Without thoroughly scanning the literature on this issue, I believe that this is true. The more urgent or life threatening the health-related incident, the more emotions are likely to be incorporated into the experience. I certainly know from personal experience that my daughter's doctor visits with shots resulted in more crying than doctor visits without them. Treatments for life-threatening cancer are likely to be highly charged with emotion, which would mean that memories associated with them may be affected by that emotion.

A third explanation of why attitudes toward health and health care are mostly positive may come from the theory of cognitive dissonance. This theory suggests that people change their attitudes and memories about events to be consistent with their values (Festinger, 1957). A major health care event is powerful in the life of a person. Patients are uniquely dependent upon medical staff to get them through a difficult time. They work cooperatively with health care providers to jointly overcome the enemy. If they do not, they tend not to survive the incident. The sample of patients who were treated and survived is therefore a biased sample as to those who would tend to be satisfied. Working jointly in overcoming the enemy allows the patient to buy into the process, to trust the medical staff, and to be pleased with the outcome at a conscious level. Applying the theory of cognitive dissonance to health care ratings would mean that the patients change their attitudes about the care to be consistent with their values of self-esteem. Would the patient have worked so cooperatively with someone who was not providing quality care? Is the patient such a poor decision maker as to decide to undertake treatment that resulted in impotence and incontinence? "Certainly not! The care was good, there are some uncomfortable complications but I'm alive," the patient reasons.

The point of this discussion is that the role of affect must be considered when measuring attitudes toward health or health care. One's health is integrated with the cognitive representation of one's identity (Sehulster, 1994). This is especially true for persons with poor physical, intellectual, or emotional health. Certainly, the emotional state of the patient may influence the ratings. If stable attitudes toward health care are desired, then researchers should utilize psychometric principles of reliability and validity in developing the questionnaire items.

Perhaps as researchers, we shouldn't worry about why health and health care ratings are, in general, high. It is possible that the quality of the medical care is in fact very high. After all, only 10% of Fowler and Bin's respondents had their prostate cancer recur after 2 or 3 years. Respondents who rate their health or health care as low become very interesting. Perhaps the care has to really be horrible for a patient to become dissatisfied. The tolerance level in the patient's dissonance cannot exceed the threshold at which other values become compromised. Fowler and Bin found significant associations between negative medical outcomes and patients' attitudes toward both quality of medical care and outcome. The major exception was the process variable for no alternative treatments having been mentioned by the physician. The distributions of ratings of both quality of medical care and attitude toward the outcome were the same for patients presented with treatment alternatives as for those who were not. After the fact, this distinction may not be salient to the core attitude, compared with other variables such as poor health, cancer recurrence, and nasty complications. In fact, Fowler and Bin show that

these three outcomes accounted for 30% of the variance in satisfaction with current health and 17% of the variance in the rating of quality of medical care. After correcting for attenuation in the dependent variable, it may be that more of the reliable variance is explained.

Nonresponse Bias

The third paper on patient satisfaction, by Harris, Tierney, and Weinberger, examines some of the methodological issues involved in evaluating quality of care from urban teaching hospitals. One of the unique considerations of such hospitals is that the clientele tends to be less educated, less likely to have insurance, and more likely to experience problems in the transition to home (Cleary et al., 1991). Harris and her colleagues modified a rather detailed instrument from the Picker/Commonwealth Program for Patient-Centered Care, which shows that minority and economically disadvantaged patients had higher problem scores. Quality of care was assessed through a multidimensional survey instrument covering various topics in patient care.

One of the main obstacles the researchers faced was obtaining responses from people who were more likely to be dissatisfied with the quality of care provided—the economically disadvantaged. Because this group was differentially harder to contact for inclusion in the survey, the resultant estimates of health care satisfaction were prone to nonresponse error. In their survey pretest, Harris et al. found that 9% of the patients discharged had no phone and 29% were ineligible because of disconnected phones or family and friends' phone numbers being given. The total response rate was only 30% for patients discharged to home. The response rate was dramatically improved in the next test by approaching patients while they were still in the hospital to obtain consent to participate and current phone numbers and addresses. Patients were provided assurance that the information was only to be used to contact them for the survey. In comparing a mail with a telephone protocol, the telephone protocol resulted in both higher survey response and item response rates (73% versus 51% response rate, 1% versus 11% item nonresponse). The researchers conclude that it is imperative to get good contact information for the patients while they are hospitalized and to use a telephone protocol with mail follow-up.

Data presented elsewhere at this conference from a methodological patient follow-up study that we conducted at the National Center for Health Statistics corroborate the fact that using patient contact information from the hospital record is, in fact, insufficient to contact large portions of the patient population. Unfortunately, trying to obtain better contact information not only leads to a more costly survey, but it also increases the respondent burden. Those of us in the federal sector must be concerned now with reducing the burden by 10% in the next 2 years in order to comply with the new Paperwork Reduction Act of 1995. Yet if you cannot get good contact information, you will undoubtedly be left with a biased sample and nonresponse error.

As to economically disadvantaged patients rating quality of care lower, I believe this is a very real phenomenon because they do receive less care. While I was in the hospital delivering my baby, I was shocked and appalled at the lack of attention and care that my roommate received. She was a teenage mother with no health insurance and no private doctor. They gave her a bed, but that was it. The nurses were rude and insensitive to her desire to relieve pain. When I complained to the head nurse about her treatment, I was told that she had no insurance and this was typical. The nurses knew that she had no doctor for them to answer to and that her bill would probably not be paid. This was typical treatment for those kinds of patients. In order to obtain unbiased estimates of health care satisfaction, methodological considerations that increase response from low-income and poorly insured populations must be incorporated into the survey design.

Summary

We have heard today about methodological, cognitive, and affective factors that influence measures of attitudes toward health and health care. These factors can cause both response and nonresponse bias in the resulting survey estimates. Response bias can be reduced by considering the impact of cognitive and affective theories on respondent behavior. Nonresponse bias in estimating health care satisfaction can be reduced by incorporating methodological considerations into the design to increase response from the economically disadvantaged population.

The assessment of patient satisfaction is becoming more important every year. The movement toward managed care and its impact on both cost and quality will be a determining factor in the future of health care in the United States. We should try to understand as best we can the construct of patient satisfaction and provide tools to measure it with as little bias as possible.

References

Cleary, P. D., Edgman-Levitan, S., Roberts, M., Moloney, T. W., McMullen, W., Walker, J. D., & Delbanco, T. L. (1991). Patients evaluate their hospital care: A national survey. Health Affairs, 10, 254–267.

Festinger, L. (1957). A theory of cognitive dissonance. New York: Row, Peterson.

Hoffman, M. L. (1986). Affect, cognition, and motivation. In R. M. Sorrentino & E. T. Higgins (Eds.), Handbook of motivation and cognition: Foundations of social behavior. New York: Guilford Press.

Idler, E. L. (1993). Age differences in self-assessments of health: Age changes, cohort differences, or survivorship? Journal of Gerontology: Social Sciences, 48, S289–S300.

Isen, A. M., Shalker, T. E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory and behavior: A cognitive loop? Journal of Personality and Social Psychology, 36, 1–12.

Matlin, M. W., & Stang, D. J. (1978). The Pollyanna Principle: Selectivity in language, memory, and thought. Cambridge, MA: Schenkman.

Schwarz, N., & Bless, H. (1992). Constructing reality and alternatives: Assimilation and contrast effects in social judgment. In L. L. Martin & A. Tesser (Eds.), The construction of social judgment (pp. 217–245). Hillsdale, NJ: Erlbaum.

Sehulster, J. (1994). Health and self: Paths for exploring cognitive aspects underlying the self-report of health status. In S. Schechter (Ed.), Proceedings of the 1993 NCHS Conference on the Cognitive Aspects of Self-reported Health Status (NCHS Cognitive Methods Staff Working Paper Series, No. 10). Hyattsville, MD: NCHS.

Measurement Models and Survey Research: Reliability and Validity Matter

Richard T. Campbell

Introduction

The measurement theory that most of us were taught in graduate school implicitly assumed that our goal was to develop perfectly reliable and valid measures. The mathematics of conventional reliability analysis led us to believe that Cronbach's alpha would be 1.0 if only we had enough items and validity was pursued in the form of the Holy Grail of explained variance. Even if in practice an alpha of .8 was considered a very good day's work and an R^2 of .5 was reason to stop the presses at our favorite journal, we assumed that someday, somehow, we would do better. These papers remind us, once again, that the gap between theory and practice is broad and deep and that measurement in the health arena is even more challenging than we thought. In what follows, I will discuss the three papers I have been assigned individually and then close with some general comments.

Verbrugge, Merrill, and Liu

This paper raises an extremely important question: How much detail is enough? The authors note that there have been cyclical shifts in attempts to measure health. Earlier attempts to replace the standard single item measure of self-rated health with batteries of disease specific items and elaborate scaling procedures have generally led to extremely expensive protocols that don't do much better, either as outcomes or predictors, than far simpler measures. As the measurement of specific health outcomes and conditions has consumed more and more survey time, researchers have begun to turn back to global items. Verbrugge and her colleagues ask if the situation with regard to disability measures is not the same. Is it possible, they ask, to develop a single global measure that does as well or nearly as well as more complex inventories?

To resolve this issue, both with regard to morbidity and disability, one must ask at least two prior questions: (a) What is the purpose for which the data have been collected? and (b) What is a tolerable level of error given the purpose of the survey and given the purpose of the particular constructs? For example, a survey intended to study residential mobility will certainly need health and disability information, but a survey intended to forecast the demand for medical care presumably needs greater detail. However, it is quite unusual for survey designers to have such specific purposes in mind. More commonly, particularly for large-scale, nationally representative surveys such as Asset and Health Dynamics of the Oldest Old (AHEAD), the goal is to provide a survey instrument that meets very diverse needs ranging from detailed, univariate descriptions of complex outcomes to sophisticated modeling and forecasting. The result is a tug-of-war among the various sections of the survey for space and time.

Since the authors of this paper clearly can't resolve the most fundamental issue—what are we doing and why?—they explore the validity issue in various ways. The conclusions they reach are useful and filled with common sense. I suspect, however, that these results will not resolve the issue. The authors will certainly need to bang this drum for some time to come. What follows are some things they might consider.

First, how do you compare the effectiveness of a simple summary measure to that of more complex scales? Essentially, this is a validity issue. The authors approach this by comparing the variance explained in equations containing simple and complex measures. But R^2 may not be the best criterion. Many of the outcomes they use are either dichotomous or badly skewed, in which case R^2's upper bound is not 1.0. R^2 is also affected by reliability. Most importantly, in a modeling perspective, one wants to compare the relative effects of variables. The question is, How does disability do as a predictor relative to other variables? R^2 does not answer this question directly. At the very least, one would want to compare regression solutions on various criteria other than R^2. For example, one could classify cases. This is a case in which Woodbury's Grade of Membership (GoM) model (Manton & Woodbury, 1991) might be put to good use. One could define fuzzy classes based on detailed health and/or disability items and show how the global items predict class membership.

Secondly, the paper ignores the reliability issue. Single indicators do not permit reliability computations. Thus, if one were to base an analysis on single global indicators of health and disability, it would not be possible to either assess the effects of unreliability or correct for them. I

Richard T. Campbell is a Professor in the Department of Sociology at the University of Illinois at Chicago.

would argue that the best solution would be to develop global measures based on four or five congeneric items.

Ofstedal, Lentzner, and Weeks

If the previous paper suggests that global items for morbidity and disability may do well, this paper strongly suggests that one pay close attention to reliability. Although the authors do not focus their attention on measurement error in any formal sense, the results they obtain show that the concordance between two measures of change is rather low. Thus, Verbrugge and her coauthors may be right that a single item is conceptually sufficient, but they may not have anticipated the reliability problem.

Unfortunately, the basic conceptualization of this paper is a bit muddled. The key building block is the concordance between a directly derived measure of change (from single-item reports of activity-of-daily-living [ADL] difficulties) and a measure of perceived change based on a single item. Each of the single-item direct measures of change, of course, is a less than perfectly reliable indicator of the underlying construct. Thus, the turnover table (their Table 1) between the first time point and the second overstates the degree of change—there are more discordant cases than there should be. At the same time, the perceived measure of change, which itself is based on a single item, is less than perfectly reliable. As a result, one does not know exactly how to interpret the results. Is the low level of concordance due to the unreliability of the ADL measures, the unreliability of the single item measure of perceived change, or both?

With two waves of data, there is precious little that one can do to resolve this question. True change is inextricably confounded with unreliability. However, the Longitudinal Study of Aging (LSOA) contains four waves of data and thus permits more sophisticated attempts to separate stability from change. Nearly 25 years ago, Wiley and Wiley (1970) showed how to do this using simple computations based on observed covariances, and their work has been extended in a variety of ways using formal measurement models (Bollen, 1989). Although applying these models to the LSOA data would be challenging because the data do not meet the multivariate normality assumptions of standard structural equation programs, the issue would certainly be worth pursuing. The problem is that the result would be more appropriate estimates of reliability and change (stability) with a simple structural equation model, but it is not clear how this information could be used to compute corrected estimates of concordance at the individual level. It may be that a multiwave log linear model would be an effective tool.

Disregarding the technical issues for the moment, one has to ask why it should be that we find subjects reporting increased difficulty with activities when objectively, they are better off. The results seem to be too systematic to be dismissed as the effects of random error of measurement. One suspects that the problem is that although the perception-of-change item refers to specific activities, subjects actually respond on the basis of broader changes in their health and mobility. In other words, there is also a validity issue in the perceived change measure. As the authors note, ceiling effects in the ADL measures may also present a problem.

The analysis of discordant responses that the authors provide is quite interesting. The results suggest that the quality of survey reports varies in systematic ways. If the analysis could be extended in such a way as to get a better handle on discordant results, the paper would be quite useful.

Flocke, Stange, and Zyzanski

This paper attempts to apply classical factor analysis to the problem of determining the components of primary care and relating them to specific outcomes of patient satisfaction and preventive service delivery. The authors begin by describing a theory-driven process of instrument development in which they attempted to tap seven specific dimensions of care. For each dimension, with one exception, at least three items were developed. The exception is "first contact," for which there is only one item. This process of moving from conceptualization to item development is a textbook example of how things should be done. Having developed the items, the authors collected their data and subjected the results to an exploratory factor analysis. Using the standard criterion of an eigenvalue greater than 1 (which they mistakenly call a measure of statistical significance), they found four factors. Thus, of the seven factors that the measurement instrument was designed to capture, the analysis detected four. A fifth is based only on one item, and two, comprehensiveness of care and longitudinality, were not found.

There is a disjuncture here between the theoretical concerns of the paper and the basic analysis. Based on prior conceptual and theoretical work, the authors set out to measure seven specific components of practice, carefully writing items to tap each of them. The analysis, then, should directly answer the following question: Can the observed variance-covariance matrix be described in terms of these seven underlying dimensions? Instead, the authors asked, How many dimensions are there to be found in this matrix? In other words, rather than a confirmatory analysis, the authors used exploratory methods.

This is not a mere technical issue regarding choices between alternative methods of factor analysis—it directly relates to the authors' theoretical concerns. Suppose they had fit a specific confirmatory model, specifying exactly which items were presumed to "load on" which factors. Suppose that model failed to fit the data. Further, suppose that the results showed that the items specified to load on two of the seven factors simply did not work as advertised. The authors would then have to consider two alternatives: (a) The items they chose to operationalize those two dimensions were not very good, or (b) those dimensions of

primary care do not exist. What the authors have done instead is to take the results of the exploratory analysis as an indication of what's really there. In other words, they have allowed the results of their exploratory analysis to override their original, theoretically driven conceptualization of the measurement domain.

Does this matter? I believe that it does. From a theoretical point of view, one is on much firmer ground in trying to show that the correlation structure of a set of items can be explained in terms of prior hypotheses than via an essentially post hoc analysis. It is almost always possible to explain the results of an exploratory analysis once one has seen them. But exploratory analyses typically conflate substantive and methodological phenomena. Not uncommonly, one or more of the factors result from "methods effects," that is, common question formats. Often, despite the power of confirmatory analysis, researchers find it very difficult to specify hypotheses. In the present case, the authors have done an excellent job in this regard, and I wish that they had carried through with the analysis.

Another strong feature of this paper is the authors' attempt to deal with construct validity by exploring how the patients' perceptions of primary care are related to specific outcomes. They show that controlling for age, the primary care scale scores are correlated with patient satisfaction but do not correlate with the degree to which screening and counseling services have been provided. They then ask whether there are "threshold effects" and dichotomize the outcome variables into low versus high level of services. In the course of this, they reverse the roles of the dependent and independent variables, attempting to show that the means of the primary care scales vary between cells based on age and adequacy of service delivery. A much more straightforward approach to this would be to model the odds of getting appropriate levels of preventive and counseling services as a function of the scale scores, age, and their interaction. This is easily done using logistic regression.

Survey Research and Psychometrics

Many of the points made in this discussion have to do with psychometric issues. Although survey researchers are obviously aware of reliability and validity issues, the truth is that many of the most important developments of the last 20 to 30 years have had relatively little impact. Whether one considers confirmatory factor analysis (Jöreskog, 1969), generalizability theory (Cronbach, Rajaratnam, & Gleser, 1972), or scaling approaches based on something more sophisticated than Likert scales, such as Rasch models (Rasch, 1966), survey researchers have tended to ignore psychometric innovation. This is less true for confirmatory factor analysis, but even there, the published applications are confined to a few areas, such as market research.

Why should this be true? I don't think it's because of the intellectual or moral failings of investigators. Rather, it reflects the fact that sociology departments, in which many survey practitioners are trained, have not done an adequate job of teaching these materials. In part, this reflects the fact that most of the interesting work in psychometrics is being done in psychology and education. Whatever the reason, we learn once again that interdisciplinary training is difficult and that it is easier to reward provinciality than prevent it.

References

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. Wiley.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum-likelihood factor analysis. Psychometrika, 34, 183–202.

Manton, K. G., & Woodbury, M. A. (1991). Grade of membership generalizations and aging research. Experimental Aging Research, 17, 217–226.

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49–57.

Wiley, D. E., & Wiley, J. A. (1970). Estimating measurement error using multiple indicators and several points in time. American Sociological Review, 35, 112–117.

Discussion Themes From Session 1

Elinor Walker, Rapporteur, and Daniel Walden, Chair

Measurement Themes

Patient satisfaction is not a new methodological topic for these meetings. But it is has acquired new significance in the debate surrounding the changing health care system as one domain of quality of care. However, the papers in this session point to a number of measurement difficulties and to a lack of clear conceptualization of the construct of patient satisfaction. As a result, there was disagreement regarding how the construct should be used. Some attendees considered the construct to have limited usefulness and warned of potential danger if it is used broadly for making policy determinations. Others pointed out, however, that so-called objective measures of quality have their own pitfalls. They argued that HMOs and governmental agencies will be as concerned about patient satisfaction as they are about the now current "report card" movement (initiated by the HMOs) and that serious use of such measures seems inevitable. Therefore, it is of the first order of importance to understand the construct and how to access and use it, the relationship of design of a survey to the response, and the meanings of objective and subjective measures.

Since patient satisfaction measures seem to be so consistently favorable, their use in evaluation may tend to obscure other measures of quality that reflect a more balanced assessment. The conceptualization of patient satisfaction in a survey format must take into consideration (a) the precision of measurement required, (b) the dimensions of quality that are relevant, and (c) the cognitive processes that are likely to affect how satisfaction is expressed and that might be influenced by the format and ordering of the specific questions in the survey. For example, a quality improvement effort based in a single health care delivery setting is likely to have different measurement standards than is an effort to develop a statewide or even national industry "report card." On the other hand, even in a limited assessment, the way in which the issue of patient satisfaction is raised in a questionnaire or interview will undoubtedly influence how it is reported. Therefore, the validity of such measures of satisfaction is likely to be critical in decisions about how care is delivered and probably needs to be explicitly addressed in evaluating the assessment process.

Elinor Walker and Daniel Walden are with the Agency for Health Care Policy and Research, Rockville, Maryland.

A central theme in the discussion of satisfaction was the importance of context or frame of reference for the expression of satisfaction and the timing of when it is assessed as aspects of patients' expressions of satisfaction with care provided. It was noted that there is a large literature on consumer satisfaction with services or products, of which patient satisfaction is a subset. From the standpoint of context or frame of reference, some of this literature suggests that when patients identify the dimensions of quality to be used in measurement, factors such as improvement of one's condition may not produce a commensurate increase in satisfaction. The patient's response to a question about satisfaction may be based upon a self-defined context; hence, implicit in the response may be the sense that "I am doing fine considering . . . ."

This interpretation is consistent with the results reported here by Fowler and Bin; Schwarz, Mathiowetz, and Belli; and Harris, Tierney, and Weinberger. It appears from their data that expectations and stakes may change due to the context in which they are reported. If they do, then patients may adapt standards as situations improve such that the perception of benefits may change based on experience. Hence, it becomes necessary to recalibrate expectations in the context of that experience. While it may seem expedient to reduce or enhance the standard, this imparts artificiality. If what you want is to capture what would be said in a conversation with a spouse or associate, it may be necessary to allow for the fact that in such contexts, the subject establishes the frame of reference. If the frame of reference is manipulated, the results may have no relationship to the reality in which they are viewed by the patient. Thus, if patients are interviewed after a hospital stay, the patient may infer a context, and the response may mean, "I'm doing fine considering the fact that I have just been discharged from the hospital or considering how I was doing when I entered the hospital."

The issue of time also becomes important in the response. If context is important, then it must also be assumed that patient reports about satisfaction contain retrospective as well as prospective elements. Hence, as with any type of question for which time is relevant, such questions may need to provide the patient with a time reference point in the interview at which to "chunk" or segregate the response in terms of then versus now. But in making measurements over time, there is the risk of confounding true change (or stability) and measurement

error. There is a literature on this topic. Also, many issues such as time and context can be looked at via multitrait-multimethod analysis, as in Mulaik (1972).

Time is an issue also in terms of the episodic nature of many conditions, whose fluctuations may have no temporal relationship to the intervals of a longitudinal survey and may not be captured by it. A related concern is the question of what is, indeed, an objective measure. For example, instrumental activities of daily living (IADLs) are not considered objective because they are not reliable over time. The paper by Ofstedal, Lentzner, and Weeks attempts to address this problem using data from the Longitudinal Study of Aging. However, the authors base their analysis on only two waves of data and thus run into problems separating reliability from change. There is a need to carefully design research that is capable of making these kinds of distinctions if the measurement issues in this area are to be addressed. Thus, differences over time may be noise but may also indicate real change. This issue of reliability versus stability of measures over time has been in the literature for many years (Wiley & Wiley, 1970).

Again, the frame of reference could be built into the question. Another useful tool might be using several waves of data, as suggested by Campbell, or reinterviewing during a particular wave to measure unreliability. There is a considerable literature on measuring gross flows in the labor force that might be pertinent (Abowd & Zellner, 1985; Chua & Fuller, 1987; Fuller, 1989; Kalton, Kasprzyk, & McMillen, 1989; Rodgers, 1989; Skinner, 1993).

General measurement issues similar to those associated with patient satisfaction also arose in relation to the issues raised by Verbrugge around the notion of parsimony in measures of patient status. The discussion focused on the tension between simplicity and precision and was prompted in part by Campbell's discussion of Verbrugge, Merrill, and Liu's paper and his introduction of the caveat that a simple solution may be a poor solution. Her response that "one dandelion may provoke a host of blooms" referred again to the issue raised in relation to satisfaction, namely that measurement precision should be viewed in terms of the function of the survey. A multipurpose study that intends to determine the proportions of the population needing different kinds of services may require more detailed measurement than does a study attempting to determine the conditions leading to any service need for which a global measure may suffice. The more detailed measure may even be a more elegant solution in the case of a multipurpose study. Even when the data collection serves multiple purposes, it would be useful to consider including one global item, even if specific indicators are also employed (again raising the specter of expansion).

When comparing global indicators with specific indicators, examination of R^2s takes into account only one statistical tool; classification procedures such as those employed by Manton and Woodbury (1991) may also be useful, although the programming is not user-friendly. It might also be of interest to examine the variables in terms of sensitivity/specificity or predictive value.

In measuring concepts like satisfaction, quality of care, and disability, there is a need to consider how current approaches to survey methods can be enhanced by psychometric theory. There is a tension, however, between psychometric measurement requirements and the need for parsimony to reduce respondent burden and overall survey costs. There followed discussion of a specific recommendation that noted that psychometric techniques are often lacking in the design of surveys. The disagreement focused on whether psychometric techniques could indicate to the designers what a question would mean. There was skepticism about whether psychometricians would be of much help because of the issues of parsimony and question meaning. A firm opposing position emerged, however, that there have been many advances in psychometrics over the last 25 years (see Campbell's discussion paper) that have not come into use in survey work and which could, used in conjunction with standard survey design methods, address some of the issues raised in this section. The absence of integration of relevant psychometric theory into survey design results from training psychometricians in isolation from other methodologists in psychology departments. It was also noted that measurement theory and survey design need to be incorporated into continuing education for physicians who may be using concepts such as these but may lack contact with relevant methodologists.

Design and Sampling Themes

Several members of the audience spoke about nested sample designs with reference to the Flocke, Stange, and Zyzanski presentation. In that paper, the sample is first of physicians, then of patients visiting physicians on particular days—thus, of patient visits. It was noted that patients making many visits have a different probability of being included than patients making fewer visits. Patients making more visits are also likely to be different from those making fewer visits. Patients making more visits might, for example, be more likely to be up-to-date in preventive measures, may be more or less satisfied with the care they receive, may be sicker, may experience more pain or other sequelae of diseases. Moreover, several patients may have the same physician; therefore, even observations made on different patients will be conditional upon sharing a common physician. Observations based on designs of this type may, therefore, be unrepresentative due to clustering of patient reports around common physicians as well as the frequency of physician contact. This issue can be addressed with hierarchical linear modeling (HLM) or multilevel modeling techniques (Bryk, & Roudebush, 1992; Hedaker, McMahon, Jason, & Salina, 1994). The techniques are widely used in educational research to address the analogous problem of separating classroom and/or school effects from intervention effects on pupils.

Another issue raised by the Flocke et al. paper is the possibility that if the sampling dates are not randomly assigned, physicians might vary their management of patients on those days when they know patients are being interviewed. A similar issue arises with hospitalization follow-up studies, in which patients are enrolled while still in the hospital, as might be the case in the Fowler and Bin study. Blinding of the participating physicians regarding the manner of sample selection and blinding of the patients regarding the purposes of the study would be recommended to prevent these types of bias. In the case of the patient, satisfaction as an evaluative measure creates a demand situation, in which providers and evaluators may focus on this measure to the detriment, perhaps, of clinical aspects of quality.

Themes to Be Pursued in Future Research

1. There is a need for further work on conceptualizing satisfaction as a component of quality of care. It appears to be a multidimensional concept, dependent upon both time and context. Both the cognitive dimensions and the measurement issues require further specification.
2. In measuring concepts such as satisfaction, quality of care, and disability, there is a need to consider how current survey approaches can be enhanced by measurement theory and psychometrics. New psychometric approaches are not well integrated into survey methods and could be particularly beneficial on these topics.
3. Several strategies need to be considered for measurement of disability and similar concepts. Global measurement of such concepts seems intuitively interesting, but problems have been found in the cognitive research on these types of concepts (Krause & Jay, 1994). There appear to be methods to evaluate these approaches that need to be considered in the context of classification strategies and other aspects of measurement theory (cf. Manton & Woodbury, 1991). In general, there is a need to reconsider ways to resolve the tension between parsimonious measurement and psychometric theory.
4. In addition to issues of validity, there is a need to consider reliability and change. Some of the new longitudinal surveys provide opportunities for sorting out differences between unreliability and change. More attention needs to be given to these issues, especially around questions of using IADLs and activities of daily living (ADLs) as outcome measures.
5. There is a difference between understandable presentation of results of this kind of research and the complexities of measurement. Attention needs to be paid to both aspects.
6. Research and policy needs require further examination of the purpose of measurement of medical care and health status. Monitoring service needs at a particular health institution requires different measurement strategies from assessing health care delivery and client needs at the state or national level. The "report card" approach needs to be carefully considered in this context.
7. Design and analysis plans for studies of patient satisfaction need to include consideration of cluster effects due to frequency of contact with the provider and contacts by several patients with the same provider.
8. The demand effects of these kinds of studies also need to be carefully evaluated before the results are used for policy decisions.

References

Abowd, H. M., & Zellner, A. (1985). Estimating gross flows. Journal of Business and Economic Statistics, 3, 254–283.

Bryk, A. S., & Roudebush, S. W. (1992). Hierarchical linear modeling: Applications and data analysis. Thousand Oaks, CA: Sage.

Chua, T. C., & Fuller, W. A. (1987). A model for multinomial response error applied to labor flows. Journal of the American Statistical Association, 82, 46–51.

Fuller, W. A. (1989). Estimation of cross-sectional and change parameters: Discussion. D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), Panel surveys (pp. 480–485). New York: Wiley.

Hedaker, D., McMahon, S. D., Jason, L. A., & Salina, D. (1994). Analysis of clustered data in community psychology: An example from a worksite smoking cessation project. American Journal of Community Psychology, 22, 595–615.

Kalton, G., Kasprzyk, D., & McMillen, D. B. (1989). Nonsampling errors in panel surveys. D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), Panel surveys (pp. 249–270). New York: Wiley.

Krause, N. M., & Jay, G. M. (1994). What do global self-rated health items measure? Medical Care, 32, 930–942.

Manton, K. G., & Woodbury, M. A. (1991). Grade of membership generalizations and aging research. Experimental Aging Research, 17, 217–226.

Mulaik, S. (1972). The foundations of factor analysis. New York: McGraw Hill.

Rodgers, W. L. (1989). Comparisons of alternative approaches to the estimation of simple causal models from panel data. D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), Panel surveys (pp. 432–456). New York: Wiley.

Skinner, C. J. (1993). Logistic modeling of longitudinal survey data with measurement error. Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys (pp. 269–276). Statistics Canada.

Wiley, D. E., & Wiley, J. A. (1970). Estimating measurement error using multiple indicators and several points in time. American Sociological Review, 35, 112–117.

Research on Survey Questions

The six papers in this session focus broadly on survey question design and data quality. These themes were introduced in the preceding session but there, they focus on issues related to measuring satisfaction with medical care and assessing health status, particularly disability. The papers in this session focus more generically on the design of health questions. There are two general themes that run through the papers in this session. The first is how cognitive information processing can be facilitated by designing questions that address this process. The second issue focuses on the interaction among the respondent, the interviewer, and the interview form and how elements of these interactions affect the resulting data quality. As noted, these themes are present in the preceding session, but the discussion centers around psychometric issues and, particularly, question reliability and validity rather than on the interaction between the respondent and the survey process, which is the main focus of this session.

Measuring and Improving Data Quality in Children's Reports of Dietary Intake

Karin A. Mack, Johnny Blair, and Stanley Presser

Introduction

The goal of this pilot study was to suggest alternative questionnaire approaches for collecting information from children for the U.S. Department of Agriculture (USDA) Human Nutrition Information Service's Continuing Survey of Food Intakes by Individuals (CSFII). The CSFII is a national survey of the household population which obtains—along with other information—reports of all foods eaten by each household member on the day prior to the interviewer's visit. The instrument used in the CSFII 1989–91 is called the Day One Individual Intake Record. For children under the age of 12, Day 1 data were obtained from a proxy respondent, usually a parent. Children, however, may frequently have meals or snacks when the parent is not present, such as at school, in day care, or while visiting friends or relatives. In such instances, the parent often cannot accurately report some or all of the things the child ate that day. This produces substantial problems of missing or incorrect data for children, especially those of school age. Our project focused on the possibility of having children serve as respondents to the CSFII. The major research objective was to design approaches to structuring a survey instrument for obtaining dietary reports from children aged 6 through 11 that USDA could test in more rigorous fashion.

The Development of Three Test Protocols

The 1991 USDA Day 1 questionnaire, which was designed for adult respondents, used a series of closed-response items in which all "eating occasions" on the reference day were asked about in chronological order. For each time the respondent ate something, he or she was asked the name of the eating occasion, who else was there, what was eaten, and details about the nature of each food item and where it was obtained. Our review of the Day 1 questionnaire suggested that the language was too stilted for children (and perhaps for adults as well), that the sentence construction was too complex, and that the strategy for recall was likely to be problematic for children.

Based on a literature review and pretesting, we developed three protocols as alternatives to the closed, highly structured, chronological Day 1 format. These three protocols, called open, meal, and location, were designed to investigate different notions about how children's recall might be aided, while also making the interview situation less forbidding.

In the open protocol, children were allowed to report foods eaten the previous day without any imposed structure. Indeed, it was developed to be quite the opposite of the CSFII Day 1 instrument. The children were able to choose the pattern of reporting they preferred, without the possibly inhibiting task of answering a series of formal questions. Children were given an introduction and then asked, "Now tell me all the things you ate or drank yesterday." Interviewers were instructed to probe for details concerning food items consumed. We developed this format in response to several concerns raised in the literature. For example, Wood and Wood (1983) found that children's length of response decreases the more frequently they are questioned. Additionally, it has been suggested that children may not have a structured sense of their day and may therefore be more comfortable reporting in a free-form format (Medrich, Roizen, Rubin, & Buckley, 1982).

The second protocol followed a meal/nonmeal format. Based on research by Rasanen (1979) and Frank, Berenson, Schilling, and Moore (1977), we hypothesized that foods eaten may be organized in memory by the schedule of regular meals. If so, then an interview organized in the same way may be effective for aiding recall. The meal/nonmeal instrument asked directly about each traditional meal of the day: breakfast, lunch, and dinner. It began, "Did you eat or drink anything before breakfast yesterday? . . . Did you have breakfast yesterday? . . . What did you have for breakfast?" It also asked specifically about eating between breakfast and lunch, between lunch and dinner, and after dinner.

The third protocol, location, used the child's activities and locations on the previous day as the basis for asking about foods eaten. Reporting what they did on the previous day may be a more natural and engaging task for children than trying to remember a list of foods eaten and may serve as a good memory trigger for those foods (Baranowski & Domel, 1992). The food reports were then obtained as a

Karin A. Mack is an Assistant Professor in the Department of Sociology, Anthropology and Social Work, Mississippi State University, Mississippi State. Johnny Blair is Associate Director and Stanley Presser is Directo of the Survey Research Center, University of Maryland, College Park.

component of each activity. After an introduction, the interview began with the statement, "I'd like you to start with when you got up yesterday and tell me each place you were." After the child reported where they had been, several follow-up questions and probes were used to find out whether the child ate or drank anything at that location.

We used the 1991 USDA Day 1 questionnaire as a control instrument.

The Sample

Children were recruited for the study by two methods. Some were recruited from nearby day care centers. In addition, telephone directory numbers were screened to identify households with children aged 6 through 11. Two rounds of interviews were conducted. Thirty-six children were interviewed in the first round, which was a 4 × 2 design, with nine cases for each of the interview protocols, divided about evenly between children aged 6 through 8 and those aged 9 through 11. Nine children and their parents were interviewed in the second round, which also included a joint parent-child reconciliation interview. All interviews were audio- and videotaped. We also conducted postinterview debriefings (described below). Observation data for validation were collected for 27 children by trained observers who recorded the midday meal eaten by the children at day care centers.

Children were paid $5.00 and parents $20.00 for participating in the study. Child care centers and community centers were compensated $50.00 for granting access to the children. Interviews ranged in length from 9 minutes to 35 minutes, with an average length of 17 minutes. After each interview, a different interviewer asked debriefing questions of the child, of the parent (where applicable), and of the child's interviewer.

Interview Structure Additions

In addition to developing the three protocols discussed above, we tested other methods of adapting the interview situation. In pretesting, we tried out think-alouds and a drawing exercise. In the think-alouds, the children were asked not only to answer the survey questions, but to report how they came up with their answers. For the drawing task, the children were asked to draw or list the food items they reported eating. They were given large sheets of paper and a set of colored markers. Because the combination of drawing and thinking aloud seemed difficult to do simultaneously, the think-alouds were dropped prior to the first round of interviews. Since the drawing exercise did not yield the expected results in round one, it was dropped for the second round of interviews, and the think-alouds were reinstated.

In the warm-up phase of the interview, children were asked about polls and surveys. Each child was asked, "Tell me a little about what you've heard about polls and surveys. Why do you think polls and surveys are done?" The overwhelming majority of children (30 of 36) were not familiar with polls or surveys. Of the 6 children who had heard of surveys/polls, 1 responded that they were like Family Feud! All of the children except one answered, "I don't know," to the question of why they thought polls and surveys were done.

A second part of the process, used in both rounds, was an extensive debriefing session. Children were asked whether they liked the interview, whether anything was hard to understand, what they did when they had trouble remembering what they ate the day before, if there were any items they thought they weren't supposed to mention, and whether they had left anything out of the interview.

All the children reported that the dietary recall instructions were easily understood and generally said that they reported their intake from the previous day and did not confuse yesterday with the day before or the current day. When probed for a response they could not recall, children either said things like they "thought back and just tried to remember what [they] actually ate" or "guessed"; said what they usually ate; or simply said, "I forget."

The children's behavior and interaction with the interviewer were coded from the videotapes.[1] Here, we focused on the child as a survey respondent in general, rather than on differences by protocol. We found that the younger children were more likely to express difficulty or to qualify their answers, while the older children were more likely to elaborate on their answers. The younger children were more likely to smile and make eye contact with the interviewer. The younger children were also more likely to show signs of confusion or distraction.

Evaluation of Designs

We compared the effectiveness of the protocols along three dimensions: (a) the completeness of the reports, (b) the ease of administration, and (c) the children's reactions.

Completeness of the reports focuses on the total number of food items reported, as well as the accuracy of reported items. Table 1 shows the number of food items reported by the children for each protocol and for the two age groups. The items reported are broken down by meal, if that information was given, or are listed in the "no meal named" row if the meal could not be determined. Each time the child reported something eaten, it was counted as one item; for example, a hamburger is one item, a milk shake is one item, and so on. Condiments on the hamburger do not count as individual items. The table shows that the Day 1 interviews yielded the lowest average number of items (10.3). The highest average number of items reported

[1]Details of this procedure are available in Presser, Blair, Mack, Ryan, and Van Dyne (1993).

Table 1. Number of food items children reported by protocol and by age

	Day 1 (n = 9)	Open (n = 9)	Meal (n = 9)	Location (n = 9)	Ages 6–8 (n = 19)	Ages 9–11 (n = 17)
Breakfast	19	8	20	14	32	29
Lunch	22	10	22	27	44	37
Dinner	19	9	31	29	54	34
Snack	17	4	15	4	25	15
Total for meals named	77	31	88	74	155	115
No meal named	16	71	17	39	70	73
Total items	93	102	105	113	225	188
Average per child	10.3	11.3	11.7	12.6	11.8	11.1

(12.6) came from the location protocol. The younger children reported slightly more items on average than the older children (11.8 compared with 11.1). Though these are not large differences, we note that all of the alternative protocols did better in total reporting of items than the Day 1 protocol.

For a subsample of children in round one, observers visited the child's day care center and listed the foods eaten during one meal. The children were not aware that their meal was being observed. The children were then interviewed the following day by an interviewer who was unaware of the observational data. Table 2 compares the children's reports for one meal, either lunch or a snack, with the observer's record of that meal. For this partial day report, the location protocol yielded the greatest accuracy, with 58% of the children's reported items matching the observer's record. The open protocol was roughly equal at 57% of the items matching. The meal protocol was the lowest at only 30% matching. The Day 1 instrument matched 50% of the items recorded. The older children were more accurate than the younger children, with 55% of their items matching compared with 44% for the younger children.

Table 3 compares the children's reports with the parents' reports. The Day 1 protocol yielded the lowest match rate, with only 31% of the children's items matching the parents' reports. The highest agreements were in the open and the meal protocols, with 72% of the items matching. The location protocol fared less well, at 43%, though still better than the Day 1 protocol. The older children were more in agreement with the parents' reports, with 64% of their reported items matching, compared with 37% of the younger children's reports matching.

Table 2. Comparison of reported food items to observer's recorded food items for the meal observed

	Day 1 (n = 6)		Open (n = 7)		Meal (n = 7)		Location (n = 7)		Ages 6–8 (n = 15)		Ages 9–11 (n = 12)	
	Child	Observer	Child	Observer	Child	Observer	Child	Observer	Child	Observer	Child	Observer
Total items reported	12	14	6	23	13	20	13	19	30	48	14	31
Matches	7		13		6		11		21		17	
% items matched	50		57		30		58		44		55	

Table 3. Comparison of children's reports to parents' reports by food items matched and not matched

	Day 1	Open	Meal	Location	Ages 6–8	Ages 9–11
Total items child	25	23	18	29	45	50
Total items parent	36	18	18	21	49	44
Child reported, parent did not	14	10	5	20	27	22
Parent reported, child did not	25	5	5	12	31	16
Matches	11	13	13	9	18	28

In the second round of interviewing, following each individual interview, a joint interview was conducted in which the child, the mother, and the child's interviewer discussed the items reported in their individual interviews and agreed on a common joint list. The interviewer went through the entire list, noting agreement as well as disagreement. There are two areas of interest from the reconciliation interviews: the joint interview methodology itself and the findings comparing parent and child reporting.

Methodologically, the joint parent-child interview was potentially threatening, requiring the child to admit errors in front of the parent. We used several procedures to reduce the potential threat. To help put the child at ease, we used the child's interviewer to conduct the joint interview. We also prefaced the joint interview with a statement that "mothers and children do not always mention the same food and beverage items" and that our objective was simply "to see where they were similar and different" and to determine the likely cause of any differences. We stressed that the objective was not to find out who was right and who was wrong. These strategies seemed to be effective in putting the child at ease and avoiding, for the most part, confrontational interaction.

Table 4 shows that some of the foods on the joint list were reported only by the child (30% of total items listed in the joint interview) and others were reported only by the parent (37.5% of the total). Overall, of the items reported by parents, 78% were accurate (were on the joint list), as compared with 72% of the items reported by children. This underscores the concern that the parent interview cannot be taken as truth. The fact that the error rates were only slightly different shows that reliance solely on the child's report would not be greatly different from reliance only on the parent's report.

One way to measure the ease of administration is to compare the amount of time each interview required. The open format, on average, produced the shortest interviews, and the Day 1 interviews were the longest. Debriefings with the interviewers suggested that the meal and open formats were the easiest to administer. The interviewers also felt that these two protocols were the most agreeable formats to the children. The Day 1 interview was considered too cumbersome and difficult for the children to understand, and the location format appeared to generate a fair amount of discussion about things other than foods eaten.

Conclusions

Although we can reach no firm conclusions with these small samples, manipulation of interview structure appears to influence both the amount of reporting and its accuracy. Although the differences are small, more food items, on average, are reported in all three alternative protocols than in Day 1 interviews.

The results are less clear when child reports are compared with those from another source, though it does appear that our protocols also improve accuracy. All three of the developed protocols performed better for reports about the entire day than the Day 1 protocol in the parent-child comparison. For the observer-child comparison, the Day 1 questionnaire performed better than only one of the developed protocols, the meal protocol. But that comparison was only for one meal during the day, and the goal of the CSFII is to obtain a complete report for the entire day.

When comparing children's reports with observer reports, we take the observer report as the true measure. If we take the parent report as the true measure, children do much worse in Day 1 (31% matched) than in the alternative protocols (43% for location and 72% for both open and meal). This is a useful comparison, since the 1989–91 USDA survey interview rules accept the parent's proxy report as the true measure. It suggests that under some conditions, the quality of the child's report approaches that of the parent's proxy report. Further work on these alternative protocol structures seems justified. As noted at the outset, however, the parent may not know about some of the things the child ate during the reference day. So while it is useful to use the parent report as a base, we must keep in mind that parent reports are also subject to error. This is suggested by the relatively large number of instances in which the child accurately reported food items not reported by the parent. More puzzling, however, were the instances in which the parent reported items not mentioned by the

Table 4. Comparison of food items reported individually to joint list (total items = 112)

	Total reported by parents		Reported only by parents		Total reported by children		Reported only by children	
	No. items	%	No. items	%	No. items	%	No. items	%
Total items discussed	98		42		90		34	
Correct	87	89	31	74	81	90	25	74
Incorrect	11	11	11	26	9	10	9	26
	Total correctly reported by parents				Total correctly reported by children			
Joint list total	87/112 = 78%				81/112 = 72%			

child, which also happened frequently. These reports may have resulted from the parent either assuming the child had eaten something the parent gave the child or resorting to reporting usual eating behavior. Finally, the results seem to indicate that older children do better at reporting than younger children, even though the younger children reported slightly more items per day. Our results suggest that interview structures may have important effects on young respondents. Given the number of mismatches between the parent and child reports, a fuller examination of child reports and child and parent comparisons appears to be a fruitful avenue to pursue.

References

Baranowski, T., & Domel, S. (1992). A cognitive model of child's reporting of food intake. Paper presented at the First International Conference on Dietary Assessment Methods, St. Paul, MN.

Frank, G., Berenson, G., Schilling, P., & Moore, M. (1977). Adapting the 24-hr. recall for epidemiologic studies of school children. Journal of Clinical Nutrition, 71, 31–35.

Medrich, E., Roizen, J., Rubin, V., & Buckley, S. (1982). The serious business of growing up: A study of children's lives outside school. Berkeley, CA: University of California Press.

Presser, S., Blair, J., Mack, K., Ryan, C., & Van Dyne, M. A. (1993, August). Final report on the University of Maryland-USDA cooperative agreement to improve reporting for children in the Continuing Survey of Food Intakes by Individuals. Unpublished report, Survey Research Center, University of Maryland.

Rasanen, L. (1979). Nutrition Survey of Finnish Rural Children. American Journal of Clinical Nutrition, 32, 2560–2562.

Wood, H., & Wood, D. (1983). Questioning the preschool child. Educational Review, 35(2), 149–162.

Cultural Variations in the Interpretation of Health Survey Questions

Timothy P. Johnson, Diane O'Rourke, Noel Chavez, Seymour Sudman,
Richard B. Warnecke, Loretta Lacey, and John Horm

Background

The United States is rapidly evolving into a culturally heterogeneous society. Within the next 50 years, it is likely that those ethnic and racial groups currently identified as minorities will collectively represent more than half of the country's population. This demographic transition is likely to challenge the survey research community to reassess long-held assumptions regarding how survey questions are developed, administered, and analyzed. Central to these concerns will be the issue of similarities and differences in the validity of survey data collected across multiple cultural groups. For several years, questions have been raised regarding the appropriateness of uncritically applying traditional survey research methods to the study of culturally diverse populations (Aday, Chiu, & Andersen, 1980; Milburn, Gary, Booth, & Brown, 1991). It has been further suggested that some of the commonly reported cultural group disparities in health-related indices found in the United States may in fact be attributable to culturally mediated differences in perceptions of the meanings of health-related survey questions (Andersen, Mullner, & Cornelius, 1987; Angel & Thoits, 1987).

These problems may be understood in terms of the distinction between "etic" and "emic" constructs in social inquiry (Berry, 1969). Concepts with a shared meaning across many cultures are considered to be etic in nature. In contrast, concepts and ideas that are unique to a given culture, or which vary considerably in meaning across cultures, are defined as emic. When concepts that are emic within a researcher's culture are employed in cross-cultural studies and assumed to be universally understood, they are referred to as "pseudoetic" (Triandis, 1972). This practice results in what Kleinman (1977) refers to as "category fallacy," the assumption that survey questions are being comprehended and interpreted in an equivalent manner by all respondents, irrespective of cultural values, norms, and experiences. In this paper, we investigate (a) the degree to which a sampling of survey questions routinely used in national health studies is comprehended and interpreted in an etic (i.e., consistent) versus an emic (i.e., differential) manner across four distinct cultural groups residing in the United States and (b) the degree to which any variability of interpretation influences substantive findings.

Timothy P. Johnson and Richard B. Warnecke are with the Chicago office of the University of Illinois at Chicago's Survey Research Laboratory. Diane O'Rourke and Seymour Sudman are with the Urbana office. Noel Chavez is with the University of Illinois at Chicago's School of Public Health. John Horm is at the National Center for Health Statistics in Hyattsville, Maryland. We wish to dedicate this paper to the memory of our colleague Loretta Lacey.

This study was funded by award number U83/CCU508663 from the National Center for Health Statistics.

Methodology

The research presented here is part of a larger inquiry into cultural differences in social cognition (Johnson et al., 1995). The study population consisted of the four largest cultural groups in the Chicago metropolitan area: African Americans, Mexican Americans, Puerto Ricans, and non-Hispanic whites. The two largest Latino communities were selected in recognition that although there is a core culture common to all persons of Hispanic origin, there are also many differences (Marín & Marín, 1989). A total of 423 adults aged 18 through 50 participated in laboratory interviews with structured probes conducted by a research team of investigators and research assistants. Respondents were stratified such that approximately one-quarter were representatives of each culture (111 African Americans, 112 Mexican Americans, 92 Puerto Ricans, and 108 non-Hispanic whites were interviewed). Respondents were further stratified by gender, age (18–30 and 31–50), and education (high school or less and more than high school). Respondents were recruited using several means, primarily through media ads and community organizations.

Substantive questions included in the survey instrument were selected from a large pool of health questions previously used in national health surveys. Items were chosen to produce variation in terms of question topics and formats. As part of the planned laboratory interview protocol, sets of specific probes designed to obtain insights into the underlying cognitive processes used by participants when answering the substantive questions were developed for use with each question (Belson, 1981; Bradburn, Sudman, & Associates, 1979). Unique probes were designed to examine various cognitive tasks, including question interpretation, memory retrieval, judgment formation/response formatting, and

response editing. Interpretation probes, which will be the focus of this presentation, examined several dimensions of respondent comprehension, including the meaning of survey questions as a whole, the meaning of specific words and phrases within questions, terminology preferences, and perceived difficulty of understanding. Responses to 21 health questions were probed to investigate respondent interpretation.

Interviews were completed between July 1993 and April 1994, conducted in English, and averaged approximately 1 hour in length. With respondent consent, each interview was tape-recorded and transcribed. During transcription, responses to unstructured probes were reviewed by several members of the research team and assigned codes representing the content of each respondent's answer. For most probes, multiple codes were used in an effort to capture as much information as possible from each respondent's answer.

Results

Our first objective was to determine whether cultural variations in question interpretation could be identified empirically. Of the 21 substantive health questions for which respondent interpretation was assessed, 18 provided evidence of cultural differences after controlling for the effects of other respondent characteristics, including gender, age, education, and income. Our second objective was to evaluate the extent to which cultural differences in responses to health questions may be a consequence of differential interpretation. Of 17 questions that could be assessed for these linkages, 8 provided such evidence. Because of space limitations, we present information only from a subset of these findings.

Global Health Ratings

One of the most commonly used health survey items in the United States is the global health rating question, "Would you say your health is excellent, very good, good, fair or poor?" After answering this question, respondents in our study were asked, "In answering this question, what kinds of things did you think about?" Responses were classified along five health dimensions (Krause & Jay, 1994): health problems (e.g., the presence or absence of various health conditions), health behaviors (e.g., their presence or absence), general physical functioning (e.g., references to physical condition or energy level), health comparisons (to previous self or to others), and mental health (e.g., references to positive or negative moods or symptoms). No cultural differences were found in the frequency with which respondents referenced four of these dimensions (see Table 1). African Americans, however, were less likely than other respondents to make references to health comparisons.

We subsequently estimated regression models to assess the independent effects of respondent culture on references to health comparisons and the effects of health comparison referents on global health ratings. A logistic regression model (see Table 2, column 1) indicated that after controlling for other demographic characteristics, Mexican American and non-Hispanic white respondents (in contrast to African Americans) remained more likely to compare themselves to other persons when thinking about their health. Perhaps not surprisingly, older persons were also more likely to report having made health comparisons when answering this question.

Column 2 of Table 2 presents a multiple linear regression model that regressed the same set of demographic indicators, along with the indicator of having made health comparisons, on global health ratings. In this model, comparing one's health to that of other persons was found to produce less positive self-evaluations, net of other variables. Consistent with previous studies, education and income were also predictive of health rating. When examined together, these regression models suggest that culture may indirectly affect global health assessments by influencing how individuals think about their health.

Table 1. Responses to global health rating probe by cultural group (percentages)

Question: "Would you say your health is excellent, very good, good, fair or poor?"

Probe: "In answering this question, what kinds of things did you think about?"

Concept mentioned	African American (n = 109)	Mexican American (n = 114)	Puerto Rican (n = 90)	Non-Hispanic white (n = 108)
Health problems	63.3	70.2	68.9	67.6
Health behaviors	46.8	45.6	42.2	49.1
General physical functioning	22.0	17.5	16.7	21.3
Health comparisons**	0.9	10.5	5.6	11.1
Mental health	8.3	12.3	8.9	9.3

NOTE: Percentages do not sum to 100% due to multiple answers.

**p < .01.

Disease Labeling

Cultural differences in the use of labels for two chronic diseases, hypertension and diabetes mellitus, were also examined. For each disease, respondents were shown a short list of synonyms and asked to identify which label they would be most likely to use in discussing this condition. The most commonly selected label for hypertension was "high blood pressure," endorsed by 78.8% of our sample (17.3% selected "hypertension," 2.3% selected "high blood," and 1.6% volunteered some other term). African Americans were more likely than other groups,

Table 2. Regression analyses of global health ratings and probe (n = 411)

Question: "Would you say your health is excellent, very good, good, fair or poor?"

Probe: "In answering this question, what kinds of things did you think about?"

	Made health comparisons (1 = yes)[a]	Health rating (1 = poor; 5 = excellent)[b]
Made health comparisons (1 = yes)	—	–0.37*
Male (1 = yes)	–0.07	0.00
Mexican American (1 = yes)	1.34**	–0.06
Puerto Rican (1 = yes)	0.91	–0.24
Non-Hispanic white (1 = yes)	1.23*	–0.06
Age	0.06**	0.01
Education	0.29	0.21***
Income	–0.08	0.09**
Model X^2	25.16***	—
F value	—	8.83***
R^2	—	0.15

[a]Logistic regression model.

[b]Multiple linear regression model.

*p < .05. **p < .01. ***p < .001.

particularly those of Hispanic origin, to prefer the term "hypertension" (see Table 3). These findings were in part confirmed in a logistic regression analysis (not shown) that revealed Mexican American respondents to be less likely to endorse "hypertension" than were African Americans, after controlling for other variables. This model also revealed that females, more educated respondents, and persons with the disease were more likely to prefer the term "hypertension."

A similar analysis of label preferences for the disease diabetes mellitus revealed a large majority (88.3%) preferred the term "diabetes." The total proportion selecting other labels, including "sugar diabetes," "sugar," "high sugar," or something else, was 11.7%. The proportion endorsing "diabetes" varied across cultural groups (see Table 3). African Americans were most likely to employ an alternative label. A logistic regression model (not shown) confirmed that African Americans were more likely than non-Hispanic whites to prefer labels other than "diabetes." These findings support the notion that culture influences the terminology used to describe common medical conditions.

Health Care Access

Cultural differences in the interpretation of health care access questions were also investigated. One item from the National Health Interview Survey asks about health care visits: "During the last year, how many times did you see or talk to a medical doctor?" Respondents in our study were

Table 3. Responses to disease label probes by cultural group (percentages)

Probes

[a] "Next I'd like to ask you about a disease that is referred to by many names. Please look at this card and tell me which name you would call it. [diabetes, sugar diabetes, sugar, high sugar, something else (specify]."

[b] "Now please look at this card and tell me which name you would actually call this problem. [high blood pressure, hypertension, high blood, something else (specify)]."

	African American	Mexican American	Puerto Rican	Non-Hispanic white
[a] Selected "diabetes"*	81.7	90.3	87.6	93.6
n	109	113	89	109
[b] Selected "hypertension"*	25.7	11.3	13.1	18.8
n	101	106	84	101

*p < .05.

probed regarding the types of doctors they thought about in answering this question. In addition to physicians, responses included a variety of other health care providers, including dentists, chiropractors, psychologists, optometrists, physical therapists, and social workers. The proportions citing these other types of health care providers, by cultural group, were used to assess potential group differences in the definition of "medical doctor." Overall, non-Hispanic whites were more likely to cite at least one nonphysician provider than were minority group respondents (22.0% vs. 14.7% of African Americans, 11.6% of Mexican Americans, and 7.8% of Puerto Ricans). A logistic regression analysis (not shown), however, indicated that once respondent income was taken into account, cultural differences in having considered nonphysician providers were eliminated. This example reveals apparent cultural differences in the interpretation of a common health care survey item to be in fact a reflection of greater access to various (nonphysician) health care professionals among persons with the economic resources to afford them.

Health Behaviors: Physical Activity

Respondent interpretations relevant to health behaviors were also investigated. The example we present is concerned with respondent definitions of "physical activity." After answering the question, "What types of physical activity or exercise did you perform during the past month?" respondents were read several examples and asked if they would or would not consider each to be a physical activity. Each of these items was designed to assess the

boundaries of the respondent's definition of physical activity. Bivariate analyses identified cultural differences in responses to two of the four examples (see Table 4). There was relative consensus across cultural groups that walking and work-related activity could be defined as physical activity. Non-Hispanic whites, though, were less likely to agree with the three minority groups that housework and yard work were examples of physical activity. These differences were generally confirmed in a series of logistic regression models (not shown). These analyses also revealed significant gender differences: Males were less likely to consider walking, housework, and yard work to be examples of physical activity and more likely to consider work-related activities as such. Work-related activities were less likely to be thought of as physical activity by persons with greater incomes and, presumably, more sedentary occupations.

Nutrition Questions

Similar methods were used to examine cultural differences in conceptualizing two nutrition questions. For example, respondents were asked, "About how many times do you eat potatoes per day or per week?" followed by structured probes that asked, "In answering this question about potatoes, did you think about: [a] French fries or any frozen potatoes; [b] potato chips?" These results, shown in Table 5, again revealed significant cultural differences. African Americans were most likely to have thought about both French fries/frozen potatoes and potato chips when reporting their potato consumption. Logistic regression analyses (not shown) confirmed the finding that African Americans were more likely than others to have considered these specific food products. These models also revealed that males were more likely to have thought about French fries and more educated respondents were less likely to think about potato chips when considering their potato consumption. Consideration of these food products, however, did not influence reported frequency of potato consumption.

Table 4. Responses to physical activity probe by cultural group (percentages)

Question: "What types of physical activity or exercise did you perform during the past month?"

Probe: "Which, if any, of the following would you (also) consider to be physical activity?"

	African American (n = 109)	Mexican American (n = 112)	Puerto Rican (n = 85)	Non-Hispanic white (n=109)
Walking	89.9	93.8	96.5	89.9
Housework**	78.0	73.7	82.2	63.3
Work-related activity	90.0	94.6	92.1	90.8
Yard work*	92.7	91.2	92.1	82.6

*p < .05. **p < .01.

Table 5. Responses to potato consumption probes by cultural group (percentages)

Question: "About how many times do you eat potatoes per day or per week?"

Probe: "In answering this question, about potatoes, did you think about: [a] French fries or any frozen potatoes; [b] potato chips?"

	African American (n = 102)	Mexican American (n = 106)	Puerto Rican (n = 84)	Non-Hispanic white (n = 98)
[a] French fries/frozen potatoes***	87.3	45.3	59.5	57.1
[b] Potato chips**	33.3	17.0	17.9	18.4

**p < .01. ***p < .001.

Depressive Symptoms

Cultural differences in interpretation of depressive symptoms are widely suspected (Kleinman & Good, 1985). One item from the Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977) asked, "During the past week, how often have you felt that you could not shake off the blues, even with help from family or friends?" After answering this question, respondents were queried using a projective probe that asked, "Do you feel this is a question that people would or would not have difficulty understanding?" Respondents of Hispanic origin were much more likely to believe that people would have difficulty understanding the question, compared with non-Hispanic whites and African Americans (among Puerto Ricans, 42.0% believed people would have difficulty understanding the question; among Mexican Americans, 33.6%; among African Americans, 23.9%; and among non-Hispanic whites, 21.1%).

These results were confirmed in a multivariate analysis that controlled for other respondent characteristics (see Table 6, column 1). A multiple linear regression model was also estimated (see Table 6, column 2). Respondents feeling that others might have difficulty understanding this question reported increased symptom frequency, suggesting that comprehension problems may be associated with greater willingness to endorse this item. Puerto Rican respondents and those with lower incomes also reported greater frequencies of this symptom. These data indicate that culture may

Table 6. Regression analyses of probe and depression symptom question

Question: "During the past week, how often have you felt that you could not shake off the blues, even with help from family or friends? Would you say most of the time, occasionally, a little of the time, only rarely, or none of the time?"

Probe: "Do you feel this is a question people would or would not have difficulty understanding?"

	Difficulty understanding (1 = yes; n = 410)[a]	"Shake off the blues" (1 = none of time; 5 = most of time; n = 404)[b]
Difficulty understanding (1 = yes)	—	0.28*
Male (1 = yes)	–0.04	0.04
African American (1 = yes)	0.13	0.18
Mexican American (1 = yes)	0.34*	0.14
Puerto Rican (1 = yes)	0.54***	0.38*
Age	0.03*	0.01
Education	0.01	–0.04
Income	0.00	–0.16***
Model X^2	19.17**	—
F value	—	3.46**
R^2	—	0.07

[a]Logistic regression model.

[b]Multiple linear regression model.

*p < .05. **p < .01. ***p < .001.

influence symptom reports both directly and indirectly through question comprehension.

Discussion

These findings provide evidence that a number of questions frequently asked in national health surveys are interpreted differently across cultural groups. That most of these differences persisted after controlling for years of formal education and other respondent characteristics provides further evidence that cultural perceptions are moderating the meanings being assigned to many questions by respondents. We can only speculate as to whether or not our results would have been even more divergent had persons from other cultural groups, such as Native Americans or Asian Americans, also participated. We suspect, though, that additional variability of interpretation would have been revealed. It would thus appear that shared stimuli, or question wording, alone is not sufficient to ensure reliable measurement, as the meanings assigned to specific stimuli may consistently vary across cultures.

Developing and evaluating health questions that are more etic than emic will pose a considerable challenge to the survey research community. The first step, however, is developing an awareness that cultural differences are a reality that influences survey data. Although a satisfactory solution is likely years away, what follows are some recommendations for standard practice in question design and testing that we believe are steps in the right direction. Many of them may seem obvious but are important enough to outline below. These suggestions would be applicable to any survey that anticipates interviewing individuals from more than one cultural group.

1. Review draft survey questions with substantive experts from each culture to be surveyed. Wherever possible, a broader goal should be to include experts from the various cultures to be surveyed as part of the research team, enabling them to participate in the development and revision of survey instruments. Such consultation and collaboration would at the very least help researchers avoid making some gross mistakes (e.g., using phrases or concepts totally absent within some cultural traditions) and has the potential to uncover more subtle problems and also moving toward the development of questionnaires that are more etic in scope.

2. Before a questionnaire is tested in a field situation, conduct laboratory interviews with representatives of each culture to be surveyed. In addition to providing firsthand information regarding varying perceptions and interpretations of survey questions, these interviews are also a well-documented approach to the general refinement of survey questionnaires that would benefit all studies.

3. Consider including a small number of probes within the body of questionnaires to be field tested in order to gather additional information on items of special concern. In our research, for example, we were successful in using open- and closed-ended probes to detect cultural differences in understanding and open-ended probes to obtain insights into the perceived meaning of questions. Continued experimentation and refinement of these techniques for integration into field interviews is needed.

4. Through each of these successive phases, items identified as emic can be either revised or replaced with questions hypothesized to be more etic, or pancultural. Consequently, we also recommend initially including multiple versions of some items with the expectation that less etic versions will be subsequently deleted.

This last point reflects our belief that the cultural specificity of any given survey question might best be viewed as existing along a continuum that ranges from mostly emic (i.e., unique to a single culture or variably interpreted across several cultures) to mostly etic (interpreted in a

consistent manner across several cultures). Although our results clearly identified cross-cultural differences in interpretations of most questions examined, we were unable to identify unambiguous examples of either etic or emic questions. These concepts may thus represent ideal types that are rarely, if ever, actually used in practice, yet which provide us with clear conceptual models against which real-life survey items can be assessed.

This paper provides evidence, if any is needed, that cultural differences can no longer be ignored. Few, of course, should be surprised to learn that respondent culture may influence question interpretation, a topic that has received much discussion but little empirical analysis. We believe that more comprehensive models that account for cultural influences on all cognitive phases of the response process (interpretation as well as recall, judgment formation, formatting, and editing) will provide additional insights into how individuals process and respond to survey questions.

References

Aday, L. A., Chiu, G. Y., & Andersen, R. (1980). Methodological issues in health care surveys of the Spanish heritage population. American Journal of Public Health, 70, 367–374.

Andersen, R. N., Mullner, R. M., & Cornelius, L. J. (1987). Black-white differences in health status: Methods or substance? Milbank Quarterly, 65(Supplement 1), 72–99.

Angel, R., & Thoits, P. (1987). The impact of culture on the cognitive structure of illness. Culture, Medicine and Psychiatry, 11, 465–494.

Belson, W. A. (1981). The design and understanding of survey questions. Aldershot, England: Gower.

Berry, J. W. (1969). On cross-cultural comparability. International Journal of Psychology, 4, 207–229.

Bradburn, N. M., Sudman, S., & Associates. (1979). Improving interview method and questionnaire design: Response effects to threatening questions in survey research. San Francisco: Jossey-Bass.

Johnson, T. P., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R. B., Lacey, L., & Horm, J. (1995). Social cognition and responses to survey questions among culturally-diverse populations. Paper presented at the International Conference on Survey Measurement and Process Quality, Bristol, England.

Kleinman, A. (1977). Depression, somatization, and the new cross-cultural psychiatry. Social Science and Medicine, 11, 3–10.

Kleinman, A., & Good, B. (1985). Culture and depression. Berkeley, CA: University of California Press.

Krause, N. M., & Jay, G. M. (1994). What do global self-rated health items measure? Medical Care, 32, 930–942.

Marín, G., & Marín, B. V. (1989). Research with Hispanic populations. Newbury Park, CA: Sage.

Milburn, N. G., Gary, L. E., Booth, J. A., & Brown, D. R. (1991). Conducting epidemiologic research in a minority community: Methodological considerations. Journal of Community Psychology, 19, 3–12.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 3, 385–401.

Triandis, H. C. (1972). The analysis of subjective culture. New York: Wiley-Interscience.

Behavioral Contagion in the Health Field Survey

Daniel H. Hill and James M. Lepkowski

Introduction

The reliability of survey data is, through the stimulus/response framework, linked to how consistently interview-ers present questions to the respondent and record their responses. The survey interview is expected to yield reliable results only when the stimulus is tightly controlled and the full response recorded. When an interviewer changes question wording, variance across interviewers is introduced and item reliability declines. This may be true even if the change is for the better, in a way that reduces bias. Other interviewer behaviors may have undesirable effects as well, including obviously undesirable behaviors, such as reading the wrong question.

Observation of respondent behaviors is also important for understanding the reliability of survey results. Through techniques such as behavior coding, respondent behaviors have been linked to problems with question wording and structure. Interruption of question reading, expressing uncertainty about an answer, giving a "don't know" answer, or refusing to answer are considered to be indicators of problems with the respondent's ability to perform the survey task.

The dynamics of the interview situation may be such that changes or erroneous questioning or respondent behaviors at one point in the interview may induce subsequent changes and behaviors. In this case, the interviewer's or the respondent's behavior may be said to be contagious. Contagion induces considerably more variation in the presentation of questions to respondents both across interviewers and on the part of individual interviewers.

This paper applies a new type of regression model to a unique set of data on interviewer and respondent behaviors, with the intent of seeking evidence of contagion in behaviors affecting the presentation of questions to respondents and of determining whether there are systematic associations of respondent or interviewer characteristics with either the incidence or contagion of these behaviors. The principal question is whether interviewer, respondent, or a combination of interviewer and respondent characteristics account for the behaviors observed in a survey interview. It is possible that a combination of respondent and interviewer characteristics determines respondent and interviewer behavior. A more thorough understanding of the relationship between interviewer and respondent characteristics and survey interviewing behavior may lead to improved interviewing techniques and more reliable data.

In this paper, we present the results of an investigation of the association of interviewer and respondent characteristics with interviewer and respondent behaviors in a sample of interviews from one survey. The presentation includes results from bivariate associations as well as multivariate models in which multiple interviewer and respondent characteristics are examined simultaneously.

In the next two sections, we outline the analytic approach (including an example from the epidemiological literature) and the stochastic models used to examine interviewer and respondent behaviors as contagious processes. The source and nature of the data used in the analysis are then described. Results are then presented, followed by a brief discussion of the implications of the findings.

The Analytic Approach

Consider the entire interview as a single observation in which the outcome of interest is the number of interviewer or respondent behaviors observed in an interview. Counts of behaviors will be modeled as a contagious stochastic process with characteristics of the respondent and interviewer predicting the frequency of occurrence and size of clusters of various types of behaviors. Since contagious process regression models are not commonly used, a simple illustration will be useful.

In a contagious process, the occurrence of one event can trigger subsequent events. Events will not occur randomly and independently in time or space but will appear in clusters. The contagious process used here is from Thomas (1949), who used it to model the spacial distribution of plants in which clusters consist of at least one parent plant and a number (possibly 0) of offspring. The distribution of clusters is assumed to follow a Poisson distribution, and the number of plants within a cluster (minus 1) is also assumed to follow a Poisson distribution. It is assumed that there will be a minimum cluster size of 1 and the size of clusters will exceed 1 only to the extent that there is contagion.

Daniel H. Hill, Associate Research Scientist, is with the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. James M. Lepkowski is a Senior Study Director at the Institute for Social Research and an Associate Professor in the Department of Biostatistics at the University of Michigan, Ann Arbor.

There are two parameters in the Thomas process that have convenient and useful interpretations. The first, cluster incidence intensity, is equivalent to the expected number of primary "infections" for the unit of analysis. The second, cluster size, is equivalent to the expected number of secondary infections per primary infection.

Consider a classic epidemiological example of infection based on the 1954 Polio Vaccine Field Trials (Francis, 1957). There were subjects assigned to treatment (i.e., polio vaccine) and placebo groups at 127 field sites. Counts of the number of paralytic cases of poliomyelitis at each site are available. The Thomas regression model assigns intensity parameters for both the occurrence of a cluster of the disease (i.e., poliomyelitis at one of the sites) and the size of the clusters (i.e., the number of cases occurring at one site). The size of these intensity parameters can be examined with respect to a single predictor: whether the site was in the experimental or the placebo group. Table 1 presents the results of noncontagious (Poisson) and contagious regression models of these data.

The first row of Table 1 presents the estimated intensity from a Poisson regression obtained when the counts of paralytic cases of polio are regressed on the treatment variable "vaccine." The results suggest that the vaccine significantly reduces the intensity of paralytic cases. (This test, by the way, is very close to the one employed by the original evaluators). The remaining two rows of Table 1 present the corresponding contagious (Thomas) regression results, in which the treatment is allowed to affect both the incidence of clusters and cluster size. The log likelihood increases from –316.96 in the noncontagious model to –188.84 in the contagious, implying a X^2 of 256 with 1 degree of freedom for the no-contagion hypothesis implicit in the Poisson model. That is, contagion is evident and strong. Further, the Thomas regression results show that all the effect of the vaccine is in the size or contagious portion of the model. This result suggests that the impact of the vaccine is to reduce the number of secondary infections within sites. In fact, this is a huge effect. In the placebo or unvaccinated population, one exogenous primary infection in a site could be expected to bring about nine secondary infections. In treatment or vaccinated sites, only one in four primary infections could be expected to induce a single secondary infection. Primary infections in the treatment still occurred but in a pattern more or less consistent with the "accidental" assumptions of the Poisson process.

Table 1. Estimated numbers of paralytic polio cases: Contagious and noncontagious regression models

	Placebo	Experimental	t test
Noncontagious model	4.004	1.711	–5.60**
Contagious model			
Incidence	1.046	1.584	0.05
Size	9.198	.244	–4.36

**Significant at the 99% level.

The contagious regression model provides inferences about the effects of the vaccine on primary and secondary infections, even though we cannot separate them in the data. This is a reflection of the fact that the separate effects are implicit in the assumed stochastic process and its fit to the empirical distribution of the total number of infections.

The Stochastic Process

The Thomas model posits that the probability of observing exactly j clusters in a randomly selected area is

$$P_j = \Pr(x = j) = e^{-\lambda}\lambda^j/j! \quad (1)$$

where λ is a Poisson parameter. Furthermore, the conditional probability of observing k secondary infections within the j clusters is

$$P_{k|j} = \Pr(y = k|x = j) = e^{-j\theta}(j\theta)^k/k! \quad (2)$$

where θ is the Poisson parameter for the distribution of secondary infections. Since $E(y|j) = j\theta$, θ can be interpreted as the expected number of significant wording changes or other behaviors per episode or cluster. Since it was impossible for Thomas to distinguish parents from offspring, her interest was in the total number of plants (j + k). Under this "Thomas distribution," the probability of observing 0 objects is

$$P_0 = \Pr(x + y = 0) = e^{-\lambda} \quad (3)$$

while the probability of observing exactly $j + k > 0$ is

$$P_{j+k} = \Pr(x + y = j + k) = \sum \frac{\lambda \cdot e}{r!} \quad (4)$$

The probability of an initial interviewer or respondent behavior occurring is a function of the nature of the interview, particularly the number of questions asked. Long interviews provide greater opportunity for behaviors to occur (i.e., greater risk exposure) than shorter ones. To account for variation in interview length, we can express the overall intensity of initial behaviors as

$$\lambda_j = Q_j\lambda_0 \quad (5)$$

where Q_j is the number of questions in the sequence and λ_0 is the per question intensity.

The intensity parameter λ_0 for interviewer or respondent behaviors may vary from one interview to the next,

depending on the respondent and interviewer characteristics. The intensity of initial behaviors can be expressed as

$$\ln(\lambda_i) = \lambda_{00} + Q_i + \lambda_X X_i + \lambda_I I_i \quad (6)$$

where X and I are vectors of respondent and interviewer characteristics, respectively. A similar expression for the intensity of secondary behaviors is

$$\ln(\theta_i) = \theta_{00} + Q_i + \theta_X X_i + \theta_I I_i \quad (7)$$

An expression for the interview level joint probability of j primary and k secondary interviewer or respondent behaviors can be obtained by substituting the exponentials of (6) and (7) into (3) and (4) to form the log likelihood function

$$\ln(L(\theta_j,\Lambda_{ij})) = \sum \ln(P_{j,k}(Q_j,X_i,I_i)) \quad (8)$$

where Θ and Λ are vectors of parameters from (6) and (7), respectively. Consistent and fully efficient estimates of the parameters of this discrete contagious regression model can be obtained by maximizing this likelihood function with respect to the elements of Θ and Λ.

Because interviewers conducted an average of 10 interviews, standard independence assumptions underlying the computation of standard errors is not justified. Standard errors were computed for a limited number of models using a jackknife estimation procedure (see, e.g. Wolter, 1985) to account for interviewer grouping. These jackknife standard errors were not computed for all models. Results indicate that standard error of estimated coefficients are increased by an average of 50% by the interviewer grouping of interviews. Test statistics presented subsequently are based on independence assumptions and thus should be increased by 50% to account for the effects of interviewer grouping of interviews. Inferential statements have been adjusted for this design effect.

Data and Field Procedures

Survey Research Center staff at the University of Michigan selected a sample of members of an HMO in the Detroit metropolitan area for a methodological study of health and health care utilization in 1993. Youths aged 14 through 17 years and persons aged 65 years and older were selected at higher rates to provide adequate numbers in the study for comparative purposes. A total of 2,006 members completed 1-hour face-to-face interviews (67% response rate) with study staff from April through August 1993. Sampled HMO members provided information about themselves concerning hospital stays, health care visits, usual sources of care, details of their last visit to a health care facility, health care coverage, injuries and poisonings, chronic conditions, and mental well-being. One-half of the respondents were assigned to an experimental interviewing condition that was designed to test hypotheses about commitment, cognitive devices such as detailed instructions, and motivational features such as short and long feedback to reward hardworking respondents. The remaining one-half of the sample received a standard survey interview.

Interviews were, with respondent permission, tape-recorded. Respondents were also asked for permission for study staff to access their medical records at the study HMO. Nearly all subjects (95% or 1,900) provided medical records release permission, and nearly all giving release permission gave permission for the interview to be tape-recorded. There were a total of 1,834 usable tape recordings obtained from the survey.

A sample of 455 usable tape recordings was selected, controlling for respondent age, race, and gender and for interviewer. A staff of six was trained to listen to the tape recordings and code standard interviewer and respondent behaviors into a microcomputer-based data entry system. Coder reliability was established through careful training, group sessions, and individual coding of the same interviewers and discussion of discrepancies among coders.

The data used in this investigation are the behavior codes obtained from the first three sections of the interview for the 455 sample subjects. A total of 54,199 questions were coded on interviewer and respondent behaviors. Interviewer behaviors included in this analysis are whether the interviewer made significant changes to the question wording and whether the interviewer read the wrong question. Respondent behaviors included here are whether the respondent interrupted the question reading with an answer, expressed uncertainty about the answer to the question, gave a "don't know" response, or refused to answer the question. A number of other behaviors were coded but have not been used in this analysis.

Respondent demographic characteristics obtained from the survey (e.g., age, gender, race, education) and interviewer demographic characteristics and employment information (e.g., age, gender, race, education, and length of service) were merged with behavior coding data for the analysis presented subsequently.

Results

Table 2 presents descriptive statistics for the two interviewer and four respondent behaviors of interest. Significant wording changes are far more common than reading the wrong question, with interviewers making changes nearly a dozen times per interview. Since the average number of questions asked per interview is 119, roughly 10% of all questions were modified by the interviewer. Significant wording change is also the most variable of the interviewer behaviors coded. There was one interview with 94 changes, far more than any other interview. The variance is so large that it is unlikely that noncontagious Poisson regression will fit the data well.[1]

[1] Under the Poisson, both the mean and variance are equal to the intensity parameter. For significant changes, however, the variance is nearly 20 times the mean. Some of this excess variance might be explained by the covariates, but not this much.

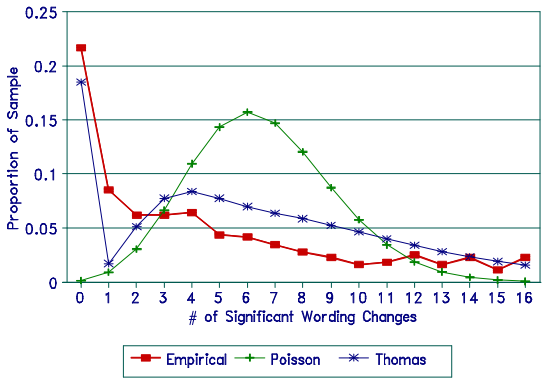
Table 2. Univariate statistics for interviewer and respondent behaviors

	M	SD	Maximum
Interviewer behaviors			
Significant wording changes	11.12	14.74	94
Reading of wrong question	0.50	2.87	49
Respondent behaviors			
Interruption	4.25	6.26	48
Uncertainty	5.65	4.95	31
"Don't know" response	4.06	4.23	30
Refusal	0.88	2.18	18

Figure 1 provides visual confirmation of this observation by plotting the actual empirical distribution (truncated at 16) along with the maximum likelihood Poisson and Thomas distribution fitted distributions. In over 20% of the interviews, the interviewer made no significant wording changes, and in nearly 10% made only one change. The distribution is skewed with a tail stretching out all the way to the maximum 94 changes. The dotted line represents the maximum likelihood Poisson model for the data, which is not very good at all. The Poisson predicts very few cases at either extreme of the distribution and has virtually all of its mass between 3 and 10 changes per interview. In order to increase the likelihood of cases in the right-hand tail, the Poisson intensity parameter is being increased so that the Poisson distribution is approaching its normal limit. The Thomas distribution does a much better job at both ends of the distribution. Having two cluster parameters allows a prediction of substantial numbers of cases with no changes, while also allowing the prediction of some interviews with very a large number of changes.

Although space precludes our listing the results here, a series of bivariate Thomas regression models was estimated for counts of the number of significant wording changes and reading of the wrong question by interviewers. A wide variety of respondent and interviewer variables was statistically associated with both the cluster incidence and cluster size portions of the model. Every interviewer and respondent characteristic except respondent cooperation has a significant association with either the incidence or size of clusters. More interviewer and respondent characteristics are associated with significant wording changes than reading the wrong question due to the higher frequency of significant wording changes.

Figure 1. Empirical and theoretical distributions: Counts of significant changes



The occurrence of clusters for significant wording changes is positively associated with interviewer age (i.e., as interviewer age increases, the occurrence of clusters increases), male interviewers, black interviewers, and experience and negatively associated with interviewer education. Respondent race, understanding of the interview, and use of records are also positively associated with cluster incidence, while respondent education and level of effort are negatively associated. The size of the clusters, as a measure of contagion of significant wording changes, is positively associated with interviewer age, gender (male), race (black), and experience and with respondent age, race (black), understanding, and use of records. Significant-wording-changes cluster size is negatively associated with respondent education and level of effort.

Another series of bivariate models was estimated for each respondent and interviewer characteristic for each of four respondent behaviors: interrupting question reading, expressing uncertainty about an answer, giving a "don't know" response, and refusing to answer. Every interviewer and respondent characteristic except respondent use of records is associated with either the incidence or size of clusters of these respondent behaviors. Respondent age, education, and understanding of the questions are associated with cluster size for all four respondent behaviors. Cluster incidence and size for interrupting the question reading is associated with more of these interviewer and respondent characteristics than the other three behaviors. The overall impression is of a substantial level of cluster incidence and substantial variation in cluster size that is strongly associated with many interviewer and respondent characteristics.

Table 3 presents t values for coefficients in multivariate models for the two interviewer behaviors (i.e., significant wording changes and reading the wrong question) in which all interviewer and respondent characteristics are included simultaneously. In these models, separate cluster incidence and size parameters are estimated and are all statistically significant. That is, once interviewer and respondent characteristics are taken into account, there is a substantial level of clustering for both interviewer behaviors. Perhaps not surprisingly, only a handful of interviewer and respondent characteristics remain associated with cluster incidence or size in the multivariate model. It appears that many of the bivariate effects discussed above are explained by a subset of characteristics. For example, interviewer race and education continue as predictors of significant-wording-change cluster incidence, while interviewer age, gender, and pay rate and respondent cooperation are associated with

Table 3. t values for multivariate Thomas regression coefficients for interviewer behaviors by cluster incidence and size and by respondent and interviewer characteristics

Predictor question	Significant wording changes	Reading the wrong
Cluster incidence	–3.90	–4.06
Interviewer		
Age	–2.54	1.44
Gender (male)	–0.50	0.72
Race (black)	5.20	0.72
Education	–4.03	–4.21
Experience	2.46	–1.23
Pay rate	0.90	0.17
Respondent		
Age	1.04	–3.73
Gender (male)	0.68	1.11
Race (black)	–1.54	–0.11
Education	–0.99	2.87
Understanding	–2.91	3.89
Cooperation	1.92	0.62
Use of records	1.88	0.82
Effort	–2.06	–3.99
Cluster size	–4.54	–3.23
Interviewer		
Age	8.60	–0.34
Gender (male)	6.36	–2.43
Race (black)	1.79	–0.28
Education	–0.26	1.40
Experience	1.95	–0.57
Pay rate	–3.44	2.86
Respondent		
Age	1.48	–4.41
Gender (male)	–0.19	0.14
Race (black)	–0.31	1.70
Education	–0.38	–2.98
Understanding	3.43	–2.23
Cooperation	–1.05	–1.17
Use of records	2.42	0.00
Effort	–1.45	2.07

cluster size. Only this latter respondent characteristic remains associated with cluster incidence or size for significant wording changes.

Table 4 presents the multivariate models for the four respondent behaviors. The overall cluster size parameter is not significant for expressing uncertainty or refusing to answer the question. As in the interviewer behaviors, only a few interviewer and respondent characteristics remain associated with cluster size or incidence. For example, for interrupting the question, only respondent age is associated with cluster incidence. No other interviewer or respondent characteristic is associated with interrupting-question-reading cluster incidence or size.

Table 4. t values for multivariate Thomas regression coefficients for respondent behaviors by cluster incidence and size and by respondent and interviewer characteristics

Predictor Refusal	Interrupts	Uncertain	"Don't know" response	
Cluster incidence	–4.74	–7.16	–4.58	–5.96
Interviewer				
Age	2.18	–0.11	–2.16	–0.46
Gender (male)	0.80	–0.37	1.04	–0.04
Race (black)	2.13	–0.72	0.37	0.32
Education	1.02	0.70	0.67	–0.61
Experience	0.74	1.59	1.75	–0.26
Pay rate	–1.45	0.05	–0.68	1.08
Respondent				
Age	3.57	1.27	0.77	–1.19
Gender (male)	–1.64	–1.72	0.39	–0.25
Race (black)	–1.39	–0.67	–2.55	–2.24
Education	0.94	2.74	–0.63	–1.90
Understanding	0.52	3.00	1.28	2.86
Cooperation	–1.39	–0.81	0.90	–0.56
Use of records	1.05	–3.70	0.06	0.66
Effort	1.69	–1.32	–0.53	2.34
Cluster size	–4.53	–2.33	–5.92	–0.43
Interviewer				
Age	–0.65	0.97	1.49	2.45
Gender (male)	–1.50	3.45	0.48	0.98
Race (black)	–1.02	2.19	–0.85	3.81
Education	1.10	–3.51	–1.82	–0.61
Experience	1.25	–0.95	–2.14	–1.81
Pay rate	1.04	–1.28	1.61	–1.89
Respondent				
Age	0.51	2.71	3.07	1.40
Gender (male)	1.13	2.62	0.41	5.24
Race (black)	–1.25	0.38	0.90	–4.00
Education	1.69	–4.79	0.57	–6.50
Understanding	0.03	0.26	2.13	–0.70
Cooperation	2.37	1.39	0.51	–2.13
Use of records	–0.24	3.95	1.76	–1.09
Effort	–0.02	0.89	0.35	–0.03

Discussion

In this paper, we have applied a contagious regression model to data on interviewer and respondent behaviors in a health survey to answer two related questions. First, are interviewer behaviors that affect data reliability contagious? That is, do behaviors like making significant wording changes occur independently and randomly within the interview, or do they cluster with one incident leading to another? Second, are the incidence and clustering of behaviors systematically related to respondent and/or interviewer characteristics?

Our analyses suggest that both interviewer and respondent behaviors are contagious, with significantly more

clumping than could reasonably be accounted for by chance. The analyses suggest rather strong associations, primarily with interviewer rather than respondent characteristics.

There are more results to examine here than space permits us to review. We illustrate the interpretation of findings for two behaviors only: significant wording changes and expressing uncertainty about an answer. For example, black interviewers have higher frequency of clusters of significant wording changes than nonblack interviewers. One possible explanation is that black interviewers conducted more interviews in black respondent homes than nonblack interviewers and that they altered question wording to fit different respondent interpretations of health characteristics that were asked about. Previous research has shown cultural differences in the interpretation of terms about health between black and nonblack respondents. Why respondent race is not significant instead of interviewer race for cluster incidence is not clear. Not unexpectedly, interviewers with higher levels of education in our study have a lower incidence of clusters of significant wording changes. Somewhat surprisingly, older interviewers have larger clusters of changes than younger interviewers. Older interviewers are more experienced interviewers on average and had interviewer training on basic techniques longer ago. The effects of training to read questions exactly as written may have diminished among older interviewers. Male interviewers also have larger clusters of question-wording changes. Male interviewers appear to have a tendency to make significant wording changes in larger clusters than female interviewers. Interviewer pay rate is negatively associated with the size of clusters. That is, as pay rate increases (as a function of experience and quality of performance), the size of clusters of significant wording changes diminishes. Conversely, interviewers with lower pay rates have larger clusters of significant wording changes. Finally, respondents who in the opinion of the interviewer had greater difficulty understanding questions have larger clusters of significant wording changes. One explanation for this finding is that interviewers were altering question wording to make the questions more understandable for these respondents.

Cluster incidence of respondent uncertainty about a question or answer is positively associated in the multivariate model in Table 4 with respondent age. That is, older respondents are more likely to have clusters of questions for which they express uncertainty than younger respondents. This may reflect more complex patterns of health care use, or it may reflect the effects of poorer recall or cognitive ability for older persons. Male interviewers have larger clusters of respondent uncertainty; there is no apparent explanation for this finding. Less educated interviewers and respondents have, not surprisingly, larger clusters of uncertain responses. And lastly, as respondents use records to answer questions, they have smaller clusters of uncertain answers.

Many of these findings confirm our expectations about what is going on during the interview and how it is associated with interviewer and respondent characteristics. The contagious regression models provide a powerful new tool for exploration of these associations. It can and should be used to explore a number of other issues with these particular data. For example, we have not examined interactions of respondent and interviewer characteristics that might aid in explaining some of the findings observed in the multivariate analyses. Using the entire interview as the unit of analysis is not very sensitive to changes in behavior that may occur throughout an interview. Further analysis of sections of the interview or sequences of questions may provide greater insight into the nature of the clustering behavior observed here.

References

Francis, T. F. (1957). Evaluation of the 1954 Field Trial of Poliomyelitis Vaccine—final report. Ann Arbor, MI: University of Michigan.

Hill, D. (1993). Response and sequencing errors in surveys: A discrete contagious regression analysis. Journal of the American Statistical Association, 88, 775–781.

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. Biometrika, 36, 18–25.

Wolter, K. (1985). Introduction to variance estimation. New York: Springer-Verlag.

Behavior of Survey Actors and the Accuracy of Response

Robert F. Belli and James M. Lepkowski

A critical component of surveys that ask about health behaviors and conditions is the accuracy of response. It is well known that respondents tend to underreport the extent to which they visit health care professionals (Cannell, Marquis, & Laurent, 1977; Cannell, Miller, & Oksenberg, 1981; Means & Loftus, 1991). Such underreporting is often attributed to respondents who may be unclear about the objectives of the questioning or who may not be motivated to meet the memory and cognitive demands necessary to develop an adequate response.

Behavior coding (Fowler & Cannell, 1996; Mangione, Fowler, & Louis, 1992; Oksenberg, Cannell, & Kalton, 1991) is a technique that can provide clues concerning the extent to which survey questions are unclear and tax the cognitive capacity of respondents. Trained staff listen to audiotapes of survey interviews and code interviewer and respondent behaviors. The behaviors of interviewers can be coded for changes to question wording and appropriate or inappropriate probing and feedback. Oftentimes, these codes indicate the extent to which interviewers are seeking ways to clarify question wording and objectives. Respondent codes include question interruptions, expressions of uncertainty, qualified and uncodable answers, and "don't know" and refusal responses. Respondent codes indicate the extent to which questions are cognitively demanding either during their interpretation or in the degree to which they involve considerable effort for recall and response formulation. Behavior coding provides insight regarding those cognitive aspects of the survey process that may be associated with the accuracy of response.

A number of behavior coding studies have indirectly indicated that the behaviors of survey actors are associated with response accuracy. For example, undesirable interviewer and respondent behaviors have been shown to occur most often only with particular questions, leading to the inference that these questions are likely to encourage response errors (Oksenberg et al., 1991). In addition, revisions to such questions not only will lead to reductions in problem behaviors, but also have been shown to lead to different—and putatively more accurate—answers than those given to the original questions (Fowler, 1992). Mangione et al. (1992) report that certain interviewer behaviors, particularly the manner in which interviewers probe respondents following inadequate answers, are associated with variance in response. Apparently, the manner in which interviewers probe affects the answers that respondents provide. In research on interviewer feedback, Cannell et al. (1981) found that positively valenced feedback from interviewers (e.g., "you're doing fine") is more likely to occur for undesired respondent behavior (such as refusal to answer) than for desired behavior (such as an adequate answer). Although interviewers are probably intent on maintaining rapport, an inadvertent result of providing feedback to undesirable responses is the continued encouragement of less-than-adequate answers. In fact, the report of health-related behaviors is significantly increased, suggesting more complete and accurate reporting, when interviewers are trained in appropriate feedback techniques than when they are not (Cannell et al., 1981).

Although the evidence from these studies is provocative regarding an association between the behavior of survey actors and the quality of response, all of the indicators of response accuracy have been indirect. None have directly examined the association of behavior coding with a measure of response accuracy, as can be done by checking survey responses against external records. The present research compares the agreement between respondent reports of hospital stays and health care office visits with the medical records available through respondent participation in an HMO. We examine the degree to which interviewer and respondent behaviors are directly related to the accuracy of the survey answers and whether these behaviors continue to have an influence, while controlling for other interviewer and respondent characteristics (such as length of interviewer service and age of respondent) that may be associated with the accuracy of response. The behaviors of survey actors are examined at the time of asking and answering survey questions. Further, behaviors that occurred in preceding survey questions are also examined to provide an indication of whether prior patterns of behavior have an influence on ongoing questions.

Methods

A total of 2,006 members of an HMO in the Detroit metropolitan area responded to a face-to-face survey

Robert F. Belli, Assistant Research Scientist, is at the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor. James M. Lepkowski is a Senior Study Director at the Institute for Social Research and an Associate Professor in the Department of Biostatistics at the University of Michigan, Ann Arbor.

interview about health and health care utilization during the period of April to August 1993. Interviews lasted approximately 1 hour, and self-report topics included hospital stays and the number of visits to health care providers. The design of the study included the random assignment of an experimental version of the survey to half of the respondents in order to introduce some special interviewing techniques. The remaining interviews used a standard questionnaire.

Nearly all interviews were, with respondent permission, tape-recorded. In addition, respondents were asked if research staff could obtain medical records data from their HMO. A total of 1,834 gave permission on both requests. Subsequently, through stratified random sampling controlling for respondent age, gender, and race, a sample of 455 interviews were selected and behavior coded. All six behavior coders had interviewing experience as well as experience monitoring interviewer performance or behavior coding. The coders were given specific training commensurate with their experience. Further, all coders participated in group training sessions involving the coding of selected interview passages and also were trained by independently coding a subset of interviews with follow-up discussion of disagreements.

Dependent Variables

Three dependent variables created from a comparison of survey reports and medical records are investigated here. Hospital stays involved two survey questions: (a) a "yes"/"no" response to "Since (CURRENT MONTH) 1st, 1992, have you been a patient in a hospital overnight?" and (b) given a "yes" response to this question, "How many different times did you stay in any hospital overnight or longer since (CURRENT MONTH) 1st, 1992?" For computing the survey report on number of hospital stays, "no" responses to the first "yes"/"no" question were coded as 0. In the few instances in which respondents gave a range even after interviewer probing (e.g., "1–3 times"), the midpoint of the range was used.

There were two dependent variables associated with health visits: The 6-month visits variable was based on the report of visits in the previous 6 months, and the 12-month visits variable was based on a 12-month reference period. Respondents were given either the 6- or 12-month reference period, with a random one-half of subjects assigned to each. Data were collected in a series of questions that asked about health visits to hospital emergency rooms, urgent care centers, doctors' offices, and any other health care facilities. Following these questions, the interviewer summed the number of reported visits (skipping those responses that were not ascertained or ranges) and asked, "I see that the total number of visits you had to medical doctors or assistants during the past (reference period, 6 or 12) months is (TOTAL). Is this number correct?" If the respondent answered "no," the interviewer read a scripted probe to determine the exact number of visits. For the purposes of this analysis, if the respondent confirmed the interviewer's summed number of visits, that sum was used as the survey report. If the respondent answered "no" to the confirming question, the respondent's final answer to the scripted probe was used as the survey response.

For all three measures, the corresponding value from the medical records was subtracted from the survey report to create the dependent variable. Regression analyses reported below are based on the absolute value of this difference.

The medical records are known to be incomplete by failing to capture visits made to facilities outside of the HMO (see Jay, Belli, & Lepkowski, 1994). One sequence of questions involved detailed queries concerning the last visit to a medical care provider, including where care was received. Approximately 16% of respondents reported having a last visit to a medical facility that was outside the HMO. Although the level to which respondents reported out-of-system visits is not known for the three dependent variables reported here, we assume that they are present and would lead respondents to report visits that would not appear in the medical records. These dependent variables are thus not complete measures of accuracy, but nevertheless are adequate for our purposes.

Predictor Variables

Interviews were behavior coded on several characteristics associated with interviewer question asking, respondent answering, interviewer probing, and interviewer feedback behaviors. Figure 1 presents the codes and summary criteria by which they were applied.

Two different classes of predictor variables were created from the behavior codes. At-the-question predictors were assigned values of 0 or 1 depending on whether a code was assigned during the interviewer-respondent interaction at the time that the question was administered. Preceding-question predictors were computed as the proportion of times particular behavior codes were assigned in prior questions.

At-the-question predictors were created for all the behavior codes in Figure 1 except for the feedback codes. These feedback codes were not included as at-the-question predictors, since feedback ordinarily follows the response. Thus, any influence from feedback should affect only those responses to subsequent questions. Accordingly, preceding-question predictors were created to correspond to all of the behavior codes in Figure 1; they can be recognized by the designation of "PQ," which follows the ordinary letter code sequence (e.g., NORPQ represents no respondent behavior, preceding-question).

The hospital stays variable was based on two survey questions, complicating specification of behaviors for analysis. For those respondents who answered "no" to the initial "yes"/"no" question, at-the-question predictors were derived from this initial question. The preceding-question predictors were derived from the three prior questions. For

Figure 1. Behavior codes

Interviewer question-asking codes

QE Exact: reads exactly as written or makes insignificant changes
QS Significant changes: makes wording changes that can affect meaning
QO Other changes: verifies, states, or suggests an answer; reads nonapplicable question; skips applicable question

Respondent answering codes

R1 Interruption: interrupts question with an answer
R2 Uncertainty: expresses uncertainty, requests question repetition, or seeks clarification
R3 Qualified response: qualifies answer with phrases such as "about," "I guess," "maybe," etc.
R4 Uncodable/inadequate response: response does not meet question objectives
R5 Item nonresponse: "don't know" responses or refusals
NOR No respondent behavior: none of the R1 to R5 respondent behaviors were assigned

Interviewer probing codes

PA Adequate probing: probing is nondirective and sufficient
PI Inadequate probing: at least one probe is directive or under- or overprobes

Interviewer feedback codes

F1 Acceptable short: neutral and appropriate short phrase (1–3 words), such as "thank you"
F2 Acceptable long: neutral and appropriate longer phrase, such as, "Thanks. That's useful information for our study."
F3 Unacceptable short: offers short phrase that may indicate approval for the content of the response
F4 Unacceptable long: offers longer phrase that may indicate approval for the content of the response
F5 Unacceptable reward: indicates approval for a "don't know" response, refusal, digression, interruption, or inadequate final answer

those respondents who answered "yes," at-the-question predictors were derived from the follow-up question on number of stays, and the preceding-question predictors were derived from the prior questions, including the "yes"/"no" hospital stays question. Since the hospital stays questions were asked early in the interview, preceding-question predictors were based on three or four previous questions. For the 6- and 12-month visits, the number of prior questions ranged from 6 to 16.

Model Construction

The simple difference between survey reports and medical records has negative as well as positive values associated with inaccurate reports, and the dependent variables were computed as absolute values of the difference. The distribution of the absolute values for these data has two undesirable features: a concentration of values at 0 and skewness to the right. Both of these features can be handled through transformations such as a Tobit (for the concentration of values at 0) or a logarithmic (for skewness) transformation. These transformations detract from simplicity of model interpretation. In this initial descriptive and exploratory analysis, we chose to retain the simplicity of the absolute difference scale. For each dependent variable, there were several values that were found, through an examination of standard influence statistics, to have an overly influential effect on regression models. These were excluded as outliers: one for hospital stays, three for 6-month visits, and four for 12-month visits.

For each of the dependent variables, we first computed the bivariate regression of each dependent variable on all of the at-the-question and preceding-question predictors. Those predictors with statistically significant associations with the dependent variable ($p < .10$) were then included in a multivariate regression model. The multivariate models also contained interviewer (age, gender, race, education, and length of service) and respondent (age, gender, education, and marital status) characteristics, as well as an indicator of questionnaire version, as control variables.

Results

Number of cases, means, and standard deviations for raw and absolute difference scores are reported in Table 1 for the three dependent variables. The raw difference scores are notable in that each variable yields overall overreporting, in contrast with the general tendency for people to underreport their visits to health care professionals (see, e.g., Means & Loftus, 1991). This apparent anomaly is due to the failure of the medical records to capture all of the health care visits of respondents. Despite this failure to capture out-of-system visits, means of the absolute difference scores do indicate that the differences measure respondent-reporting accuracy. The mean absolute difference for hospital stays is lower than that for 6- or 12-month visits, as expected, since respondents are better able to accurately report events that tend to be less numerous and distinctive. Further, 6-month visits show greater correspondence between survey reports and medical records in comparison with 12-month visits, exactly what we expected, given that longer reference periods are associated with higher degrees of forgetting.

Results concerning the predictor variables; their means; and their performance in bivariate and multivariate association with hospital stays, 6-month visits, and 12-month visits are presented in Tables 2, 3, and 4, respectively. The bivariate statistics reveal that reports of hospital stays are

Table 1. Description of dependent variables

		Raw difference		Absolute difference	
Variable	n	M	SD	M	SD
Hospital stays	451	0.06	0.29	0.07	0.29
6-month visits	240	0.62	3.10	2.00	2.45
12-month visits	202	0.26	5.51	3.69	4.10

Table 2. Means of behavior codes and bivariate and multivariate associations with hospital stays

Predictors	M	Bivariate	Multivariate[a]
At-the-question			
QE	0.692	0.033	
QS	0.181	–0.030	
QO	0.007	–0.074	
R1	0.069	0.025	
R2	0.040	0.097	
R3	0.002	–0.074	
R4	0.045	0.185**	0.041
NOR	0.842	–0.097**	–0.003
PA	0.033	0.200**	0.207**
PI	0.049	0.209**	0.222**
Preceding-question			
QEPQ	0.731	0.039	
QSPQ	0.149	–0.051	
QOPQ	0.000	0.000	
R1PQ	0.031	–0.093	
R2PQ	0.079	0.198**	0.155*
R3PQ	0.039	0.029	
R4PQ	0.091	–0.048	
R5PQ	0.036	0.037	
NORPQ	0.769	–0.031	
PAPQ	0.098	0.100	
PIPQ	0.051	–0.012	
F1PQ	0.262	0.002	
F2PQ	0.002	1.824**	1.843**
F3PQ	0.012	–0.080	
F4PQ	0.004	0.179	
F5PQ	0.006	–0.236	

[a]Controlling for interviewer and respondent characteristics.
*0.10 < p < 0.05. **p < 0.05.

Table 3. Means of behavior codes and bivariate and multivariate associations with 6-month visits

Predictors	M	Bivariate	Multivariate[a]
At-the-question			
QE	0.717	0.059	
QS	0.104	–0.625	
QO	0.035	0.586	
R1	0.043	–0.872	
R2	0.087	–0.093	
R3	0.057	0.395	
R4	0.030	–0.227	
NOR	0.765	0.157	
PA	0.070	0.540	
PI	0.039	–0.395	
Preceding-question			
QEPQ	0.673	0.089	
QSPQ	0.176	–0.500	
QOPQ	0.011	–0.152	
R1PQ	0.039	–0.167	
R2PQ	0.079	1.918	
R3PQ	0.037	8.538**	8.141**
R4PQ	0.072	3.406**	2.515
R5PQ	0.037	0.464	
NORPQ	0.786	–2.289**	–0.535
PAPQ	0.070	1.923	
PIPQ	0.057	3.244*	1.520
F1PQ	0.157	0.509	
F2PQ	0.007	–2.708	
F3PQ	0.008	10.352**	8.357*
F4PQ	0.006	0.291	
F5PQ	0.013	–3.469	

[a]Controlling for interviewer and respondent characteristics.
*0.10 < p < 0.05. **p < 0.05.

significantly associated with at-the-question predictors more often than they are with preceding-question predictors, whereas reports of 6- and 12-month visits are significantly associated with preceding-question predictors more often than they are with at-the-question predictors.

On none of the dependent variables did the interviewer question-asking codes (QE, QS, QO, and their preceding-question counterparts) reach statistical significance (even at the 0.10 level). In contrast, there was always at least one predictor from each of the remaining code types (i.e., respondent-answering codes, interviewer-probing codes, and interviewer feedback codes) that demonstrated a significant association with at least one of the three dependent variables.

Each of the multivariate models (which always included control variables, described previously) demonstrated a considerable reduction in the number of significant predictor variables in comparison with the bivariate associations. Two respondent behaviors reached significant levels of association: Preceding-question uncertainty (R2PQ) was positively associated with hospital stays, and preceding-question qualified response (R3PQ) was positively associated with both 6- and 12-month visits. That is, the presence of these behaviors is associated with greater disagreement or inaccuracy in the dependent variables. Curiously, both adequate and inadequate at-the-question probing codes (PA and PI) were positively associated with hospital stays (see Table 3), indicating that the occurrence of either was marked by increased inaccuracy. Although occurring infrequently, the presence of preceding-question acceptable long feedback (F2PQ) was significantly associated with increased inaccuracy for both hospital stays and 12-month visits, and the infrequently occurring preceding-question unacceptable short feedback (F3PQ) was likewise positively associated with 6-month visits.

Table 4. Means of behavior codes and bivariate and multivariate associations with 12-month visits

Predictors	M	Bivariate	Multivariate[a]
At-the-question			
QE	0.803	–0.236	
QS	0.083	0.827	
QO	0.010	0.325	
R1	0.031	–2.421	
R2	0.073	–0.039	
R3	0.093	2.254**	–0.062
R4	0.073	0.731	
NOR	0.731	–0.071	
PA	0.052	0.022	
PI	0.078	–0.664	
Preceding-question			
QEPQ	0.763	0.415	
QSPQ	0.125	–0.324	
QOPQ	0.006	–6.073	
R1PQ	0.041	1.703	
R2PQ	0.080	6.425*	5.300
R3PQ	0.055	7.378*	9.252*
R4PQ	0.077	5.799**	7.351
R5PQ	0.064	5.504**	3.795
NORPQ	0.749	–5.161**	4.126
PAPQ	0.079	9.315**	1.363
PIPQ	0.065	–1.196	
F1PQ	0.167	0.062	
F2PQ	0.004	34.856**	37.643**
F3PQ	0.013	11.265*	9.937
F4PQ	0.002	–10.373	
F5PQ	0.009	17.095*	–1.765

[a]Controlling for interviewer and respondent characteristics.

*$0.10 < p < 0.05$. **$p < 0.05$.

Discussion

Findings indicate that behavior codes uniquely reflect aspects of the quality of survey response that are not accountable by other measures. Moreover, the patterns of results provide insights regarding those aspects of the behavior of survey actors that are the best indicators of the likely accuracy of survey reports.

Whereas questions presented early in the survey yielded greater degrees of association with at-the-question codes, questions presented later yielded higher associations with preceding-question codes. With early questions, there is very little prior interaction between the survey actors, apparently making the at-the-question behaviors more informative with regard to response accuracy. On the other hand, at-the-question behaviors are not informative of response accuracy with later questions, perhaps because the behavior coding of prior questions reveals a behavioral pattern among the survey actors that is more diagnostic of accuracy than those behaviors that occur at the question.

The question-asking behavior of interviewers (i.e., whether the question was asked exactly as written or was read with significant changes) never reached statistical significance in any of the analyses. This suggests that interviewer variance from scripted performance did not affect accuracy. Given that there are significant associations of respondent behavior with accuracy, these results indicate that respondent behavior is more diagnostic of response accuracy than anything over which the interviewer has direct control. In comparison with question asking, the other interviewer behaviors (interviewer probing and feedback) are likely to depend more on respondent behavior. Suggestive of this dependency are the results for hospital stays, which showed that regardless of whether an interviewer probed adequately or inadequately, the occurrence of probing was associated with greater inaccuracy. Interviewers are instructed to probe whenever respondents provide less-than-adequate answers. Our results suggest that the behavior of interviewers will be ineffective following an inadequate report. Brenner (1982) has noted that while inadequate directive probing can lead the respondent to a firm but inaccurate final response, nondirective adequate probing, in which respondents are encouraged to make another response attempt, will often result in a final inaccurate, inadequate response (see also Mangione et al., 1992).

The interviewer feedback results are particularly puzzling. The occurrence of preceding-question adequate long feedback was found to be associated with greater inaccuracy on hospital stays and 12-month visits. There are at least two possible reasons for this finding. Consistent with the interpretation that interviewer behavior has little influence on accuracy, interviewers may be more likely to use long feedback with the more troublesome respondents, with little or no effect. Alternatively, the finding that preceding-question adequate long feedback is only associated with those dependent variables that were based on the longer 12-month reference period, and not the shorter 6-month period, may indicate that the use of long feedback is having an undesirable effect. Burton and Blair (1991) found that encouraging respondents to engage in a more effortful enumeration strategy actually increased inaccuracy, in comparison with a less effortful estimation strategy when the frequency of events that occurred during the reference period were more numerous. Similarly, by providing feedback to earlier questions, interviewers may be successful in encouraging respondents to try hard. However, with the longer reference periods and the higher frequency of events that longer periods contain, interviewers may also be inadvertently encouraging the use of an ineffective strategy.

In summary, the results lend little support to the notion that indicators of adequate and inadequate interviewing technique are associated with the quality of survey response (see Fowler, 1992, and Oksenberg et al., 1991). Nevertheless, caution must be taken before concluding that reducing problematic interviewer behaviors will be ineffective for improving survey response. The null findings with interviewer

question asking only indicate that the exact or inexact reading of the same questions is not likely to have much influence on response quality. If interviewers tend to misread poorly written questions (Fowler & Cannell, 1996), and poorly written questions, in turn, are likely to lead to poor responses, then despite our results, providing interviewers with better questions will improve response accuracy.

References

Brenner, M. (1982). Response-effects of "role-restricted" characteristics of the interviewer. In W. Dijkstra & J. Van Der Zouwen (Eds.), Response behavior in the survey interview (pp. 131–165). London: Academic Press.

Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. Public Opinion Quarterly, 55, 50–79.

Cannell, C. F., Marquis, K. H., & Laurent, A. (1977). A summary of studies of interviewing methodology. Vital and Health Statistics (Series 2, No. 69; DHEW Publication No. HRA 77-1379). Washington, DC: U.S. Government Printing Office.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), Sociological methodology (pp. 389–437). San Francisco: Jossey-Bass.

Fowler, F. J. (1992). How unclear terms affect survey data. Public Opinion Quarterly, 56, 218–231.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), Answering questions: Methodology for determining cognitive and communicative processes in survey research (pp. 15–36). San Francisco: Jossey-Bass.

Jay, G. M., Belli, R. F., & Lepkowski, J. M. (1994). Quality of last doctor visit reports: A comparison of medical record and survey data. American Statistical Association 1994 proceedings of the Section on Survey Research Methods: Vol. 1, 362–367.

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question characteristics and interviewer effects. Journal of Official Statistics, 8, 293–307.

Means, B., & Loftus, E. F. (1991). When personal history repeats itself: Decomposing memories for recurring events. Applied Cognitive Psychology, 5, 297–318.

Oksenberg, L., Cannell, C. F., & Kalton, G. (1991). New strategies for testing survey questions. Journal of Official Statistics, 7, 349–365.

Heuristics Used by Older Respondents to Answer Standardized Mental Health Questions

Bärbel Knäuper and Hans-Ulrich Wittchen

Introduction

The use of standardized diagnostic interviews in assessing the prevalence of mental disorders in the general population has increased in recent years. In these diagnostic interviews (e.g., the Composite International Diagnostic Interview [CIDI; World Health Organization, 1990]), a long series of questions is asked to assess the diagnostic criteria of various disorders according to modern classification systems, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R; American Psychiatric Association, 1987) or the International Classification of Diseases, 10th revision (ICD-10; World Health Organization, 1991). The translation of diagnostic criteria into questions, however, results sometimes in lengthy and complex sets of symptom and probe questions. Symptom questions include all severity and time-related information needed to distinguish psychopathological symptoms from less severe, subclinical symptoms. Furthermore, to allow the estimation of lifetime prevalence rates, some of these instruments, like the CIDI, ask the questions retrospectively for the whole life span of respondents. For example, in the CIDI, one of the symptoms for a major depression syndrome, trouble concentrating, is assessed by the following question: "Has there ever been two weeks or more when nearly every day you had a lot more trouble concentrating than is normal for you?" In this example, the expressions "two weeks or more," "nearly every day," and "a lot more . . . than is normal" are used to discriminate more normal, everyday complaints from long-lasting, severe symptoms that possibly reflect a clinically significant mental illness symptom. Furthermore, if respondents answer this question in the positive, they are probed as to whether the symptom was "ever" (and, in a second step, "always") caused by a physical illness or injury or by taking drugs, medication, or alcohol. This probing serves to exclude from diagnosis symptoms that were not entirely psychologically caused.

Bärbel Knäuper is with the Institute for Social Research, University of Michigan, Ann Arbor. Hans-Ulrich Wittchen is with the Max-Planck Institut für Psychiatrie, Munich, Germany.

The reported research was supported by grant Wi 709/3-1 from the Deutsche Forschungsgemeinschaft. We would like to thank Ron Kessler and Norbert Schwarz for valuable comments on a previous draft of this paper.

From a cognitive perspective, answering these symptom and probe questions requires complex comprehension, retrieval, and judgment tasks. These cognitive tasks might be particularly difficult for older respondents, reflecting that cognitive capacities decrease with age. Additionally, older respondents have to review a longer time period than younger respondents to find possible symptoms in memory. This further increases the demands of the memory task. Finally, determining if an experienced symptom was psychologically rather than physically caused could also be a particularly challenging—and perhaps unfeasible—task for older people. In older age, individuals are more likely to experience physical complaints and illnesses (see, e.g., Blazer, 1989), and it may be difficult to decide if the physical illness caused the depressive feelings or vice versa. Altogether, the complexity of the comprehension, memory, and judgment tasks presented by the questions point to the possibly limited validity of standardized diagnostic interview questions for older respondents.

These issues have been of particular interest in recent years because several epidemiologic population studies, conducted in different countries, found surprisingly low lifetime and current prevalence estimates of depressive and other mental disorders in the elderly—or conversely, high rates in younger age groups (cf. Wittchen, Knäuper, & Kessler, 1994). This phenomenon is described in the literature as an age or birth cohort effect of depression (Cross-National Collaborative Group, 1992; Klerman & Weissman, 1989). For example, the data from the Epidemiological Catchment Area Study (ECA; Robins & Regier, 1991) revealed that only 1.4% of people over 65 experienced a major depressive disorder at least once in their life, compared with 7.5% of people aged 30 to 44, based on the Diagnostic Interview Schedule. Moreover, only 0.9% of people over 65 years of age were diagnosed as currently depressed, compared with 3.9% of people aged 30 to 44. The validity of these findings has been questioned because older people had a longer time at risk for developing the disorder; thus, their lifetime prevalence rates should be higher than those of younger people. Moreover, most studies which use self-reported symptom scales rather than standardized interviews find increasing current depressive symptomatology with increasing age (e.g., Berkman et al., 1986; Blazer, Burchett, Service, & George, 1991; Gaitz &

Scott, 1972; Gurland, Kuriansky, Sharpe, Simon, Stiller, & Birkett, 1977; Roberts, Lee, & Roberts, 1991).

A number of explanations have been suggested for the unexpectedly low prevalence estimates in the elderly. Among them are the inappropriateness of the diagnostic criteria in old age (e.g., Blazer, 1989), sample selection effects due to differential mortality and institutionalization, and a failure to recall episodes that occurred in the remote past (Giuffra & Risch, 1994; Simon & VonKorff, 1992). It was also suggested that older respondents may be less likely to view a symptom as a psychological problem or to think in psychological terms (Davies, Sieber, & Hunt, 1994; Hasin & Link, 1988) and may be more likely to view it as physically caused (Ray, Raciti, & MacLean, 1992). Blazer (1989) suggests that cases of depression in the elderly may be missed or labelled incorrectly due to a masking of depression by somatic symptoms. As described above, physically caused symptoms of depression do not count towards a diagnosis of depression; thus, this latter possibility could contribute to an underestimation of depression in older age. In a reanalysis of epidemiological data generated by the 1981 Munich Follow-up Study, based on a representative sample of the Western German population (Wittchen & von Zerssen, 1987), we demonstrated that older respondents indeed tend to attribute experienced symptoms to a physical illness (Knäuper & Wittchen, 1994). Furthermore, this response behavior was not restricted to individuals with poor current physical health, but occurred independently of the current physical health status of the respondents. It was concluded that the attribution of symptoms to a physical illness or condition, at least in part, explains the low depression prevalence rates in the elderly.

But why do older respondents often report a physical causation of symptoms? Recent research indicates that respondents tend to resort to the use of heuristics when they are not sufficiently motivated or able to perform the complex comprehension and memory tasks involved in questions (Bless, Bohner, Hild, & Schwarz, 1992; Krosnick, 1991). Attributing vague emotional symptoms to a physical illness may be a heuristic strategy respondents use when the complexity of the question exceeds their cognitive capacity. Initial exploratory investigations support this assumption: Respondents with low working memory capacity were particularly likely to attribute symptoms to a physical illness (Knäuper & Wittchen, 1994).

The present paper further investigates this assumption by examining the interaction between respondents' age and question difficulty. The heuristic strategy of attributing symptoms to a physical illness should be particularly likely to be used for difficult symptom questions. Questions can be assumed to be increasingly difficult the more time-related and severity information they contain that has to be considered. It is predicted that older respondents use the physical attribution strategy more often for difficult than for easy questions. The response behavior of younger adults, on the other hand, should not be affected by question difficulty.

Methods

Assessment of Question Difficulty

To assess question difficulty, the questions were analyzed according to a propositional complexity analysis suggested by Kintsch and van Dijk (1978). This procedure basically counts the number of information units in a text or sentence. Thus, every additional severity or time-related piece of information increases the propositional complexity score of the question.

Sample and Control Measures

Thirty-one older (55–75 years) and 32 younger (25–40 years) German adults from the community participated in a laboratory study. Half in each group were female. There were no differences between age groups in verbal ability, the tendency to answer in a socially desirable manner (as assessed by the K-scale of the Minnesota Multiphasic Personality Inventory), and self-reported current depressive symptomatology as measured by a depressive symptom screening scale (Depressionsskala, D-S' [von Zerssen, 1976]). Younger adults had about 2 years' more formal education than older adults. Several aspects of physical health and medication for somatic illnesses were assessed to serve as a measure of physical disability. More physical health problems were reported with increasing age ($r = .53$, $df = 61$, $p < .000$).

Results

First, the frequency of endorsements of the symptom questions was explored as a function of question difficulty (see Table 1). As can be seen, older people's responses vary only slightly and nonsignificantly as a function of question difficulty ($X^2 < 1$). Younger respondents, however, are less likely to endorse symptoms assessed by questions that are high, rather than moderate or low, in difficulty (X^2 [1, N = 32] = 2.84, $p < .05$). This suggests that older respondents are less likely to consider all of the information provided in the question. Of course, questions

Table 1. Percentage "yes" responses to symptom questions by question complexity and age (N = 63)

	Younger (n = 32)		Older (n = 31)	
Question complexity	%	SD	%	SD
Low	27.4	18.1	27.4	20.4
Middle	28.0	18.4	29.3	24.3
High	24.1	19.0	30.1	26.1

that contain a high number of time-related and severity criteria (i.e., difficult questions) by nature refer to something more serious and should in general be less likely to be endorsed.

As can be seen in Table 2, older respondents were overall more likely to report a physical cause for experienced symptoms than were younger respondents. However, as expected, older respondents showed this response behavior primarily for highly difficult questions compared with the other conditions (X^2 [1, N = 31] = 5.19, $p < .023$). For younger adults, no variation by question difficulty was found ($X^2 < 1$). These results support the assumption that attributing symptoms to a physical illness is a heuristic that older respondents use to simplify complex answering processes when the complexity of the task exceeds their cognitive capacity.

Discussion and Implications

The findings demonstrate that older respondents are more likely than younger respondents to attribute experienced symptoms to a physical illness. That this response behavior was found primarily for difficult questions supports the assumption that physical causation is a simple, plausible, and highly available explanation they resort to because they are overtaxed by the complex comprehension, memory, and judgments tasks posed by the standardized diagnostic interview questions. As described above, the attribution of symptoms to physical instead of psychological causation is crucial for a diagnosis for major depression. Symptoms that are entirely attributed to a physical illness are excluded from a diagnosis. Thus, the use of this response simplification strategy contributes to the most likely erroneous conclusion that the lifetime and current prevalence of major depression of older people is lower than in other age groups.

Fully standardized diagnostic interviews such as the CIDI reflect only the respondent's own subjective judgment—and perhaps his or her misunderstanding of questions. These interviews are therefore vulnerable to attribution biases, such as the one reported here. The observed response behavior should be less pronounced or should even not occur in more loosely structured clinical diagnostic interviews. These interviews allow the investigator to use a wider range of flexible and possibly individualized questions in an attempt to adapt the symptom questions to the respondents' capabilities. Also, the investigator can use clinical judgment to weight the respondents' answers about whether or not a reported symptom was caused by a physical illness or condition. The use of clinical interviews, some of which have been developed especially for the assessment of depression in the elderly (e.g., the Comprehensive Assessment and Referral Evaluation Depression Scale [CARE; Gurland et al., 1977] or the Geriatric Mental State Interview [GMS; Copeland et al., 1976]), often results in higher prevalence estimates of major depression (Blazer & Williams, 1980; Kay, Henderson, Scott, Wilson, Rickwood, Grayson, 1985; see Wittchen, Knäuper, & Kessler, 1994, for a review).

Table 2. Percentage of symptoms entirely attributed to physical illnesses by symptom question complexity and age (N = 63)

Question complexity	Younger (n = 32)	Older (n = 31)
Low	0.7	2.6
Middle	0.5	2.2
High	0.6	8.1

In sum, the age cohort hypothesis of depression suggested by several recent epidemiological studies reflects, at least in part, differences in respondents' task performance. This illustrates how an insufficient understanding of the cognitive processes that underlie respondents' reports may bias epidemiological data. So far, the primary focus in the development and improvement of diagnostic interviews has been to achieve consensus on the diagnostic criteria and to create questions that contain all these criteria. Possible contributions of cognitive psychology and survey methodology to question answering in general and the understanding of symptom reports in particular have been largely ignored in this process. Cognitive and other methods should be used to identify and remedy problems such as those illustrated in this paper. The validity of mental health self-reports could most likely be substantially improved by considering and applying some basic methodological principles in revising the diagnostic instruments. For example, the complex and often vague diagnostic criteria need to be translated into easier and nonambiguous symptom and probe questions. Based on insights of research into autobiographical memory, improved methodologies for the retrospective assessment of lifetime prevalence rates of mental disorders need to be developed. In doing so, the limitations of respondents' memory and motivation have to be taken into account.

References

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd ed., revised). Washington, DC: American Psychiatric Association.

Berkman, L. F., Berkman, C. S., Kasl, S., Freeman, D. H., Leo, L., Ostfeld, A. M., Cornoni-Huntley, J., & Brody, J. A. (1986). Depressive symptoms in relation to physical health and functioning in the elderly. American Journal of Epidemiology, 124, 372–388.

Blazer, D. (1989). The epidemiology of depression in late life. Journal of Geriatric Psychiatry, 22, 35–52.

Blazer, D., Burchett, B., Service, C., & George, L. K. (1991). The association of age and depression among the elderly: An epidemiologic exploration. Journal of Gerontology: Medical Sciences, 46, M210–M215.

Blazer, D., & Williams, C. D. (1980). Epidemiology of dysphoria and depression in an elderly population. American Journal of Psychiatry, 137, 439–444.

Bless, H., Bohner, G., Hild, T., & Schwarz, N. (1992). Asking difficult questions: Task complexity increases the impact of response alternatives. European Journal of Social Psychology, 22, 309–312.

Copeland, J. R. M., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, J. L., & Sharpe, L. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule. I. Development and Reliability. Psychological Medicine, 6, 439–449.

Cross-National Collaborative Group. (1992). The changing rate of major depression: Cross-national comparisons. Journal of the American Medical Association, 268, 3098–3105.

Davies, R. M., Sieber, K. O., & Hunt, S. L. (1994). Age-cohort differences in treating symptoms of mental illness: A process approach. Psychology and Aging, 9, 446–453.

Gaitz, C., & Scott, J. (1972). Age and the measurement of mental health. Journal of Health and Social Behavior, 13, 55–67.

Giuffra, L. A., & Risch, N. (1994). Diminished recall and the cohort effect of major depression: A simulation study. Psychological Medicine, 24, 375–383.

Gurland, B., Kuriansky, J., Sharpe, L., Simon, R., Stiller, P., & Birkett, P. (1977). The Comprehensive Assessment and Referral Evaluation (CARE)—Rationale, development and reliability: Part II. A factor analysis. International Journal of Aging and Human Development, 8, 9–42.

Hasin, D., & Link, B. (1988). Age and recognition of depression: Implications for a cohort effect in major depression. Psychological Medicine, 18, 683–688.

Kay, D. W. K., Henderson, A. S., Scott, L. R., Wilson, J., Rickwood, D. J., & Grayson, D. (1985). Dementia and depression among the elderly living in the Hobart community: The effect of the diagnostic criteria on the prevalence rates. Psychological Medicine, 15, 771–788.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363–394.

Klerman, G. L., & Weissman, M. M. (1989). Increasing rates of depression. Journal of the American Medical Association, 261, 2229–2235.

Knäuper, B., & Wittchen, H.-U. (1994). Diagnosing major depression in the elderly: Evidence for response bias in standardized diagnostic interviews? Journal of Psychiatric Research, 28(2), 147–164.

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

Ray, D. C., Raciti, M. A., & MacLean, W. E. (1992). Effects of perceived responsibility on help-seeking decisions among elderly persons. Journals of Gerontology: Psychological Sciences, 47, 199–205.

Roberts, R. E., Lee, E. S., & Roberts, C. R. (1991). Changes in prevalence of depressive symptoms in Alameda County: Age, period, and cohort trends. Journal of Aging and Health, 8(1), 66–86.

Robins, L. N., & Regier, D. A. (Eds.). (1991). Psychiatric disorders in America: The epidemiologic catchment area study. New York: Free Press.

Simon, G. E., & VonKorff, M. (1992). Reevaluation of secular trends in depression rates. American Journal of Epidemiology, 135, 1411–1422.

von Zerssen, D. (1976). Clinical Self-rating Scales (CSR-S) of the Munich Psychiatric Information System (PSYCHIS München). In N. Sartorius & T. A. Ban (Eds.), Assessment of depression (pp. 270–303). Berlin: Springer.

Wittchen, H.-U., Knäuper, B., & Kessler, R. (1994). Lifetime risk of depression. British Journal of Psychiatry, 165, 16–22.

Wittchen, H.-U., & von Zerssen, D. (1987). Verläufe behandelter und unbehandelter Depressionen und Angststörungen: Eine klinisch-psychiatrische und epidemiologische Verlaufsuntersuchung. Berlin: Springer.

World Health Organization (Ed.). (1990). Composite International Diagnostic Interview (CIDI): (a) CIDI-interview (Version 1.0), (b) CIDI-user manual, (c) CIDI-training manual, (d) CIDI-computer programs. Geneva: World Health Organization.

World Health Organization (Ed.). (1991). International Classification of Disease (ICD-10). Chapter V. Mental health and behavioral disorders. Geneva: World Health Organization.

Reinterview Methods for Assessing and Improving the Quality of Data From a Medicare Population

Barbara H. Forsyth, D. Kirk Pate, Timothy K. Smith, and Leslye Fitterman

Information available in the Health Care Financing Administration's (HCFA's) administrative data files has been used by HCFA and other researchers for a variety of analytic and policy purposes. While the administrative data have many strengths, they are insufficient to answer some important policy questions. Therefore, HCFA developed plans for the Medicare Beneficiary Health Status Registry, and they selected the Research Triangle Institute (RTI) to design a pilot test to determine the feasibility of key features. The pilot test was conducted in 1993. Findings presented here represent results from the pilot test as reported to HCFA in August 1994 and the opinions of the authors.[1]

The Registry will collect survey information from Medicare enrollees about their health, medical history, functional status, and quality of life. Survey information will be linked with Medicare records for use in understanding and forecasting the health care needs of older Americans. Registry analytic purposes require large sample sizes, and survey cost is an important issue. HCFA selected self-administered mail survey methods for the basic Registry design based on the hypothesis that mail survey methods will yield adequate response rates and data quality without exceeding HCFA's cost constraints. Smith and Biemer (1991) reviewed methodological research on surveys with older populations and concluded that existing evidence is largely mute on the issue of whether a mail survey with older respondents will yield response rates and levels of data quality sufficient for Registry purposes. Thus, the Registry field test was designed to collect information that could be used to test this general hypothesis. The goals of the Registry field test conducted by RTI were (a) to assess the feasibility of the recommended Registry design and (b) to collect empirical information about data quality necessary to make informed choices among alternative Registry design options.

This paper reports some results from a reinterview study conducted as part of the Registry field test. The reinterview study had two purposes. First, we used initial and reinterview responses to develop estimates of temporal response consistency. These estimates gave preliminary information about survey response quality. Second, we used responses to reinterview probe questions to identify self-reported reasons for temporal inconsistencies. We used these responses to evaluate how well questionnaire design activities enhanced item comprehension, memory recall, and response selection.

Initially, we hoped to develop estimates of response bias for Registry variables. As the field test design developed, we found it increasingly difficult to justify the assumption that our reinterview methods produced error free measures of Registry variables—an assumption necessary for estimating response bias. Instead, we focused on developing protocols for probing discrepant responses to collect detailed information about possible causes of response discrepancies.

Registry Reinterview Methods

The reinterview was administered using computer-assisted personal interviewing (CAPI) methods. The CAPI software presented the reinterview in three general steps. First, interviewers used the CAPI software to readminister the initial interview questionnaire to a reinterview respondent. Second, after completing the interview readministration, the CAPI software read a file containing the initial interview responses and identified all items that elicited discrepant responses across the two administrations. Third, after the CAPI software identified response discrepancies, reinterviewers used the software to administer follow-up probe questions tailored to the observed response discrepancies. Interviewer instructions emphasized that inconsistent responses often reflect design problems and that the respondent is the best source of information about problems caused by aspects of question presentation. We hoped that this methodological focus would reduce potential embarrassment and increase willingness to cooperate in identifying causes of response discrepancies.

Registry Field Test Sample Design

A total of 2,510 Medicare beneficiaries were selected to participate in the Registry field test. The Registry field test

Barbara H. Forsyth, D. Kirk Pate, and Timothy K. Smith are at the Research Triangle Institute, Research Triangle Park, North Carolina. Leslye Fitterman is at the Health Care Financing Administration, Baltimore, Maryland.

[1]The conclusions do not represent the position of HCFA or the Department of Health and Human Services in relation to the Registry, and Health Care Financing Administration has requested proposals to undertake further Registry design work.

sample was selected using a two-stage, deeply stratified design. Twenty-seven primary sampling units were selected, and three of these (Richmond, Virginia; Raleigh, North Carolina; and Atlanta, Georgia) were purposively selected for the reinterview study because of their proximity to RTI in North Carolina. The second-stage selection yielded two age cohorts: a younger cohort of new Medicare enrollees who were 65 years old and an older cohort of beneficiaries who were between 76 and 80 years old. The second-stage frame was stratified by race (white or nonwhite) and gender. The older cohort was also stratified by a health indicator. The final step in sample selection involved randomly assigning sampled beneficiaries to several experimental treatments embedded in the field test design. Six hundred and twenty-seven Medicare beneficiaries were selected to participate in the reinterview component of the Registry field test.

Initial interview responses were collected using self-administered mail survey methods. We contacted initial nonrespondents by mail with two additional mailings before continuing nonresponse follow-up by telephone. Interviewers completed CAPI reinterviews an average of 27 days after initial data were received from respondents. In cases in which initial interviews were completed by proxy respondents, interviewers attempted to complete reinterviews with the same proxy respondents.

Reinterview Questionnaire

The Registry field test questionnaire covered five domains: health behaviors and lifestyle, functional status, medical history, quality of life, and sociodemographics. Discrepancy follow-up probe items were designed to collect self-reported reasons for inconsistent responses, and we designed follow-up probe sets that were tailored to each of six specific types of response inconsistencies. For example, one set of follow-up probes was administered when respondents gave an answer during one interview but gave no answer during the other. A different set of probe items was administered when respondents gave two different answers during the initial and reinterviews.

In general, we used open-ended question formats to collect information about question and interview features that contributed to response consistencies. Interviewers asked the open-ended questions and recorded respondents' verbatim responses. After the interviewer entered the full verbatim response, a CAPI code screen appeared, and interviewers selected codes from the list to describe respondents' open-ended responses. We developed 21 codes to characterize discrepancy follow-up responses. Several of the codes were relevant to some types of response discrepancies but not to others. Therefore, both the probe questions and the code screen were tailored to fit the six types of inconsistencies that we identified.

Reinterview Response Rates

The original reinterview sample consisted of 627 Medicare beneficiaries. Among the sampled beneficiaries, we identified 126 who were ineligible for reinterviews, yielding a final reinterview sample size of 501. Of these sampled beneficiaries, 433 responded to the reinterview, yielding an overall reinterview response rate of 86.4%.

We selected the proportion of consistent responses as a measure of response consistency for simplicity and because it enabled general summaries and comparisons across variables. Under the assumption of independence between initial interview and reinterview responses, the proportion of consistent responses is proportional to an unbiased estimate of the simple response variance for dichotomous variables (Biemer & Stokes, 1991). Rodgers, Billy, and Udry (1982) noted additional advantages to using the proportion of consistent responses as a measure of response consistency. Notably, the proportion of consistent responses is an empirically defined measure of consistency that requires no implicit assumptions about response errors or response error distributions.

Item Consistency Rates

We used a consistency index, Cij, to examine mean response consistency. Cij was assigned a value of 1 when respondent i gave consistent responses for variable j during both interviews, and Cij was assigned a value of 0 when respondent i gave inconsistent responses for variable j during the two interviews. Then, C.j is the proportion of consistent responses for variable j computed across the i respondents. Item consistency rates were generally high. The median item consistency rate for the younger cohort was approximately 0.92, and the median consistency rate for the older cohort was approximately 0.83.

Response Consistency and Item Content

Project staff at HCFA sorted Registry questionnaire items into 14 general content areas, based on a combination of substantive and policy considerations. All items were assigned to at least one content area, and some items were assigned to more than one content area. We computed response consistency rates, Cijk, for each of i individuals and each of j variables in each of k content areas. We used these consistency rates to estimate item consistency rates, C.jk, for variables assigned to each content area. Then mean consistency rates, C..k, were computed across the items in each content area.

For both age cohorts, mean response consistency rates exceed 85.0% in the areas of surgery, medical conditions, tobacco, and male health. Additional records validation results conducted as part of the Registry field test (Smith,

Turner, & Fitterman, 1995) indicate that respondents' reports of surgical procedures closely match their medical records. Thus, for surgical procedures, respondents gave responses that were both consistent and accurate. There was less correspondence between respondents' reports of medical conditions and information in their medical reports. Thus, for these medical conditions, respondents' reports were consistent but did not match records well, due either to idiosyncracies in the way diagnosis codes are assigned and used in medical records systems, respondent difficulties understanding the questions and conditions, or both (Smith et al., 1995).

Items in the areas of prevention and female health achieved high levels of response consistency within the younger cohort (81.6% and 86.5%, respectively), but mean consistency rates for these content areas were only moderate for the older cohort (71.0% and 74.2%, respectively). The low levels of response consistency for these two content areas within the older cohort were due to two items that were assigned to both content areas. Response consistency for the item on recency of last Pap smear was 78.5% for the younger cohort and 55.7% for the older cohort, and response consistency for the item on recency of last mammogram was 83.9% for the younger cohort and 69.4% for the older cohort. Decreased response consistency with age might be expected if these items seem more sensitive to older respondents or if recency items are particularly difficult for older respondents to answer. Both hypotheses, and probably others, are consistent with the reinterview data presented here. More detailed data collection and validation efforts are necessary to distinguish them.

Response consistency rates for three content areas were low for both age cohorts. Mean consistency rates for items on current health status and current mental health status ranged from 57.9% to 68.5%, and response consistency for items on health insurance coverage were 20.8% and 23.0% for the younger and older cohorts, respectively. It is important to note that relatively low consistency rates for items on current health and mental health status do not necessarily reflect low data quality. Response inconsistencies across interviews may accurately reflect true changes in the respondent rather than failings of Registry measurement procedures.

Early questionnaire pretest research indicated that respondents were unfamiliar with some of the terminology used in the health insurance items (e.g., Civilian Health and Medical Program, Veteran's Affairs [CHAMPVA], HMO), and some pretest participants indicated they did not know much about the details of their health insurance coverage. The Registry field test questionnaires aimed to address these issues by providing simple definitions and memory cues to help respondents answer the insurance items accurately. The low levels of response consistency may suggest that these revisions were insufficient. On the other hand, low consistency rates for the health insurance items may reflect increased response accuracy in the reinterview if the initial interview prompted respondents to pay more attention to details of their health insurance coverage. The small records validation study conducted under the Registry field test did not include records documenting health insurance coverage. We expect that the full Registry design will provide opportunities for small validation studies, such as a study to validate respondent reports on health insurance coverage.

Inconsistent Response Follow-up

Two patterns of inconsistent response accounted for most of the observed inconsistencies (see Table 1). Eighty percent of the observed response inconsistencies occurred when respondents gave different answers during the initial interview and the reinterview, and 15% of the observed inconsistencies occurred when respondents left an item blank or refused to answer it in one interview but gave a substantive response to the same item during the other interview. Notably, there were very few occasions on which respondents gave uncodable mail survey responses. There were also very few cases in which respondents selected multiple responses when items required only one response.

We computed the distribution for codes assigned to self-reported reasons for discrepant responses (see Table 2). Codes for difficulties understanding the questions and codes for difficulties understanding and using item response categories were most frequent, representing 22.3% and 27.7% of the assigned codes, respectively. "Other" uncodable responses were documented for 15% of the open-ended responses, and the code assigned when respondents reported changing their mind was assigned to 13% of the open-ended responses. Relatively few open-ended responses were coded for skip errors or other instructional errors. The Registry field test questionnaires were specifically designed to minimize the impact of skip errors on data quality. The low frequencies for the skip and instruction error codes suggest that these design features were effective in enhancing data quality.

Table 1. Unweighted distribution (and standard errors) of inconsistent interview-reinterview responses by inconsistency type

Inconsistency type	No.	%[a]
Blank or refused both times	8	0.3 (0.1)
Answered once and blank or refused once	410	15.3 (2.7)
Answered once and "don't know" once	95	3.5 (0.6)
Different answers at different times	2,152	80.2 (2.6)
Uncodable response to mail questionnaire	1	0.1 (0.1)
Multiple responses to mail questionnaire	18	0.7 (0.2)
Total	2,684	100.0 (0.0)

[a]Unweighted standard errors are given in parentheses. Note that these standard errors should be interpreted with caution, since individual respondents contributed multiple observations to each estimate.

Table 2. Unweighted distribution of codes describing respondent reasons for inconsistent responses

Inconsistency code	No. assigned	% all assigned codes
Understanding question	599	22.3 (1.5)
Total	599	22.3 (1.5)
Knowledge, recall, sensitivity	74	2.8 (0.6)
Insufficient knowledge	39	1.4 (0.4)
New knowledge	98	3.6 (0.5)
Sensitivity	26	1.0 (0.3)
Total	237	8.8 (1.0)
Response categories	2	0.1 (0.1)
Multiple suitable	265	9.9 (0.9)
None suitable	26	1.0 (0.2)
Comprehension or selection	450	16.8 (1.3)
Total	743	27.7 (1.5)
Recording response	0	0.0 (0.0)
Skip	178	6.6 (2.4)
Other instruction error	60	2.2 (0.9)
Total	238	8.8 (2.4)
Assistance received	1	0.1 (0.1)
Interviewer assistance	7	0.3 (0.1)
Other assistance	26	1.0 (0.3)
Total	34	1.3 (0.3)
Respondent specific	2	0.1 (0.1)
Burden	14	0.5 (0.3)
Fatigue or physical limit	44	1.6 (0.6)
Refused	0	0.0 (0.0)
Changed mind	353	13.6 (1.5)
Total	413	15.4 (1.6)
Other	418	15.6 (1.5)
Total	418	15.6 (1.5)
Total across codes	2,682	100.0 (0.0)

NOTE: No codes were assigned to two interview-reinterview response inconsistencies.

The most frequently occurring codes are those indicating problems with question wordings and response options. However, the actual number of inconsistencies represented by these codes was small. There were approximately 24,681 pairs of interview and reinterview responses that could have produced inconsistencies. Only 10% of these response pairs actually reflected inconsistencies, and 5% of the total number of response pairs were identified as inconsistent because of difficulties with the questions or response categories. The follow-up probe responses are useful for documenting difficulties and developing improved wordings. At the same time, the relatively low frequency with which respondents identified difficult item and response wordings indicates that pretest and design activities were effective in identifying question and response wordings that contribute to high data quality.

Summary and Conclusions

We designed the Registry field test reinterview study to meet two general goals. First, we used reinterview results to estimate temporal consistency for Registry variables. These estimates were useful in evaluating overall data quality and in making decisions about alternative Registry design options. Second, we used responses to follow-up probe questions to collect information about potential causes of response inconsistencies. We used these results to evaluate questionnaire pretest activities and to identify questionnaire features that interfered with data quality.

Item consistency rates were higher for the younger cohort than for the older cohort, but mean consistency rates were acceptably high for both cohorts. The generally high consistency supported the more general conclusion that the general Registry design is feasible (Turner, Wheeless, & Witt, 1995). In addition, the generally high level of response consistency suggests that the occasionally low consistency rates observed for some questionnaire items likely reflect item specific measurement problems rather than more general problems related to survey procedures or data collection with older respondents.

Response inconsistencies occurred relatively infrequently compared with the total number of responses that could be inconsistent. A major portion of the observed inconsistencies occurred when respondents answered items differently in the two interviews. A smaller number of inconsistencies occurred because respondents refused or failed to answer an item in one interview. Very few inconsistencies resulted from uncodable mail survey responses, suggesting extensive questionnaire design efforts contributed to data quality. Respondents cited difficulties understanding questionnaire items and difficulties understanding and using response options as the most frequent reasons for inconsistent response. These codes accounted for no more than 5% of the total number of reinterview responses, further testifying to the beneficial effects of early questionnaire design activities.

References

Biemer, P. P., & Stokes, S. L. (1991). Approaches to the modeling of measurement errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement errors in surveys (pp. 487–516). New York: Wiley.

Rodgers, J. L., Billy, J. O., & Udry, J. R. (1982). The rescission of behaviors: Inconsistent responses in adolescent sexuality data. Social Science Research, 11, 280–296.

Smith, T. K., & Biemer, P. P. (1991). Literature review of mail surveys of the elderly. Research Triangle Park, NC: RTI.

Smith, T. K., Turner, C. F., & Fitterman, L. (1995). The quality of health data obtained in surveys of the elderly: A validation study of the Medicare Beneficiary Health Status Registry (MBHSR). Manuscript submitted for publication.

Turner, C. F., Wheeless, S., & Witt, M. B. (1995). Feasibility of large-scale surveys of the elderly using mail data collection with telephone follow-up: Results from the national field test for the Medicare Beneficiary Health Status Registry (MBHSR). Manuscript submitted for publication.

DISCUSSION PAPER

Strategies for Detecting Survey Error

Floyd J. Fowler Jr.

In a standard methodological section for a report of survey results, the sampling errors and response rates will be duly reported. However, virtually nothing will be reported to allow readers to evaluate issues relevant to response error: how well interviewers did their jobs and, more importantly perhaps, the quality of the survey questions. At least one reason for the absence of such reports is that for most surveys, researchers do not have any information about the quality of the questions they asked, how well their survey procedures were implemented, or how these affect survey error.

Each of the papers in this session addresses some issue relevant to the design of survey questions or data collection procedures. However, an important aspect of these papers to which I want to draw attention is that they demonstrate an innovative and valuable array of strategies for detecting error.

The paper by Johnson et al. is a good application of the cognitively oriented interview for detecting sources of potential error. The focus of the paper is on variation in the way different population groups interpret key terms in survey questions. By asking respondents to describe their understanding of certain words, the authors were able to evaluate the extent to which the meaning of questions varied across population subgroups. In essence, this is a systematic way to measure how standardized a question is in its meaning, which is a reasonable basis on which to say it is a good or poor question.

Another important tool for detecting problems in survey procedures is standardized coding of interviewer and respondent behaviors. Of course, behavior such as giving inadequate answers or not reading questions as worded are not in themselves evidence of survey error.

However, behavior coding has been demonstrated to provide evidence of problems with survey questions. The papers by Belli and Lepkowski and Hill and Lepkowski provide two interesting examples of the potential of behavior coding for studies of survey error. In the Belli and Lepkowski paper, interviewer and respondent behaviors are related to a more direct measure of survey error: correspondence between survey reports and records data. In the Hill and Lepkowski paper, characteristics of respondents and interviewers are related to question-reading behavior. Because behavior coding is simple to do, it provides an important and useful strategy for presurvey question evaluation and for monitoring the quality of data collection during a survey period. Learning more about behavior during interviews, the conditions under which variations from accepted data collection strategies are most likely to occur, and the link between behavior in interviews and error in survey data constitute a very important strategy for evaluating and improving survey data collection procedures.

The methodologies used in the paper by Knäuper and Wittchen and Mack, Blair, and Presser are more traditional. The Mack et al. paper compares four different interview protocols. Two strategies were used for evaluating the results. First, simply the rates at which children reported eating things were compared, under an oft used assumption that more reporting probably constitutes better reporting. In addition, the reports of parents and the results from systematic observation were used as standards against which to compare the accuracy of the children's reporting. Although each of the measures of errors in reporting was imperfect, the fact that the several comparisons tended to converge in showing that the experimental approaches were superior to the initial test approach provides evidence that the validity of data was improving.

The value of a theory when looking for error is particularly evident in the Knäuper and Wittchen paper. One of the contributions of cognitive evaluations of questions over the past decade is the improved sensitivity of researchers to the nature of the task they are giving to respondents. The notion that older respondents would be particularly overloaded by complex questions, which in turn would lead them to attribute reported problems to somatic causes, is not a self-evident hypothesis. The fact that the authors were thinking about question complexity led them to figure out a way analytically to test their ideas and demonstrate that an apparent substantive finding, that psychological distress most often is said to have somatic origins among older respondents, is in fact an artifact of the design of the questions. As more researchers look at questions critically, they undoubtedly can find many opportunities to examine the possible effect of question design on their substantive results.

Finally, I am particularly taken with the Forsyth, Pate, Smith, and Fitterman paper because of its innovative use of reinterview techniques for assessing the quality of survey

Floyd J. Fowler Jr. is a Senior Research Fellow at the Center for Survey Research, University of Massachusetts–Boston.

results. In fact, the strength of the Forsyth et al. project is the integration of reinterviews and cognitive interviews to produce an evaluation of the quality of survey data. This is not the first time this approach has been used. Cannell, Fisher, and Bakker (1965) reinterviewed respondents who failed to report a known hospitalization and included questions about why the known hospitalizations were not reported. However, the Forsyth et al. strategy obviously has much broader application because it can be used for any survey and any type of question, not just those for which there are records against which to compare survey responses.

The project gets high marks for an innovative application of computer-assisted personal interviewing (CAPI) technology as well. By having previous answers stored in the computer but not available to interviewers until after a reinterview was completed, the problem of making the interviewer blind to the initial answers was solved.

Interpretation of an inconsistency between two reports is not without its challenges. One issue is how different two answers have to be before they are labeled inconsistent. A second challenge, obviously, is that some things can change in a short period. Nonetheless, it is noteworthy how often the respondents' explanations for changes in answers were of the sort that could be of great benefit in designing better questions. For example, knowing when respondents said that they understood the question differently the second time or that more than one response category fit their answers is exactly the sort of information a person would like to have in a presurvey evaluation as a basis for designing better questions.

In order to improve our understanding of how question design and data collection procedures contribute to response error, we need good measures of error. The papers in this session provide six examples of how to use a variety of techniques to evaluate the survey data collection process or its results. Records check studies can make a great contribution to our methodological knowledge, but because of their complexity and cost, they will no doubt continue to be rare. In contrast, the key procedures described in these papers can be used much more routinely. Presurvey cognitive evaluation of survey questions is becoming standard in many survey organizations. Behavior coding can be used in field pretests, prior to surveys, and also, as was done in the Michigan study, can be incorporated into an ongoing survey to monitor the data collection process. The results can be used for methodological studies related to question design, interviewer behavior, and respondent behavior. As researchers become more attuned to cognitive aspects of their questions, they undoubtedly will find many opportunities to analyze their data in ways that illuminate unwanted effects of poorly designed questions. Finally, reinterview strategies, combined with respondent debriefings, seem to me to be a particularly underutilized strategy for presurvey and postsurvey evaluation of survey questions. Adding debriefing about inconsistencies to reinterviews adds a great deal to their value as aids to understanding the sources of survey error.

In conclusion, as we try to improve our survey methods, better presurvey and postsurvey evaluations of questions and data collection procedures are needed. If researchers can identify poor questions before they do surveys, they can design better questions. If they can detect sources of response error in their survey data, the results can inform analyses and alert users to limits in the data. Most importantly, when response error is measured, researchers will be more likely to invest in error reduction. These papers provide excellent examples of the kind of techniques that need to be routinely applied in order to improve the quality of our methodological generalizations and our methods.

Reference

Cannell, C. F., Fisher, G., & Bakker, T. (1965). Reporting of hospitalization in the Health Interview Survey. Vital and Health Statistics (Series 2, No. 6). Washington, DC: U.S. Government Printing Office.

Discussion of Research on Health Survey Questions

Norman M. Bradburn

We have heard six excellent papers that cover a wide variety of topics, all of which can be grouped under the rubric of research on health survey questions. Four of the papers address problems in questions with respondents who vary in important characteristics, such as age or ethnicity; the other two focus on the interaction between interviewers and respondents in the interview process. I will discuss the papers from a cognitive, information-processing point of view, concentrating more on their contributions to our understanding of the survey process than on their contributions to data quality, which is more the focus of my fellow discussant.

Because it raises some of the most general questions that those who write survey questions must deal with, I will start with the Johnson et al. paper on cultural differences in interpretation of survey questions. We think of question comprehension as the first cognitive task that respondents engage in during the question-answering process. The goal of the researcher is to find question wording that will be comprehended in the same way by all respondents. We know that this is a utopian goal, and some, like Belson (1981), are pessimistic that one can ever do better than get slightly more than a majority of respondents to understand the question in the same way. The great strides that have been made in the last decade on understanding the cognitive processes involved in answering survey questions make some of us feel that we may be able to do better than that.

Much of recent research (see, for example, Sudman, Bradburn, & Schwarz, 1996) has focused on the effects of internal aspects of the questionnaire on comprehension—aspects such as question order, answer categories, or format. The papers here shift the focus from the questionnaire to the characteristics of respondents and how shared characteristics of respondents, such as ethnicity or age, may influence question comprehension or strategies for answering questions.

When we consider group differences among respondents, we have to consider what the mechanisms are that might produce the observed effects. Several suggest themselves. First, social groups may form linguistic subcultures. The criteria for such subcultures are that they communicate frequently and develop their own "language" or dialect. Such subcultures may be based on characteristics other than ethnicity or native language as, for example, age, occupation, or region. The meanings of words may be literally different within different subcultures. "Doctor" may be restricted to MDs by some but may include a wide range of health providers by others. Or, the other way around, concepts with similar meanings may be expressed in different words by members of different linguistic groups; for example, hypertension may be described as "high blood." Idioms are particularly susceptible to subcultural differences, and English is a particularly idiomatic language, as any English speaker who has tried to learn another language knows well. Thus, idiomatic expressions, such as "the blues," are likely to be particularly sensitive to subcultural differences.

Second, the frequency of experiences among members of different subcultures may appear to affect their comprehension of questions when, in fact, it does not. For example, Johnson et al. report that African Americans are more likely than other groups to have thought about potato chips when reporting on potato consumption. This difference seems to be due to differences in consumption patterns rather than in understanding what is meant by the word "potato" in the question, "About how many times do you eat potatoes per day or per week?" There are other group differences in what respondents thought about when asked about eating potatoes. But do these differences really result from differences in the comprehension of the questions? The data suggest that the different images evoked by the question stimulus do not affect their comprehension, since they do not affect the reports of frequency of consumption.

We must be careful, then, to distinguish between differences in actual comprehension of terms in a question and differences in images evoked by a question that are reflections of differential frequency of experience with elements of the terms, such as potato chips, French fries, baked potatoes. In the first case, respondents may exclude consumption of potato chips from reports of frequency of consumption because they incorrectly do not consider potato chips to be potatoes; that is, their comprehension of the question is wrong. In the second case, there is no problem of comprehension. They do not think about potato chips because they do not, in fact, eat potato chips.

When one observes differences between subcultural groups in behavioral reports, it may be difficult to know

Norman M. Bradburn is the Tiffany and Margaret Blake Distinguished Service Professor, Department of Psychology and the Harris Graduate School of Public Policy Studies, University of Chicago, and Senior Vice President for Research at the National Opinion Research Center.

which interpretation is correct. For example, Johnson et al. report that more non-MD visits are reported by non-Hispanic whites than by other respondents. Is this because of subcultural differences in understanding the term "medical doctor"? Apparently not, because when income is controlled, the cultural differences are eliminated.

Symptom reporting, like many subjective phenomena, is a particularly difficult problem for researchers. Not only is there the comprehension question just discussed, but also, there may be group differences in strategies used to recall the information asked for in the question, particularly when questions are difficult and require considerable cognitive effort to answer. Knäuper and Wittchen show differences in strategies used by younger and older respondents in answering difficult questions about the frequency of experiencing different psychological symptoms consistent with the hypothesis that older respondents are more likely than younger respondents to use simplifying heuristics to answer difficult questions.

There are several competing hypotheses, however, that might explain their finding about differences in attributing mental symptoms to physical causes. The first lies in differential frequency of physical problems. As people age, they experience more chronic physical problems, which often have some psychological consequences. Because of the greater frequency of chronic physical problems, many of which are not treated, or, if treated, are treated by medicines that may also produce side effects on mental processes, older people may be reporting experiences related to their understanding of the questions, that is, about things that are due to physical causes.

Another possible alternative hypothesis is that there are some real cohort effects on the comprehension of the causality of mental problems. Younger people grew up in an environment that attributes many more symptoms to psychological causes than the environment older respondents grew up in. Generations might in fact represent quite different linguistic subcultures that have different understandings of the etiology of "mental" problems. I do not know if their data could shed any light on these alternative hypotheses.

Mack, Blair, and Presser address the difficult problem of recall strategies and how these might differ for children who have a less well developed sense of time than adults. While we know from studies of adult autobiographical memory (Barsalou, 1988) that experiences are stored as event sequences rather than as discrete events and that they are coded in such categories as location, activity, participants, and so forth, no category seems to be superior to any other in facilitating accurate recall of events. Even though time is a major organizing principle of memory, the time that an event happened is generally a poor retrieval cue. Mack et al. explored the use of location, activity, and free recall as strategies to contrast with the time-based strategy used in the standard U.S. Department of Agriculture surveys. As with adults, no strategy seemed to be consistently superior. All three experimental strategies produced more reports of food items eaten than did the standard format. If one is willing to accept the increased reports as valid, then the use of strategies other than time would appear to be better.

When the different strategies were compared for accuracy for the meal that could be observed by a third party, there were few differences in accuracy, except that meals as cues seemed to produce the least number of accurate matches. This difference may be due to the problem of distinctiveness that plagues accurate recall. In general, similar events that are in a series are difficult to remember accurately because respondents tend to confuse different instances of the event. Thus, events like meals in school tend to blur with one another, and it is difficult to remember what one ate on which day, unless there is some known pattern, such as pizza is always served on Wednesdays.

I think that this difficulty in differentiating similar events is the cause of some of the response inconsistencies reported by Forsyth, Pate, Smith, and Fitterman. Older respondents were particularly inconsistent in their reports of Pap smears and mammograms, which are repetitive events in a series, while everyone did best on reports of distinctive items, such as surgical procedures. The differences between the consistency exhibited by younger and older respondents in reports of Pap smears and mammograms may be due to the greater experience of older respondents with the procedures and hence their being less distinctive.

Cognitively difficult items such as assets and health insurance coverage were also less consistently reported for all respondents, as might be expected from the sheer cognitive difficulty of recalling the material.

The Hill and Lepkowski paper rightly calls attention to the dynamic nature of the interview process, that is, that the interview is an ongoing conversation. We need to look at responses as contingent on what went on before in the interview. They also introduce us to some statistical techniques for analyzing these contingencies. It is such a rich paper that I can only raise a few questions that intrigued me. One concerns the complex finding regarding the race of interviewer. When both interviewer and respondent were African American, there were more incidents of question-wording change, but the changes were smaller than when the interviewer was African American but the respondent was not. In those cases, there were fewer incidents of wording changes, but when they occurred, they were larger. Is this an indicator that there are cultural differences in understanding of questions? When both interviewer and respondent are African American, do the interviewers, recognizing a difference in understanding, alter the questions to make them more understandable to the respondent? In this case, they may recognize a number of such instances but be able to get the idea across with relatively small changes in the questions. When an African American interviewer is interviewing a non–African American, there may be only a few instances when a misunderstanding is recognized, but when a misunderstanding does occur, it may take more discussion and change to work out what is going on. If something like this were going on, analyzing the differences between the behavior of

interviewers when they interview members of their own culture and when they interview members of other groups might help pinpoint problems in comprehension of the questions.

Finally, I have one comment on the Belli and Lepkowski paper. Somewhat surprisingly, they found that each of their variables produced overreporting, which they ascribe to the failure of the medical records to capture all of the health care visits. While there are, no doubt, some errors in the records and respondents may have been reporting visits to other providers whose records the authors did not have access to, I suspect that there is significant telescoping going on. Since this is a cross-sectional survey and there was no attempt to bound the responses, we would expect a fair amount of telescoping to occur that could well produce overreporting.

In conclusion, let me say again that these are stimulating papers that, as with all good papers, settle a number of questions and raise more.

References

Barsalou, L. W. (1988). The content and organization of autobiographical memory. In U. Neisser & E. Winograd (Eds.), Remembering reconsidered: Ecological and traditional approaches to the study of memory. New York: Cambridge University Press.

Belson, W. (1981). The design and understanding of survey questions. Aldershot, England: Gower.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Discussion Themes From Session 2

Marcie Cynamon, Rapporteur, and Johnny Blair, Chair

Need for a Theory of Question-Wording Effects

Research on the measurement of question-wording effects has as long a history as almost any survey methodological area. Still, the difficulty of assessing the influence of question wording on responses persists. This difficulty is due, among other things, to the lack of a comprehensive theory of question-wording effects, the problem of separating wording effects from those due to other factors (such as the role in question interpretation of conversational norms or the social context of the interview), and vagaries of the measurement process itself. These factors have particular manifestations in health surveys that may call both for special research approaches as well as modification of the goals of health survey interviews, at least for some populations.

Psychometric Issues

The detection and measurement of question-wording changes—in the presence of confounding or competing explanatory factors—and their effect on response variance and bias is fundamental to investigating data quality. When there are competing explanations for an observed effect, it is tempting to let the choice of explanation be influenced by preconceptions about the population.

Validation Issues

In a medical records check study of physician visits, both telescoping into the reference period and forgetting visits that did occur may be present. The net effect of these factors is known, but the contribution of each may be difficult to determine when exact records matches are unclear. For example, assume that records show 10 visits in a reference period and 9 are reported in the interview. This report could result from either forgetting 6 and telescoping 5 or from forgetting 4 and telescoping 3. When the study participants are elderly, there may be an unwarranted tendency on the part of the researchers to give more weight to forgetting.

Additionally, in such a study, there may be visits that occurred that are not in the records, such as visits to a provider outside the system of records used for validation. The researcher may tend to attribute the resultant overreporting to response error when, in fact, it represents true behavior. Further, it is inappropriate to put unqualified reliance on the accuracy of medical records, which are also subject to numerous errors.

Another instance of incorrect attribution of reasons for error can occur when in an effort to be diagnostically precise, we excessively complicate the response task. For example, if respondents, as much research shows, have difficulty understanding relatively common words, they certainly will have problems with technical terms. When language borrowed from diagnostic instruments is used in survey questionnaires or when such terms have both an everyday meaning and a technical one, serious levels of response error may follow. While we may succeed in developing questions that are diagnostically precise, they are not necessarily cognitively precise. Confounding this source of error is faulty recall of autobiographical episodes, as when in a diagnostic interview, respondents are asked whether they have ever had a period in their lives when they were depressed. This may lead to the respondent not attempting such a careful memory search, but to simply reporting those instances that most readily are remembered.

Interviewer Effects on Survey Questions

Response errors also occur when interviewers misread questions, although there is disagreement about the extent to which this affects data quality. Wording changes are often identified by coding the interaction between the interviewer and the respondent. This behavior coding is itself a measurement that is subject both to variability and bias. Such measurement effects in coding could result from insufficient sensitivity of the coding protocol or to variability in coder performance. Coder variability can be an important analysis variable. It is less useful if it is subject to high measurement error. In such cases, its use as an

Marcie Cynamon is with the Division of Health Interview Statistics at National Center for Health Statistics, Hyattsville, Maryland. Johnny Blair is Associate Director of the Survey Research Center at the University of Maryland, College Park.

explanatory variable would not test significant in a regression model.

Themes to Be Pursued in Future Research

1. Future methods research on wording effects needs to address many of these complex questions. This requires developing better measurement procedures for detecting measurement errors resulting from question design and for separating them from other error sources.

2. Behavior coding and cognitive assessments can be used independently to improve data quality; however, their combined use may be even more effective. Innovative uses of known techniques such as reinterviews might also contribute to this effort. If the results of these tests were routinely made available to data users, this would provide insights into the strengths and limitations of the data.

3. Insufficient research has been done on the causes and effects of item nonresponse in mail surveys. Items with high levels of missing data are often simply not used in analysis. We need to investigate both why item nonresponse occurs and its impact on the subsequent questions and on the analysis.

4. As questionnaires grow in length and complexity, we need to remain sensitive to respondent burden, especially as it affects major subgroups. For example, the difficulty of detecting episodes of depression among the elderly may not be because the elderly are less able respondents than others but because of unrealistic demands for retrieval from long-term memory.

5. More attention needs to be given to the implications of a particular line of research for both the individual project's analysis and for what it contributes to the relevant area of methodology. Simply identifying sources of error is, in itself, of limited value. We need to spend more time helping researchers understand how particular error sources affect survey findings.

6. An important context for all of these research concerns is the recent introduction of new modes of data collection, such as computer-assisted personal interviewing (CAPI), computer-assisted self-interviewing (CASI), and audio computer-assisted self-interviewing (ACASI). Methodologies for measuring mode effects need to be expanded to encompass these new technologies.

7. The preferred strategy for dealing with error is to avoid it initially or nullify its effects in analysis. If it can't be done, then it should be measured.

Sampling and Cooperation

Sampling and cooperation, as noted in the discussion of the papers in this session, are really two aspects of the problem of matching a design that optimizes quality with costs. In this session, six papers focus on two aspects of this issue. The first aspect is the problem of selecting sampling frames that are efficient and yet minimize the errors due to both sampling and nonsampling. All of the papers describe creative ways to maximize generalizability and yet reduce costs. The second aspect, intimately related to the first, is the problem of access. These papers describe designs that attempt to obtain data from very hard-to-contact or difficult-to-interview populations. Even if problems of enumerating these populations are resolved through very creative strategies to create sampling frames, the problems of access remain and present real threats to the validity of the final results.

Use of Probability Versus Convenience Samples of Street Prostitutes for Research on Sexually Transmitted Diseases and HIV Risk Behaviors: How Much Does It Matter?

Sandra H. Berry, Naihua Duan, and David E. Kanouse

HIV infection, sexually transmitted diseases (STDs), and drug use have large health consequences both for the individuals who are directly involved and for the larger society. Researchers and public health workers are understandably interested in the health and the health-related behaviors of populations that are at risk for these problems, but these populations are often difficult to study. Barriers include structural features, such as not residing at stable addresses with listed telephones; legal impediments, such as being involved in illegal behaviors; and social factors, such as stigmatization. These factors make it difficult to locate individuals in the population of interest, to construct complete sampling frames, to gain access to individuals for data collection, and to obtain their cooperation for the research. Often, researchers must fall back on reaching these populations in ways that they know are incomplete or unsystematic, such as by using convenience samples of various kinds. Researchers know or suspect there are biases that result and often can make some guesses about direction but are usually unable to quantify the size.

Studying prostitutes is an excellent example of this problem. Prostitution is illegal almost everywhere in the United States, and it is generally viewed as socially stigmatized behavior. Further, despite the venerability of prostitution as a widespread occupation, there is very little formal organization. Street prostitution requires very little in the way of facilities to conduct business. Virtually any public area will do for meeting clients (although there are regular patterns), and there is rarely any form of official certification program. Even informal organizations of prostitutes claim to include only a small proportion of the active population. Virtually all research on street prostitutes has been carried out on convenience samples, and, for example, HIV seroprevalence estimates for Centers for Disease Control and Prevention–sponsored studies have ranged from 0% to 47%, depending on the origin of sample and the city. Samples have been obtained from jails, drug treatment programs, STD clinics, and other locations, such as legal brothels and streets selected according to convenience. Without more information, however, it is impossible to do any more than speculate on how these samples relate to each other or the larger population from which they are drawn.

In contrast, the 1989–90 Los Angeles Women's Health Risk Study collected data from a probability sample of 998 street prostitutes in Los Angeles County. Briefly, the sampling method included systematic identification of all geographic areas where street prostitution was routinely taking place and a random selection of days and times for data collectors to visit those areas and collect data. In each area, the data collector followed a protocol for randomly selecting for interviewing and drawing blood. In addition to questions needed to weight the sample and questions about background, risk behaviors, and other related issues, questions were asked that allowed classification of respondents by the degree to which they were at risk for inclusion in several kinds of convenience samples. These included questions about contact with the justice system (arrests, convictions, and incarcerations); contact with the health system (STD clinics, regular source of care); and contact with drug treatment programs, including methadone maintenance programs. We then constructed synthetic subsamples representing the sets of women who could have been sampled through various convenience sampling methods.

In this paper, we compare the estimates obtained from the probability sample[1] with estimates that might have been obtained using various convenience sampling methods to access the same population of women. We examine the resulting estimates in terms of demographic differences, differences in work characteristics (e.g., clients per week), risk behaviors (e.g., drug use), things that might minimize risk (e.g., use of condoms) and estimates of prevalence of STDs (including syphilis, and hepatitis B).[2] We find that

Sandra H. Berry is Director of the Survey Research Group at RAND in Santa Monica, California. Naihua Duan is a Senior Statistician and David E. Kanouse is a Senior Behavioral Scientist at RAND.

[1] We used selection probabilities to weight the sample to take into account the reported extent of prostitution activity. However, the resulting weights changed the estimates only slightly, and the weighted data were much more difficult to work with analytically. For this analysis, we are reporting results based on the unweighted data.

[2] Syphilis and hepatitis B are generally considered marker conditions for HIV because they are also STDs. The HIV seropositivity rate in this population was about 3.5%, too low to permit meaningful comparisons at the subgroup level.

there were generally significant differences between estimates from the probability sample and the convenience samples and many differences among the convenience samples. We discuss these differences and their implications for studies of various topics related to prostitutes and their clients.

Methodology

To interview women soliciting customers on the streets of Los Angeles County, we constructed a sampling frame by systematically interviewing informants knowledgeable about street prostitution in various parts of Los Angeles County. We randomly sampled area/day/shifts and sent field teams of interviewers, drivers, and phlebotomists to these locations at the selected times. In each area, the field staff randomly selected women on the street, screened them for study eligibility, conducted interviews, and took blood samples. Field staff also conducted an enumeration of all women who could have been approached for screening in that area/day/shift.

Sampling Frame

Our target population consisted of women soliciting customers on the streets of Los Angeles County, a 4,000-square mile area. We narrowed the frame to areas (street segments and adjacent side streets, parking lots, etc.) where street solicitation was likely to be taking place during the period of the study. We began by compiling an extensive list of possible areas of street prostitution activity, identified in two ways. First, we interviewed a broad range of expert informants, including officers, sergeants, and lieutenants in divisional vice in all 18 Los Angeles City Police Department precincts; persons at all ranks from officers to captains at each of the county's 37 other municipal police departments and 17 county sheriff stations; and STD field investigators in each of the 23 districts served by the Los Angeles County Department of Health Services and two independent health departments serving the cities of Long Beach, Pasadena, and Vernon. We also made extensive use of ethnographic consultants, including outreach workers, current and former prostitutes, and persons familiar with minority or other subcultures and institutions. We asked them to identify areas (street segments) where prostitution activity occurs and to estimate the extent of this activity (the number of women) at various times of day; over 200 interviews were conducted for this purpose. Second, we mapped the locations of marker establishments for adult entertainment or gambling that might attract prostitution activity in the surrounding area. These included gambling clubs, hostess dance halls, adult bookstores, adult theaters and arcades, and strip clubs and topless bars licensed as adult entertainment establishments under city or county ordinances.

We compiled all of this information into a data file of possible street segments and verified the information by making independent on-site enumerations. The timing of prostitution activity varied across areas. Some areas were active in the evenings, others during the lunch hour, and so on. Activity also varied by day of the week. We divided each day of the week into four 6-hour shifts (5 a.m. to 11 a.m., 11 a.m. to 5 p.m., etc.). We estimated probable densities of prostitution activity for each area/day/shift based on informants' reports and the results of our own inspection. We then constructed a sampling frame composed of area/day/shifts where and when prostitution activity occurs. Our estimates of the density of prostitution activity were updated 18 times during the 36-week field period to reflect actual field experience and new information provided by informants. We also periodically canvassed informants and continually canvassed respondents to identify new sample areas.

Selection and Screening of Women

Prostitutes were sampled by choosing a random starting point within the area and then approaching the first woman seen. The interviewer introduced herself and said she was from RAND. She then started the interview: "We're talking to working girls about their health and possible risks to their health like AIDS. We're not from the police and we can pay you for an interview. Could you answer some questions for me now? Have you traded any kind of sex, including sex talk or B and D (bondage and discipline), for money or drugs or anything else of value in the past 12 months? (If yes): Have you been interviewed already by the Los Angeles Women's Health Risk Study?"

Respondents who acknowledged eligibility and said that they had not previously been interviewed were given an informed consent form explaining that participation in the study involved (a) taking part in a 45-minute interview about the way they work, their life experiences, and their health and (b) providing a blood sample to be tested for exposure to the AIDS virus, syphilis, and hepatitis B. They were told that they would not be asked to disclose their names, addresses, or other identifying information. Participants were paid $25 for their participation, an amount that did not represent the actual value of time taken away from work for most participants. The blood test was voluntary; women who declined the blood test but completed the interview were paid the same $25. Test results and posttest counseling were made available upon request.

Enumeration of Sample Area

We enumerated most women on the street in each sampled area/day/shift, except for those that were obviously not prostitutes, for example, those pushing a stroller or carrying groceries. We preferred to rely on self-screening rather than interviewer judgment in determining potential eligibility for the study; however, some women could not, in the interviewer's judgment, be safely or discreetly ap-

proached (e.g., women in the midst of a large group of men). Some women refused screening or denied eligibility for the study. Interviewers were asked, however, to record their (probabilistic) judgments as to whether a sampled woman who refused screening or denied being eligible was actually a prostitute. Demographic characteristics were also recorded for all women approached for screening.

Interviews and Blood Samples

Interviews were conducted in various locations on or accessible from the street, including bus stops or park benches, fast food restaurants, laundromats, parking lots, or interviewers' cars. Unlike some studies of street prostitutes in other cities, we did not routinely use a study van for conducting interviews and collecting blood samples; we felt it would attract unnecessary attention and would jeopardize the safety of field staff and respondents. Beginning in September 1990, we obtained blood samples from 636 respondents, which represents most of the women interviewed during the period of time in which we were able to draw blood. Blood samples were drawn by trained phlebotomists using procedures especially adapted to difficult field situations. Study participants received pretest counseling regarding the blood test after they had completed the interview. Those who provided a blood sample could obtain their test results by calling RAND to arrange for an appointment. Results were stored and retrieved using a personalized identification code constructed at the time of the interview (reproducible if lost).

Sample Coverage Results

We believe that the sample areas identified for the study are a complete list of areas with significant street prostitution activity and that the activity we found there accounts for most of the street prostitution in Los Angeles County.

Of 164 potential street areas named by informants or associated with marker establishments, 111 were judged on the basis of enumeration visits to have sufficient prostitution activity for inclusion in the study. We successfully completed interviews in 79 of these areas, widely dispersed over the 4,000–square mile area of Los Angeles County.[3] Altogether, our interview teams made 1,033 visits to the sample areas, for a yield of about 1.0 interview per visit. For the most part, informants drawn from different backgrounds (law enforcement, health departments, and ethnographic sources) tended to agree on the general locations of prostitution activity. Additions during the course of the study mainly extended or modified the boundaries of known locations.

Although we made an effort to sample active areas at various shifts, not all shifts were sampled in each area. Many areas of street prostitution in Los Angeles County also tend to be associated with drug dealing and gang activity. Consequently, certain sample areas and the late night shift were initially assigned a selection probability of 0 because of safety concerns. However, we did complete systematic observations and counts in these areas and during the late night shift and carried out limited interviewing: 90 interviews were completed in high-risk sample areas, and 11 were completed during the late night shift.[4]

Response Rates for Interviews With Street Sample

The street field operation was quite successful. Although for safety reasons we were not able to approach about a third of the women that were selected for screening,[5] 89% of the women we did approach answered the screening question. Of those women, more than three-quarters acknowledged being eligible for the study, and we were able to complete interviews with 89% of those who acknowledged eligibility. A major question of interest is the completion rate—that is, of all women in the sample area who met our eligibility criteria and were approached for an interview, what percentage actually completed an interview? Our data do not permit a precise answer but do permit us to set bounds. A lower bound is provided by assuming that all women who refused screening or said that they were ineligible were simply refusing to participate. In that case, the completion rate is 998 completed interviews out of 1,629 eligible women approached for the first time (61%). An upper bound may be set by assuming that all women who refused screening or denied eligibility were in fact ineligible, in which case the completion rate is 89% (women interviewed/women screened as eligible). The actual

[3]Most of the areas in which we failed to complete interviews had limited amounts of prostitution activity. In 120 interview visits to these 32 sample areas, we identified a total of 65 potentially eligible women (*M* = 0.5 per visit) and approached 26 for screening (0.2 per visit). Fortunately, we were able to complete interviews in many other low-density sample areas (defined as those in which we identified 1.5 or fewer potentially eligible women per interview visit). Altogether, in 419 interview visits to 57 such low-density sample areas, we identified 272 potentially eligible women, approached 186 women for screening (68%), and completed interviews with 90 women (48% of those approached). This completion percentage is lower than in the sample as a whole (61% of those approached).

[4]We made a total of 141 enumeration only visits during the late night shift and 111 during other shifts for the purpose of estimating the amount of prostitution activity in area/day/shifts where interviews were sparse. These limited enumeration data and data from other shifts in a given sample area can be used to estimate the amount of prostitution activity in area/day/shifts that were removed from the sampling frame for reasons of safety.

[5]We encouraged interviewers not to put their safety or the respondents' safety at risk by approaching and screening women who were, for example, in the company of males at the time they were identified for screening. If the male was a client, the interference in trade might be resented. If the male was not a client, overhearing the conversation between the interviewer and the woman might put the woman or the interviewer at risk. The eligibility status of the women who could not be approached is undetermined.

completion rate probably lies closer to the lower than to the upper bound.

Comparing the Probability Sample With Convenience Samples

Taking the full sample of 998 completed interviews with street prostitutes as a base, we then used self-reported characteristics to construct synthetic samples of women who were at risk for inclusion in possible convenience samples. Table 1 shows the distribution of these characteristics in the unweighted full sample for some key characteristics of the sample that we used to construct subsamples.

We created samples of (a) 678 women who reported having been arrested for solicitation, (b) 436 women who reported having been convicted of solicitation, (c) 102 women who had been in methadone maintenance programs, (d) 178 who had been in any form of drug treatment in the past year, and (e) 232 women who had been treated in an STD clinic in the past year.

These subsamples represent several common approaches to identifying convenience samples of prostitutes to study—women who were incarcerated briefly or for longer periods in jails or prisons for prostitution-related charges, women in methadone maintenance programs or in other forms of drug treatment, and women identified through STD clinics. Note that our subsamples only approximate these convenience samples by identifying women who might have been included using such an approach. Actual convenience samples would be sensitive to other factors, such as length of the field period, exactly how eligibles are identified and approached, payment, confidentiality procedures, and so on. Because they were drawn from a systematically sampled population, we think our subsamples approximate "best case" results from convenience samples and should be interpreted as such.

Table 1. Estimates of characteristics related to inclusion in convenience samples (percentages)

Solicitation and related offenses	
Ever arrested	67.9
Ever convicted	43.1
Drug treatment	
Ever had methadone maintenance treatment	10.1
Any drug treatment last year	17.6
Used STD clinic in last year	22.9

Results

In Table 2, we compare the results for race (percentage African American), reported extent of drug use (percentage reporting intravenous (IV) drug use in the past 6 months and percentage reporting daily crack use), sexual risk be-havior with clients (percentage reporting vaginal sex without a condom in the last week), risk reduction activity (percentage who had a condom to show the interviewer), STD status (percentage syphilis reactive and percentage for whom hepatitis B surface antibody was detected), and level of prostitution activity (mean number of clients reported for the week prior to the interview) in the full sample and in each of the subsamples. Because each convenience sample in this case is a subset of the cases in the full sample, the two samples are not independent. Therefore, the significance tests are based on disjoint subsamples; we use a two-sample t test to compare the variance in the means for the cases in the convenience sample with the variance in the means for the cases that are not in the subsample.

It is readily apparent that there are noticeable and statistically significant differences between the estimates made from the full sample and the estimates of the same

Table 2. Comparison of full sample estimates with estimates from key subsamples

	Full sample (N = 998)	Arrested ever (n = 678)	Convicted ever (n = 436)	Methadone treatment ever (n = 102)	Drug treatment last year (n = 178)	STD clinic last year (n = 232)
African American	68.9	69.0	61.9***	31.4***	61.2*	67.2
IV drug use past 6 months	20.2	23.3**	32.0***	82.4***	35.6***	18.0
Daily crack users	50.1	55.8***	56.2**	40.2*	45.5	48.9
Vaginal sex without a condom with a client in last week	39.4	42.3**	41.4	49.0*	41.9	38.0
Had condom available	29.6	31.4	34.9**	37.3	27.7	28.1
Syphilis reactive	33.7	39.7***	41.3***	30.8	38.9	38.4
Hepatitis B surface antibody was detected	36.4	40.6**	42.7*	69.4***	44.6	44.6*
Mean no. clients in prior week	30.2	32.4	36.9*	42.6	31.5	33.6

NOTE. Sample sizes for each characteristic vary somewhat due to missing data. Blood tests were obtained from 619 women, so results for syphilis and hepatitis B apply to that portion of the samples only. Significance test for "mean no. clients in prior week" was based on the square root of the number of clients per week.

*p ≤ .05. **p ≤ .01. ***p ≤ .001.

parameters made from the subsamples. Out of 40 comparisons, we might expect 2 to show significant differences at the 95% level. In fact, we see 21 comparisons that show significant differences. In considering how well the results from the subsamples represent the results from the full sample, we need to consider two factors: what proportion of the full sample is included in the subsample and how different the cases that are not included in the subsample are from the cases that are included. Only the "arrested ever" subsample includes more than half the cases in the full sample. Despite the large degree of overlap, however, there are significant differences between the cases that are included in the subsample and the cases that are not. Five of the eight variables show significant differences in the means for the two groups.

The rest of the subsamples include less than half the cases in the full sample, and in two of these subsamples, many of the differences in the estimates of the subsample means differ significantly from the means for cases that did not fall into the subsamples. Seven of the eight variables show differences in the "convicted ever" sample and five in the "methadone treatment ever" sample. In contrast, while the "drug treatment last year" and "STD clinic last year" samples include smaller proportions of the total cases, fewer of the estimates of variables show significant differences—only two for the "drug treatment last year" subsample and one for the "STD clinic last year" subsample. It may be that the women who appeared in the 6 months prior to the interview in drug treatment or in an STD clinic are close to a random sample of women who were eligible to appear in these settings. For example, about 98% of the full sample have used drugs, so the ones who appear in treatment may simply be very much like the ones who could have sought drug treatment but who didn't during the 6 months prior to the interview.

The parameters most often estimated with significant error in the subsamples were the proportion who had been engaged in injection drug use in the past 6 months and the proportion of cases that tested positive for hepatitis B. Four of the subsamples yielded subsample estimates that were significantly different from the cases that did not fall into the subsamples. All of the estimates that differed significantly were higher than the full sample mean. The percentage of African Americans and the percentage of daily crack users were incorrectly estimated in three of the subsamples. The percentage who had unprotected vaginal sex with a client in the week prior to the interview and the percentage who were syphilis reactive differed significantly for two of the subsamples, and the percentage who could show a condom to the interviewer and mean number of clients in the past week differed significantly for only one of the samples.

The direction of the differences between the full sample estimates and the subsample estimates is not consistent; however, the largest differences are in the direction of yielding higher estimates of risk factors (injection drug use, a higher mean number of clients per week) or of STDs (hepatitis B) than were found in the full sample.

Discussion

Convenience samples are sometimes the only economically feasible approach to conducting research on some populations, but researchers will always be uncomfortable generalizing from them to the larger population of interest. This research indicates that they should be. It also indicates that careful selection of a convenience sampling approach and careful operationalization of the sampling procedures can yield decent results, especially if there is good information from other sources about the population being studied. However, it is difficult to obtain the kind of information that would allow a researcher to make judgments about the accuracy of possible convenience sampling approaches. One possibility would be to identify a probability sample of the desired group and interview them briefly to assess the degree of overlap with potential convenience samples. Unfortunately, obtaining a probability sample of these groups is often so expensive and difficult that if the required efforts were made, it would be better to collect all of the needed information on the spot.

While it might be tempting to generalize from these results to the problem of selecting samples of street prostitutes in other cities, the amount and direction of bias in different convenience samples might well vary from city to city, depending on local patterns of prostitution activity, law enforcement, drug use, and STD treatment, which could be quite different. The problem is that in the absence of a probability sample for comparison, it's difficult to determine which approaches will be useful and which will be inaccurate and what the direction of the error will be.

Household Seroprevalence Survey in Two High-Risk Chicago Neighborhoods: Associations Between Phone in Household and Sexual Risk Behaviors and Crack Cocaine Use

Mary Utne O'Brien, James R. Murray, Afsaneh Rahimian, and W. Wayne Wiebel

Introduction

Risk for HIV in residential samples from two Chicago inner-city neighborhoods was studied. Neighborhoods were selected for study because of their high concentration of out-of-treatment street injection drug users (IDUs). Previous research by the authors on IDUs recruited from street settings in these neighborhoods showed HIV seroprevalence among them to be approximately 30%. The nongay, non-IDU residents of these areas are among those considered to be at highest risk for HIV infection in the general population. Increased concern about the spread of HIV infection to the non-IDU heterosexual population has led to studies of the distribution of high-risk sexual behavior and partnerships in the general population. We sought as well to determine the distribution of those high-risk behaviors and the prevalence of HIV infection in these communities.

In this paper, we describe the major substantive findings from the household survey. We also examine the distribution of key study variables in households with no phone in order to assess the likely effect of data collection method—face-to-face personal interview versus telephone survey—on the observed distribution of responses.

Methods

Setting and Study Population

Two low-income Chicago neighborhoods with a high prevalence of drug abuse, including one with a relatively large gay population, were selected for a larger study of HIV transmission pathways between IDUs and the general heterosexual population. These inner-city neighborhoods are exemplars of the communities described by the National Research Council (Jonsen & Stryker, 1993) as socially and economically deprived, representing the focal points of new HIV infections in the United States. Both neighborhoods are characterized by poverty and unemployment rates twice those observed citywide, a high prevalence of drug and alcohol abuse, and rates of infant mortality, AIDS, hepatitis, syphilis, and tuberculosis that exceed citywide and national rates (Chicago Department of Health, 1994).

The population sampled was 18- through 45-year-old African American, white, and Hispanic residents in 12 Census tracts (1990 Census N = 22,272). The majority of the population in these tracts is African American or Hispanic and under age 35. Full probability methods were used to select to the household level. Equal quotas for gender, race, and age groups for individual selections were imposed using probability sampling with quotas (Sudman, 1966; Stephenson, 1979).

Sample Selection

Each block within designated Census tracts was listed for dwelling units. Specifically included were single resident occupancies and group quarters (e.g., halfway houses). A systematic selection of dwelling units was then made at the central office with a random start and unique selection interval for each tract. The tract-specific selection interval was chosen to yield a number of hits per tract proportional to the 1990 Census counts of eligible adults.

The target sample sizes were set to equality over all 18 cells defined by crossing the three age groups, the sexes, and three race groups (total N = 264). Thus, the selection of dwelling units deviates from a simple random sample only through the use of systematic selection intervals (i.e., there is no clustering). The results of this combined probability sampling with quotas have been shown to be similar to those obtained by full probability sampling with response rates of 75% (Stephenson, 1979).

Interviewer quota sampling instructions were designed to avoid clustering and high rates of substitutions. Within each race group, second interviews in any one cell were not allowed until an interview was completed with persons in each of the other sex and age cells in order to minimize geographical clustering. We limited the number of not-at-homes and refusals per household selection to any

Mary Utne O'Brien is Associate Professor of Epidemiology and Biostatistics at the School of Public Health, University of Illinois at Chicago. James R. Murray, Afsaneh Rahimian, and W. Wayne Wiebel are also in the Department of Epidemiology and Biostatistics at the School of Public Health, University of Illinois at Chicago. This research was supported by grant R01 DA 06589 from the National Institute on Drug Abuse.

combination of six; at that point, interviewers were required to make repeated callbacks to complete an interview, just as in full probability sampling.

During the field period, attempts were made at a total 1,311 dwelling units. Of these, 448 dwelling units were persistent "not at homes" after repeated attempts. There were 26 "known eligible" refusals and 119 "eligibility unknown" refusals. A total of 492 dwelling units were established as out of scope ("vacant/not eligible"). The cooperation rate was 64%.

Data Collection

Two hundred sixty-four face-to-face, structured interviews with finger-stick blood specimens were collected by trained interviewers in Chicago between September 1992 and January 1993. Questionnaires assessed respondents' demographic characteristics, health-related history, recent drug use and sexual behavior, and details of prior HIV antibody testing (site, year, receipt of results, and self-reported sero-status). Respondents were paid $15 for participation. Counseling with voluntary notification of HIV serologic results was provided to participants at the conclusion of each interview. Study protocols and written informed consent forms were approved by the institutional review board at the University of Illinois at Chicago.

Serologic Testing

Serologic screening for HIV antibodies was performed using paper-absorbed finger-stick blood specimens. Specimens repeatedly reactive in whole-virus lysate enzyme-linked immunosorbent assays (ELISA; Genetic Systems, Seattle, Washington) were confirmed by Western blot (Page Blot Systems, Genetic Systems). Specimens were considered seropositive if they had viral specific bands to any two of p24, gp41, or gp120/gp 160. No indeterminate results were observed, nor did we observe negative results in any respondent who claimed to be HIV seropositive.

Statistical Analysis

Results presented are weighted proportions. Statistical estimates are weighted in terms of the 1990 Census cell totals. Weights reflect the sampling population of 22,272 African American, white, and Hispanic adults aged 18 through 45 in the Census cell total.

Results

Study Sample Characteristics

As expected from City of Chicago data on community area characteristics, the study sample was overwhelmingly poor, with almost half (48.7%) reporting annual household incomes under the official poverty level. About one-quarter of the study group received public aid, and only 6 in 10 were employed (compared with 75% citywide). A slight majority of study respondents reported a spouse, and about half had children. The large majority (93% in one community, 80% in the other) described themselves as sexually active in the last 6 months.

Prevalence of HIV Risk Behaviors

High levels of risk behaviors and membership in classic risk groups for HIV infection were reported. In Community 1, 5.3% of males reported they were gay or engaged in male-male sex, and in Community 2, immediately adjacent to Chicago's most densely gay neighborhood, 12% of males reported they were gay or had ever engaged in male-male sex. Almost 6% of respondents in each neighborhood reported current injection drug use or said they were on methadone maintenance. About 10% of respondents reported that crack cocaine was used in their household (this indirect question asked whether the respondent or anyone else in the household smoked crack cocaine in the last 4 weeks). Almost 12% of the study sample reported a history of gonorrhea, and more than 5% had been diagnosed with syphilis. Ten percent of respondents reported activities that met the Centers for Disease Control and Prevention (CDC) criteria for high risk for HIV.

Prevalence of HIV Infection

The overall rate of HIV infection found among residents of the two communities was 4.4%: 3.8% in Community 1 and 5.5% in Community 2 (see Table 1). The higher overall rate in Community 2 was driven by the larger number of gay males there.

The observed rate of HIV infection among gay males (about 50%) is similar to the prevalence rate observed in Chicago's Multi-City AIDS Cohort Study (MACS; gay male cohort), in which approximately 40% of cohort members are infected (MACS Project Staff). The rate of HIV infection among IDUs was 31%, almost exactly the same as that found in the all IDU samples recruited from these same neighborhoods in the authors' earlier studies (Wiebel et al., 1993).

We next examined the rate of infection and risk behaviors among the nongay, non-IDU population—the group that has not been the focus of major public health interventions and testing and counseling initiatives. In this group, we found the rate of HIV infection to be 2.5%. HIV infection was solely among the 35 through 45 year olds, for whom the infection rate was 5.5%.

The most dramatic variations in HIV prevalence among the nongay, non-IDU population were predicted by residents' social location. Even in neighborhoods whose nongay, non-IDU residents were as a group overwhelmingly poor, HIV was concentrated among the least well-off of them (see Table 2). Thus, while the HIV infection rate was

Table 1. HIV Seroprevalence in subgroups of residents in two Chicago neighborhoods

	%	95% CI
Total community HIV seroprevalence (estimated no. seropositives = 790)	4.4	
Community 1 only		
Males	4.1	
Females	3.5	
Total	3.8	
Community 2 only		
Males	7.7	
Females	2.8	
Total	5.5	
Gay males	51.6	20.9–82.0
IDUs	30.6	0.0–68.6
CDC high risk	35.7	10.3–61.1
Nongay, non-IDU	2.5	1.2–3.8

NOTE: For residents 18 through 45, black, white, and Hispanic. Community 1 and Community 2 data are observed, weighted proportions. Remaining data in this table are jackknifed estimates.

Table 2. HIV seroprevalence among socioeconomic subgroups of nongay, non-IDU residents in two Chicago neighborhoods (percentages)

Overall HIV seroprevalence among nongay, non-IDU residents	2.5
In poverty	2.9
Not in poverty	0.0
Receiving public aid	3.3
Not receiving public aid	0.8
Not employed	3.6
Employed	0.0
In household with no phone	1.6
In household with phone	1.4

NOTE: Jackknifed estimates for residents 18 through 45, black, white, and Hispanic.

3% among those living below the poverty level, we found no cases of infection among those above the poverty level. Similarly, the HIV infection rate was almost three times greater among those nongay, non-IDU residents receiving public aid than among those not receiving such assistance. The rate was 3.6% among the unemployed, while no cases of infection were found among the employed.

Sexual Risk Behaviors by Risk Group

It is through its members' sexual behavior that HIV will enter the group not traditionally considered at high risk for HIV infection—nongay, non-IDU community residents. Therefore, we examined the current sexual activities and patterns of association of members of this group to shed light on the possible future of HIV among them. As a point of contrast, we begin by examining the prevalence of several aspects of sexual conduct among the traditional risk groups, gays and IDUs, as well as the nongay, non-IDU community residents. As expected (see Table 3), gay males and IDUs have far higher rates of risky sexual behavior, including more sex partners overall and more new sex partners in the recent past, and three times the likelihood of nongay, non-IDU residents to have sex partners who have sex with others. Gays and IDUs are far more likely to have sex partners who are bisexual and are (apparently accurately) more likely to believe their sex partners have a moderate to high chance of HIV infection. Interestingly, even in the relatively low-risk nongay, non-IDU group, more than a quarter believe this to be true of their sex partner, and that portion varies little by age, race, or gender (data not shown).

The strongest predictor of high-risk sexual behavior among the nongay, non-IDU residents was the use of crack cocaine in the household, reported by 8% of this group (vs. 10% in the total sample). Table 4 illustrates this association. Over 43% of those reporting crack use in the household also reported having three or more sex partners in the last 6 months, versus fewer than 5% of non–crack users. While crack cocaine users were more likely than others to report always using a condom, 90% acknowledged that they fail to use one all the time. Crack users were also almost

Table 3. Sexual risk behavior among residents of two Chicago neighborhoods by risk group

	Gay	IDU	Nongay, non-IDU
More than 4 sex partners in last 6 months	32.7	28.6	4.2
No. sex partners in last 6 months			
0	31.2	13.0	14.1
1	25.2	39.1	63.6
2	9.6	19.3	14.7
3+	34.0	28.7	7.6
No. new sex partners, last 6 months			
0	46.8	51.7	75.3
1	19.2	19.7	14.2
2	1.3	0.0	4.6
3+	32.7	28.7	5.8
Has sex partner who			
Has sex with others	30.0	28.6	10.4
Is bisexual	22.2	16.9	0.3
Is IDU	1.3	18.1	0.1
Has "moderate" to "great" chance of HIV infection	37.2	64.0	26.7
Reports always uses a condom	36.6	11.8	9.4

NOTE: Jackknifed estimates for residents 18 through 45, black, white, and Hispanic

Table 4. Crack cocaine use in the household and own sexual risk behavior among nongay, non-IDU residents of two Chicago neighborhoods (percentages)

	Cocaine use in household	No cocaine use in household
More than 4 sex partners in last 6 months	36.2	1.3
No. sex partners in last 6 months		
0	4.0	14.9
1	32.6	66.3
2	20.1	14.2
3+	43.3	4.5
No. new sex partners, last 6 months		
0	32.2	79.1
1	29.0	13.0
2	5.9	4.5
3+	33.0	3.5
Has sex partner who		
Has sex with others	31.7	8.5
Is bisexual	2.8	0.1
Is IDU	0.0	0.1
Has "moderate" to "great" chance of HIV infection	43.8	25.2
Reports always uses a condom	10.3	0.0

NOTE: Jackknifed estimates for residents 18 through 45, black, white, and Hispanic.

four times more likely than non–crack users to report having sex partners who have sex with others, the avenue through which HIV will be transmitted to this group. Interestingly, no HIV infections were found among those reporting household crack use, indicating that although conditions are ripe for rapid transmission, the virus has not yet been introduced to this risk group.

No-Phone Households

Households with no phone constituted almost one-third of the study sample (30%), higher than the rates reported in the 1990 Census. No-phone households were significantly more likely to represent minority households, mainly African American and Hispanic (see Table 5). Also, men were more likely to report having no phone than were women. Respondents from these households were also significantly more likely to report being on welfare and not having regular employment. In fact, more than one-third of the surveyed households that fell below the poverty line were households with no phone. More than half of those who reported ever being jailed were also among those in households with no phone.

Table 6 shows the experiences with violence and drugs of respondents from no-phone households. Although reports

Table 5. Telephone coverage of sample demographic and economic subgroups

	No phone		p value
	No.	%	
Overall	78	30	
Age			
18–24	23	31	0.968
25–34	29	30	
35–45	26	24	
Race			
White	13	14	0.001
Black	28	39	
Hispanic	37	37	
Gender			
Male	44	33	0.002
Female	34	23	
Partner status			
Spouse	30	28	0.657
No spouse	48	29	
Welfare			
Receiving welfare	30	50	0.002
Not receiving welfare	48	29	
Poverty			
Above poverty	26	21	0.007
Below poverty	50	36	
Employment			
Not employed	41	36	0.056
Employed	37	25	
Jail			
Ever in jail	8	67	0.038

of injection drug use were fairly rare, almost three-quarters of those who reported injection drug use were in households with no phone. In addition, more than half of those who reported someone using crack cocaine were in households with no phone.

Respondents from these households were significantly more likely to report having observed an increase in drug use in their neighborhood and being injured from assault or beating. More than one-third reported being physically assaulted or raped. Had this been a phone survey, we would have missed a major portion of these occurrences.

Table 7 shows the risk behaviors associated with HIV infection among people in households with no phone. In general, these households reported more risky behavior, placing them at higher risk for HIV infection. They were significantly more likely to report having exchanged sex for money ($p \leq 0.05$) and/or drugs ($p \leq 0.03$). They were also significantly more likely to have had multiple sex partners in the 6-month period prior to the interview. These partners were also significantly more likely to have been an IDU and/or to have had sex with others.

The picture that emerges from these analysis is that the households with no phone are more likely to be poor minority households whose members experience more

Table 6. Telephone coverage by drug use, exposure to violence, and exposure to AIDS prevention activity

	No phone		
	No.	%	p value
Nobody in household injected drugs, last 6 months	69	27	
Anyone in household injected drugs, last 6 months	7	73	0.004
Nobody in household used crack in last 6 months	62	24	
Anyone in household used crack in last 6 months	14	61	0.001
Did not observe increase in drug use in neighborhood	25	23	
Observed increase in drug use in neighborhood	47	33	0.019
Never injured from assault or raped	48	25	
Ever injured from assault or beating	27	36	0.040
Never physically assaulted or raped	50	26	
Ever physically assaulted or raped	26	34	0.157
Have not heard of AIDS prevention in community	54	23	
Heard of AIDS prevention in community	22	45	0.028

violence and are engaged in higher risk behavior, both in terms of sexual risk and drug risk, which places them at higher risk for HIV and other sexually transmitted disease infections. In a phone survey, this population would have been missed, and thus, the extent of its high risk behavior as well as its other experiences with violence and drugs would have been significantly underrepresented.

These findings raise the question of the usefulness of telephone surveys focused on HIV risk behavior to predict HIV transmission (e.g., Catania et al., 1992) because they seriously underestimate the numbers of individuals at highest risk, that is, those who engage in high-risk behaviors and do so in communities in which the virus is already present at substantial levels. In the United States, telephone coverage is highest in communities in which the virus is least likely to be present—in effect, the microscope is working best where there are not any microorganisms.

Table 7. Telephone coverage by HIV risk behavior or risk subgroup

	No phone		
	No.	%	p value
Overall	78	30	
Risk groups/behaviors			
Gay male	3	23	0.866
IDU	7	48	0.103
Had sex for money, last 4 weeks	4	73	0.056
Had sex for drugs, last 4 weeks	3	100	0.032
Has had syphilis	6	42	0.109
Has had gonorrhea	11	44	0.459
No. sex partners in last 6 months			
0	7	14	0.043
1	49	28	
2	11	32	
3+	11	46	
Has sex partner who			
Has sex with others	9	47	0.025
Is bisexual	1	12	0.841
Is IDU	2	100	0.027
Has "moderate" to "great" chance of HIV infection	25	33	0.504
Reports always uses a condom	63	35	0.875
HIV status			
HIV positive	4	35	0.613
HIV negative	74	28	

References

Catania, J. A., et al. (1992). Prevalence of AIDS-related risk factors and condom use in the United States. Science, 258, 1101–1106.

Chicago Department of Health. (1994). Community area health inventory—Volume I: Demographic and health profiles. Chicago: Chicago Department of Health.Jonsen, A. R., Stryker, J. (Eds.). (1993). The social impact of AIDS in the United States. Washington, DC: National Academy Press.

MACS Project Staff. Personal communication.

Stephenson, B. C. (1979). Probability sampling with quotas: An experiment. Public Opinion Quarterly, 43, 477–496.

Sudman, S. (1966). Probability sampling with quotas. Journal of the American Statistical Association, 61, 749–771.

Wiebel, W., Jimenez, A., Johnson, W., Ouellet, L., Jovanovic, B., Lampinen, T., O'Brien, M. U., & Murray, J. R. (1993). Positive effect on HIV seroconversion of street outreach intervention with injection drug users in Chicago, 1988–1992. Presented at the IXth International Conference on AIDS, Berlin.

Aggregating Survey Data on Drug Use Across Household, Institutionalized, and Homeless Populations

Robert M. Bray, Sara C. Wheeless, and Larry A. Kroutil

Studies of drug abuse have generally been targeted to specific subpopulations, but few efforts have been made to combine data across these groups into an aggregate population. Since 1971, the National Household Survey on Drug Abuse (NHSDA) series has provided key information about the extent of drug use and drug-related problems among people living in households in the United States (Substance Abuse and Mental Health Services Administration [SAMHSA], 1993). Recently, some estimates from the NHSDA have been criticized because the survey has excluded or has been limited in its ability to adequately represent populations that are potentially at high risk for abusing alcohol or using illicit drugs, such as incarcerated or homeless people. For example, a report to the Senate Judiciary Committee suggested that the 1988 NHSDA underestimated the prevalence of frequent, heavy cocaine use in the United States because the survey did not count drug users who were homeless, in prison, or in treatment (U.S. Senate, 1990). More recently, a report by the U.S. General Accounting Office noted the potential in the NHSDA for noncoverage or undercoverage of groups at increased risk for using drugs (1993).

This paper represents an initial step toward addressing some of these concerns. It describes one effort to combine data obtained from members of households, institutionalized populations, and homeless populations aged 12 and older into an aggregate population in the District of Columbia metropolitan statistical area (DC MSA) and presents findings about drug use prevalence (percentages and numbers of users) for both household and aggregate populations. These results show what effect adding data from nonhousehold populations has on estimates of prevalence of drug use and numbers of users compared with those obtained from the household data alone. This research is part of the Washington, DC, Metropolitan Area Drug Study (DC*MADS), sponsored by the National Institute on Drug Abuse (NIDA).

Robert M. Bray, Senior Research Psychologist, Sara C. Wheeless, and Larry A. Kroutil are with the Research Triangle Institute, Research Triangle Park, North Carolina.

Data Sources and Response Rates

Findings in this paper are based on data from three separate surveys conducted in the DC MSA in 1991. These surveys are the DC MSA oversample of the 1991 NHSDA, the DC*MADS Institutionalized Study, and the DC*MADS Homeless and Transient Population Study.

The NHSDA surveyed the civilian, noninstitutional population, including civilians living on military bases and persons living in noninstitutional group quarters (e.g., rooming houses, dormitories, shelters for homeless people, and group homes). There were 2,547 respondents from a sample of 5,399 households and selected group quarters in the DC MSA (SAMHSA, 1993).

The DC*MADS Institutionalized Study surveyed persons in institutional and noninstitutional group quarters. Institutional group quarters included correctional facilities, mental or psychiatric hospitals, and other institutions, such as noncorrectional facilities for juveniles. Noninstitutional group quarters included group homes for people who are mentally retarded, homes for people with physical disabilities, and transitional homes for people leaving treatment for alcohol or other drug abuse. Nursing homes and hospitals or wards providing treatment for alcohol or other drug abuse were excluded. There were 1,203 interviews with residents of 42 institutions stratified into four groups: 868 interviews from 20 correctional institutions; 207 interviews from 6 psychiatric institutions; 55 interviews from 7 noncorrectional institutions for juveniles; and 73 interviews from 9 group homes (NIDA, 1994).

The DC*MADS Homeless and Transient Population Study surveyed persons who were either literally homeless or at imminent risk of becoming homeless, including persons who spent the previous night in an emergency shelter or in a nondomicile (i.e., in a vacant building, public or commercial facility, city park, or car or on the street) or who were using soup kitchens or emergency food banks for the homeless population. There were 908 interviews from four overlapping sampling frames: 477 interviews with residents in 93 shelters; 224 interviews with patrons of 31 soup kitchens and food banks; 143 interviews with literally homeless people from 18 major clusters of encampments; and 64 interviews with literally homeless people from an area probability sample of 432 Census blocks in the MSA (NIDA, 1993).

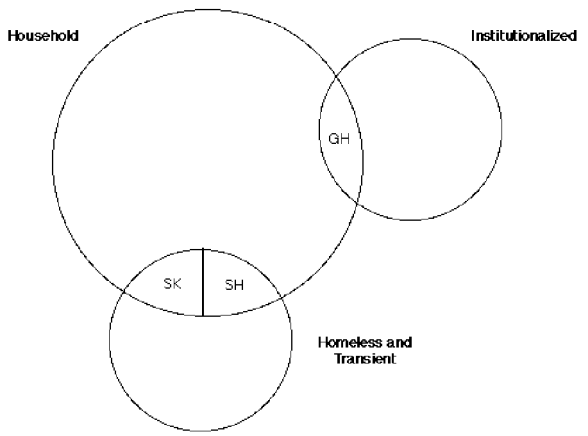
For the household, institutionalized, and homeless studies, respectively, the household/institutional response rates were 93.5%, 87.5%, and 82.6%; the individual interview response rates were 82.1%, 89.4%, and 86.1%; and the overall rates were 76.8%, 78.2%, and 75.0%. Data were combined from the household, institutionalized, and homeless populations to produce an aggregate population for the DC MSA based on interviews from 4,658 individuals. Aggregate data were adjusted for potential sampling overlap across the surveys.

Combining Data Sets for Aggregate Population Estimates

Steps were taken during the planning of these three studies to permit the integration of the data. These included coordination of timing of data collection, definition of the subpopulations, similar structure and content of questionnaires, and similar estimation procedures. Although the populations surveyed by the three studies were generally defined in terms of place of residence, there was a small overlap in the target populations for the three studies. Both the NHSDA and the Institutionalized Study included some portions of the noninstitutionalized group quarters population. Both the NHSDA and the Homeless and Transient Population Study included persons living in homeless shelters and persons who, while not literally homeless, may have been at risk of homelessness, as evidenced by their use of soup kitchens or food banks.

Figure 1 shows graphically the potential overlap in the target population for the three surveys. Of the 4,658 persons interviewed, 637 could potentially have participated in more than one of the studies. In terms of the total number of persons represented, however, the overlap was very small; less than 0.5% of the total combined population was potentially represented by more than one of the surveys. Nevertheless, to address these potential overlaps, it was necessary to make adjustments to avoid multiple counting of the subpopulations when producing aggregate estimates. Respondents were first classified according to the number of overlapping surveys for which they could have been selected. At most, the overlap occurred only in two of the three surveys (i.e., household and homeless, household and institutionalized, or homeless and institutionalized). It was not known whether persons interviewed for the NHSDA may have been at risk of homelessness, as evidenced by their use of soup kitchens. Although it was not possible to completely adjust for this potential multiplicity, it was assumed that only a small proportion of persons who were linked to the area frame used for the NHSDA were also linked to the soup kitchen frame.

Figure 1. Overlap of 1991 DC MSA household, institutionalized, and homeless sampling frames



GH = Noninstitutionalized Group Homes
SK = Not Literally Homeless Who Used Soup Kitchens
SH = Homeless Shelters

NOTE: Populations not drawn to scale. Household data source: 1991 NHSDA: DC MSA (SAMHSA, 1993). Homeless data source: 1991 DC*MADS Homeless and Transient Population Study (NIDA, 1993). Institutionalized data source: 1991 DC*MADS Institutionalized Study (NIDA, 1994).

Analysis weights were adjusted for persons who could have been selected for two surveys by multiplying them by one half. These adjusted weights were then summed with the final analysis weights for other individuals in the three surveys to form multiplicity-adjusted weights for the aggregate population.

Even though the use of multiplicity-adjusted weights in the overlapping portions of the surveys reduces the bias (ideally to 0), it is plausible that this reduction is more than compensated for by an increase in sampling variance. To assess the trade-offs involved in the use of multiplicity estimates, the variances of key estimates were examined for different options that varied in how the data in the overlapping portions of the target populations were treated:

- Option 1. Disregard the NHSDA portion of the overlap with the other two surveys (and assume that the number of interviews with users of soup kitchens in the NHSDA is negligible).
- Option 2. Use the multiplicity-adjusted overlap for all overlapping portions.
- Option 3. Disregard the interviews with soup kitchen users who were not literally homeless from the homeless survey portion of the overlap and use multiplicity-adjusted weights for shelter and noninstitutionalized group home interviews.
- Option 4. Disregard the interviews with soup kitchen users who were not literally homeless from the homeless survey portion of the overlap and disregard the NHSDA portion of the overlap with the other two surveys.

For each option, the estimated total number of persons in the union of the three populations and prevalence estimates related to past year and past month use of any illicit drug, crack cocaine, heroin, and alcohol were computed along with the estimated number of users, standard errors, and

relative standard errors of all estimates. Estimates of prevalence and numbers of users were similar for the four options. Consequently, the multiplicity-adjusted weights (Option 2) that retained all of the data were selected for use in producing estimates for the aggregate data. The aggregate data set provides unbiased estimates of the prevalence of illicit drug, alcohol, and cigarette use among the eligible population in the DC MSA. The SUDAAN software package (Research Triangle Institute, 1990) was used to compute prevalence estimates and associated standard errors.

Household and Aggregate Population Characteristics

Table 1 shows the percentages of the DC MSA aggregate population and estimated numbers of people from the household, institutionalized, and homeless subpopulations. Estimates were made after the data were adjusted for potential multiplicity in the sampling frames. As shown, the household population made up the vast majority (99.2%) of the aggregate population. However, an estimated 52% of the total institutionalized and group quarters population were not included in the Institutionalized Study, of whom the majority (46%) were nursing home residents. Had residents of these other institutions been sampled, the percentage of the aggregate population living in institutions would have been greater than the 0.6% that is shown in Table 1.

Because the majority of the aggregate population came from the household population, the demographic compositions of the two populations were virtually the same. For the household and aggregate populations, respectively, 47.7% and 48.1% were male, 61.9% and 61.6% were white, 27.2% and 27.5% were black, 33.5% and 33.8% were single, 12.9% and 13.2% of adults had less than a high school education, and 65.4% and 65.1% of adults were employed full-time.

Table 1. Estimates of people in the DC MSA aggregate population by subpopulation: 1991

Subpopulation	%	No.
Aggregate	100.0	3,198,698
Household	99.2	3,171,915
Institutionalized	0.6	19,395
Homeless	0.2	7,388

NOTE: Estimates have been adjusted for potential multiplicity in the sampling frames.The institutionalized subpopulation does not include nursing homes, residential schools for the hearing impaired, homes for the developmentally disabled, or religious group quarters. Household data source: 1991 NHSDA: DC MSA (SAMHSA, 1993). Home-less data source: 1991 DC*MADS Homeless and Transient Population Study (NIDA, 1993). Institutionalized data source: 1991 DC*MADS Institutionalized Study (NIDA, 1994).

Prevalence of Illicit Drug, Alcohol, and Cigarette Use During the Past Year

Findings from the DC*MADS Institutionalized Study and the DC*MADS Homeless and Transient Population Study indicate high rates of illicit drug, alcohol, and cigarette use in the past year in these two nonhousehold populations in 1991. Nearly half (49.9%) of the institutionalized population used an illicit drug in the past year, 36.9% used cocaine in any form, 31.7% used marijuana, and 30.4% used crack cocaine (NIDA, 1994). Among the homeless and transient population, 57.7% used an illicit drug in the past year, 48.4% used cocaine, and 44.8% used crack cocaine (NIDA, 1993). However, direct comparison of these rates with those in the household population may be misleading because of demographic differences between the household and nonhousehold populations that have also been shown to be related to substance use. For example, the two nonhousehold populations were predominantly male (90.7% for the institutionalized population covered by DC*MADS and 75.7% for the homeless and transient population) while less than half of the household population was male, and males generally have higher rates of substance use than do females. Therefore, the focus of the discussion of prevalence estimates is on comparison of rates for the DC MSA household and aggregate populations, where the demographic compositions were virtually identical.

In general, adding data from nonhousehold populations in the DC MSA raised estimates of the percentages of users only slightly, despite high rates of drug use in these populations. However, some of these increases represent an additional 10,000 or more users in the aggregate population compared with the estimated numbers based on household data alone.

Table 2 shows estimates of the prevalence of use and estimated numbers of users of illicit drugs, alcohol, and cigarettes in the past year in the DC MSA household and aggregate populations. Specific highlights include the following:

1. Adding data from nonhousehold populations to the household data raised the prevalence estimate of any illicit drug use slightly, from 11.7% in the household population to 12.0% in the aggregate population. However, this 0.3% increase represents an additional 14,252 past year illicit drug users.
2. The prevalence of any illicit drug use except marijuana in the past year was 8.1% in the aggregate population compared with 7.8% in the household population. This represents an estimated 12,179 more users in the aggregate population.
3. Adding data from nonhousehold populations raised the prevalence estimate of any form of cocaine use to 3.9% in the aggregate population compared with 3.6% in the household population. This represents an

Table 2. Illicit drug, alcohol, and cigarette use in the past year in the DC MSA household and aggregate populations: 1991

	Household		Aggregate	
	%	No. users	%	No. users
Any illicit drug use[a]	11.7	370,486	12.0	384,738
Marijuana/hashish	8.1	256,514	8.3	265,591
Cocaine	3.6	114,538	3.9	125,508
Crack	0.9	29,027	1.2	38,433
Inhalants	1.7	52,514	1.7	53,350
Hallucinogens	1.5	48,417	1.6	51,091
Heroin	0.4	12,314	0.5	15,549
Nonmedical use of any psychotherapeutics[b]	4.5	142,118	4.5	144,696
Any illicit drug use, excluding marijuana[c]	7.8	248,132	8.1	260,311
Alcohol use	73.5	2,332,210	73.4	2,349,174
Cigarette use	28.1	891,575	28.5	912,063

NOTE: Aggregate population includes the combined household, homeless, and institutionalized populations. Estimated numbers of users have been adjusted for potential multiplicity. Household data source: 1991 NHSDA: DC MSA (SAMHSA, 1993). Homeless data source: 1991 DC*MADS Homeless and Transient Population Study (NIDA, 1993). Institutionalized data source: 1991 DC*MADS Institutionalized Study (NIDA, 1994).

[a]Use of marijuana or hashish, cocaine (including crack), inhalants, hallucinogens (including PCP), heroin, or nonmedical use of psychotherapeutics at least once.

[b]Nonmedical use of any prescription type stimulant, sedative, tranquilizer, or analgesic; does not include over-the-counter drugs.

[c]Use of cocaine (including crack), inhalants, hallucinogens (including PCP), heroin, or nonmedical use of psychotherapeutics at least once.

estimated 10,970 more past year cocaine users in the aggregate population relative to the household population.

4. Adding data from nonhousehold populations increased the prevalence of crack cocaine use in the past year from 0.9% for the household population to 1.2% for the aggregate population. This represents an additional 9,406 crack users compared with the estimate obtained from household data alone.

5. The aggregate population data yielded an estimate of 3,235 more past year heroin users than in the household population.

6. The prevalence of any alcohol use in the past year was virtually the same in the household as in the aggregate population. Adding data from nonhousehold populations raised the prevalence of any cigarette use in the past year by 0.4%, from 28.1% in the household population to 28.5% in the aggregate population.

Another way of viewing data on the numbers of past year users of illicit drugs is in terms of the percentage of users in the aggregate population who would be accounted for by the household data (i.e., estimated numbers of users based on household data divided by estimated numbers based on aggregate data). These results are based on data in Table 2, but the calculations are described below. Highlights include the following:

1. Of the estimated 384,738 past year illicit drug users in the aggregate population, over 95% would have been accounted for by the household estimate (i.e., 370,486 ÷ 384,738 × 100 = 96.3%).

2. Of the estimated 260,311 users of any illicit drug except marijuana in the aggregate population in the past year, approximately 95% would have been accounted for by the household estimate (i.e., 248,132 ÷ 260,311 × 100 = 95.3%).

3. Household data would have accounted for approximately 90% of the estimated 125,508 past year users of any form of cocaine in the aggregate population (i.e., 114,538 ÷ 125,508 × 100 = 91.3%).

4. Household data would have accounted for approximately three-fourths of the past year crack users in the aggregate population (i.e., 29,027 ÷ 38,443 × 100 = 75.5%).

5. Of the estimated 15,549 past year heroin users in the DC MSA aggregate population, household data would have accounted for slightly less than four-fifths (i.e., 12,314 ÷ 15,549 × 100 = 79.2%).

Statistical tests showed that nearly all of the estimated numbers of past year users shown in Table 2 for the aggregate population were significantly greater than the corresponding numbers for the household estimates ($p < .05$). In particular, estimates of the numbers of users of any illicit drugs, any illicit drugs except marijuana, any form of cocaine, and crack cocaine were significantly greater for the aggregate population compared with the household population. The number of past year heroin users in the aggregate

population was the only estimate in Table 2 that was not significantly different from the household estimate.

Despite attempts to coordinate across studies, some differences occurred that might be viewed as potential limitations of the data. One difference was in the mode of questionnaire administration. For the NHSDA, most of the questionnaire was self-administered, whereas for the two DC*MADS studies, the instruments were interviewer administered because many of the institutionalized and homeless respondents may have had limited reading ability. Although steps were taken to reassure DC*MADS respondents of the confidentiality of their data, some may have been less likely to report drug use in the interviewer-administered questionnaire. Albeit this difference in data collection procedures introduces possible method variance, use of self-administered questionnaires in DC*MADS would likely have resulted in data of poor quality due to respondent difficulty in completing the questionnaires.

Another difference concerns some variation in the timing of data collection. These three studies were initially designed so that data collection would be conducted in the first half of 1991 (January through June 1991). Data collection for the Institutionalized Study, however, actually extended from April to December 1991. In combining data from the three studies, the assumption was made that drug use would be fairly stable in these populations over the time period. However, if drug use showed seasonal variations over the year for the institutionalized population, then the estimates obtained from combining the data could vary from the estimates that would be obtained had all three studies been conducted in the first half of 1991. Nevertheless, the actual effect of this variation in the data collection periods is likely to be small because the institutionalized population is only a small fraction of the total population.

Finally, because of some undercoverage, aggregate estimates do not reflect the entire DC MSA population. Groups excluded from the studies include those living in nursing homes and those in the military. However, these omissions are viewed as minor because these groups represent either a small fraction of the total population or are not likely to include drug users. Some of these groups are covered by other population surveys, such as the Worldwide Surveys of Substance Abuse and Health Behaviors Among Military Personnel (Bray et al., 1992).

Discussion and Conclusions

Combining data from household and nonhousehold populations resulted in prevalence estimates (i.e., percentages) that were only slightly higher than the estimates that were obtained from household data alone, even though these nonhousehold populations had relatively high rates of drug use. Adding data from nonhousehold populations also had relatively little effect on prevalence estimates within demographic subgroups for any illicit drug use, any alcohol use, heavy alcohol use, and cigarette use.

Greater variation was evident, however, when examining numbers of users. Specifically, the aggregate population data yielded past year estimates of approximately 14,000 more illicit drug users, 9,000 more marijuana users, 11,000 more users of any form of cocaine, 9,000 more crack users, and 3,000 more heroin users compared with the corresponding estimates for the household population. In addition, comparisons of numbers of past year crack users and heroin users suggest that estimates based on household data alone would fail to capture about 25% of the past year crack cocaine users and 20% of the past year heroin users in the aggregate population. These findings suggest that data from the household population alone may yield somewhat conservative estimates of the numbers of past year users of illicit drugs, including marijuana, cocaine in any form, crack cocaine, and heroin in the DC MSA.

The higher estimates of the numbers of past year crack and heroin users based on aggregate data may be of particular interest because of the low rates of use of these substances in the household population. Further, data from nonhousehold populations may be important to researchers and policy makers concerned with total counts rather than percentages, such as treatment providers trying to estimate the need or demand for services.

Findings from this study also underscore a potential limitation in reporting overall "macrolevel" estimates for a large population. Such general estimates can obscure high rates of drug use or related problems among subgroups that constitute only a small percentage of the overall population. If these problems go undetected, policy makers and service providers may fail to develop appropriate strategies to address them. These findings, of course, are specific to the DC MSA and cannot immediately be generalized to other metropolitan areas or to national estimates of drug use. However, replication of this type of study in other major metropolitan areas would help to establish whether the observed underrepresentation of crack cocaine and heroin users in the household population is unique to the DC MSA or is a more general phenomenon.

Assuming that similar findings were observed in other metropolitan areas, decisions about the importance of including nonhousehold populations depend upon the aims of the study. If the study's aim is strictly to estimate percentages of the population who have used drugs, then coverage of the household population may be sufficient. However, if the aims include estimating the size of the drug-abusing population or conducting more detailed examination of relatively rare behaviors (e.g., crack cocaine, heroin, and needle use), important subgroups of drug abusers might be missed without the inclusion of nonhousehold populations. This study demonstrates the feasibility of combining data from surveys of drug use among household and nonhousehold populations in a metropolitan area to produce prevalence estimates that cover a broader population.

References

Bray, R. M., Kroutil, L. A., Luckey, J. W., Wheeless, S. C., Iannacchione, V. G., Anderson, D. W., Marsden, M. E., & Dunteman, G. H. (1992). 1992 Worldwide Survey of Substance Abuse and Health Behaviors Among Military Personnel. Research Triangle Park, NC: Research Triangle Institute.

National Institute on Drug Abuse. (1993). The Washington, DC, Metropolitan Area Drug Study: Prevalence of drug use in the Washington, DC, metropolitan area homeless and transient population: 1991 (Technical Report No. 2). Rockville, MD: National Institute on Drug Abuse.

National Institute on Drug Abuse. (1994). The Washington, DC, Metropolitan Area Drug Study: Prevalence of drug use in the Washington, DC, metropolitan area institutionalized population: 1991 (Technical Report No. 4). Rockville, MD: National Institute on Drug Abuse.

Research Triangle Institute. (1990). Software for Survey Data Analysis (SUDAAN), Version 5.30 [Computer software]. Research Triangle Park, NC: Research Triangle Institute.

Substance Abuse and Mental Health Services Administration. (1993). National Household Survey on Drug Abuse: Main findings 1991 (DHHS Publication No. SMA 93-1980). Rockville, MD: Substance Abuse and Mental Health Services Administration.

U.S. General Accounting Office. (1993). Drug use measurement: Strengths, limitations, and recommendations for improvement (Report No. GAO/PEMD-93-18). Washington, DC: U.S. General Accounting Office.

U.S. Senate. (1990). Hard-core cocaine addicts: Measuring and fighting the epidemic (Senate Report No. 6). Washington, DC: U.S. Government Printing Office.

Sampling Medicaid and Uninsured Populations

John W. Hall

Introduction

This paper discusses strategies for sampling Medicaid beneficiaries and those without health insurance. To illustrate these strategies, the paper presents the design of 10 similar samples for surveys conducted in 1993.[1] Each sample covered the household population of a state and required oversampling of the Medicaid and uninsured populations.[2] Each sample used list-assisted random-digit dialing (RDD) to reach the population living in households with telephones and area probability sampling to reach those in households without telephones. In 9 of the 10 states, lists of Medicaid recipients were employed to facilitate oversampling those on Medicaid.

The main objectives of each state's sample design were two: to provide an adequate sample for three domains of interest (insured, uninsured, and Medicaid) and to minimize bias due to nonresponse and noncoverage without substantial and expensive in-person data collection. An additional objective was to provide statewide estimates of insurance coverage and other health-related measures. The study design called for roughly equal effective samples in each state of uninsured and insured and a somewhat smaller Medicaid sample. To enable statewide estimates, households with only Medicare recipients were included but sampled at half the rate of other insured households.

In all but one state, we were able to obtain lists of Medicaid recipients to more effectively target that group. We still had to rely on general population screening to identify those without health insurance. To minimize screening costs, we oversampled areas more likely to contain the uninsured (or Medicaid recipients).

In addition to minimizing screening costs, the sample design had to address the bias associated with noncoverage. We estimated that 22% of the Medicaid and 13% of the uninsured groups did not live in households with telephones (1990–1991 Current Population Survey [CPS] estimates), so omitting or underrepresenting nontelephone families could lead to biased survey estimates. We thus used in-person interviews to obtain data on a small sample of nontelephone households in each state.

In the remainder of this paper, we discuss these topics: First, we discuss the study population; the following section covers the sampling frames and rationale for stratification; the next two sections present overviews of the design for the telephone sample and the design of the in-person sample; the final section contains some concluding remarks and observations.

Study Population

The study population consisted of all persons living in households in selected states. Study objectives required (separate) minimum samples of (a) those covered by Medicaid or equivalent state programs; (b) those insured through private or employment-based insurance, Medicare, the Indian Health Service, other state programs, the Civilian Health and Medical Program for the Uniformed Services (CHAMPUS), or Veteran's Affairs; and (c) the uninsured. Individuals on Supplemental Security Income (SSI) or Medicare are included in the survey for the limited purpose of making generalizations to the entire household populations of each state. The sample targets of completed interviews were stated in terms of subunits of households that we called families. Within each sampled household, we collected data about each family.

The insurance family includes the head, spouse, and their dependent children up to age 18 or to age 23 if they are in school. This latter definition represents conventional practice in the private insurance market and is similar to the filing unit used by Medicaid and state-subsidized insurance programs. The Census family (U.S. Bureau of the Census, 1992) sometimes comprises more people than the insurance family. Examples of people typically included in the same Census unit but in different insurance units are adult children and their families living in the homes of their parents, adult siblings living together, and parents living in the homes of their adult children.

John W. Hall, Senior Sampling Statistician, is at Mathematica Policy Research, Inc., in Princeton, New Jersey.

[1]The surveys were conducted by Mathematica Policy Research, Inc., for the Robert Wood Johnson Foundation.

[2]By "uninsured" we mean those that have no government-supplied or private health insurance coverage. "Medicaid" also includes equivalent state-sponsored programs.

Sampling Frames and Rationale for Stratification

As discussed, the sample design in each state used one in-person and two telephone frames. In states where Medicaid lists were available, that list comprised a fourth frame.[3]

In discussing sample stratification, we distinguish strata from what Kish (1965) calls domains. A stratum is a subgroup of the study population that can be identified in the sampling frame(s) and thus sampled separately; a domain is a subgroup that will be examined during the analysis. The study focused on three domains—insured, uninsured, and Medicaid recipients. Of the three domains, only the Medicaid group (because list frames were available) could have been treated as a sampling stratum. We used stratification in our sample design to achieve the desired distribution of the sample over the three domains. By selecting strata that are correlated with one or more of these domains, we sampled more efficiently than if we merely selected a random sample of households. The specific plans for stratifying the telephone and in-person samples are discussed in greater detail below.

Telephone Survey Sample

This section discusses sampling frames, stratification, allocation of the sample among strata, and screening rates for the telephone sample.

Telephone Frames and Stratification

For telephone interviewing, we employed three frames: lists of Medicaid recipients provided by the states and two frames for general population screening. One general population frame comprised telephone numbers published in telephone directories, while the other included all potential household numbers not appearing in directories. Below, we refer to these as the published and unpublished frames.[4] We believed that the procedure of dividing the general population frame into two component frames would be more efficient than using a single frame because we could use Census data to more accurately stratify the published frame. We expected the published frame would be more efficient than the unpublished for oversampling the uninsured or Medicaid beneficiaries.

We employed the Medicaid frame because where available, it would be our richest source of Medicaid households. The published telephone household frame would make it possible to target families with specific characteristics (such as low income) in smaller geographic areas and eliminate most nonresidential telephone numbers. The unpublished frame was used to locate households with unpublished numbers using what is essentially a one-stage RDD sample from which listed business numbers have been eliminated, drawn from a frame that includes working banks.[5]

We first stratified the published frame. To do this, we stratified Census block groups using 1990 Census data. Each published telephone number that could be accurately identified with a block group[6] was assigned to that block group's stratum. Telephone numbers that could not be identified with a block group were treated as unlisted. Our goal for published strata was to stratify in terms of expected yield of the uninsured in all states but the one where a separate Medicaid frame was unavailable; in that state, we targeted the Medicaid group. The stratifying variables we used to target the uninsured were estimates of income and proportion receiving Social Security. Income correlates with insurance status, and Social Security is an indicator of Medicare, which was undersampled. We used receipt of public assistance to help us target Medicaid households in the state that did not provide a Medicaid list. For each state, we formed three, four, or five such strata based on predicted yield of uninsured or Medicaid households.

For the unpublished frame, we created strata parallel to the published strata by examining the distribution of telephone area code/exchange[7] combinations in the published frame. Then, we assigned each exchange to a stratum based on the published stratum that had the plurality of published numbers for that exchange. For example, if the plurality of published numbers in area code/exchange (609) 799 were in the published stratum having the lowest expected yield of uninsured, then unpublished numbers in (609) 799 were assigned to the lowest yield unpublished stratum. However, if the plurality was less than 40% of an exchange's published numbers, we assigned the exchange to an unpublished stratum judgmentally, based on the overall distribution of published numbers across strata.

Selecting the Samples

The samples from published and unpublished frames were selected by our sample vendor using random selection within substrata. A substratum was defined as the width of

[3] In all states where a Medicaid list was available, it was used as a frame for the telephone sample. In three states, we also used the Medicaid lists for the in-person sample.

[4] The term "frame" is used to emphasize the different procedures we used in sampling from these two lists. An alternative way to conceptualize the process would be this: We used one general population telephone frame, and dividing this frame into published and unpublished components was the first level of stratification.

[5] Working banks are defined as banks of 100 telephone numbers (consecutive numbers with the last two digits ranging from 00 to 99) with at least one listed household number, all sharing the identical area code and exchange.

[6] Some telephone numbers are listed in directories with a name and number but no address. In other cases, the vendor cannot make a reliable match between address and block group. Between them, these categories comprise 10% to 20% of a state's listed telephone numbers.

[7] In a 10-digit phone number (XXX) YYY-ZZZZ, (XXX) defines the area code and YYY the exchange.

a stratum's sampling interval (the ratio of the population of telephone numbers in the stratum to the sample size). In selecting the unpublished sample for this study, the vendor purged from the frame all numbers having a chance of selection into the published frame.

The Medicaid samples were selected by the states, using instructions we provided. The methods of selection differed in two ways: according to whether selection was random or systematic after a random start and according to whether the unit selected was a family or case or an individual. Six states provided us with cases (roughly equivalent to our families) and three gave us records of individuals. We used case ID numbers, addresses, and phone numbers to merge case or individual records into household sampling units.

None of the states provided us with telephone numbers for all Medicaid cases. In each state, we used directory assistance lookups to obtain telephone numbers for as many cases as we could before releasing the sample for interviewing.[8]

Sample Allocation

Sample allocation was done iteratively in each state. Based on our initial allocation, we first released a portion of the sample needed to complete the study in a state. As the survey progressed, our initial estimates were revised and sample allocations adjusted accordingly. Subsequent releases of sample were based on the revised allocation.

The initial allocation was based on the estimated cost of an interview, which in turn incorporated estimates for each stratum of the prevalence of the rarest domain (uninsured or Medicaid) and assumptions about the prevalence of working household telephone numbers in the published, unpublished, and Medicaid frames.

As the interviewing progressed, we observed that the differences between the published and unpublished frames were not quite what we had expected. While the screening rate for households was better for the published strata than for the unpublished, the prevalence of the uninsured and Medicaid domains was often higher for the unpublished strata. Our iterative strategy allowed us to adjust our sample allocation by stratum to take advantage of these findings.

Telephone Screening

The screening process first identified whether a telephone number reached a household. In the case of a number called from the Medicaid frame, we determined if we had reached the unit sampled. Once we determined that we had reached a household or the Medicaid unit sampled, questions were asked to classify the household as a "Medicare/SSI," "insured," "uninsured," or "Medicaid" household.[9] The household was retained for interviewing depending on the household's classification and the need for additional interviews in the category to which it belonged.[10]

In-Person Sample Design

As discussed above, the purpose of the in-person sample is to ensure that members of the three domains who do not have telephones are represented in the survey data. The insured population would be well covered by a telephone sample; however, telephone coverage is inadequate for the uninsured and Medicaid groups. In a few states, CPS estimates of telephone coverage even for these subgroups are quite high. However, because 1990–92 CPS within-state estimates are based on very few cases, it would not have been prudent to assume adequate coverage of these groups from a telephone survey. Further, our analysis shows that uninsured families with no insurance coverage at all are much more likely than families in which some members have coverage to live in households without telephones.

In-Person Frame and Stratification

We used area probability frames to conduct in-person interviews with selected nontelephone households. Using 1990 Census data (File STF3), we defined primary sampling units (PSUs) as counties or groups of counties. In general, we grouped the PSUs into strata of roughly equal size, with one stratum being defined for each PSU allocated (see "Sample Allocation" section below) for in-person interviewing. Stratifying variables included geography and degree of urbanization. For subsampling in large PSUs, we used within-PSU stratification.

Sample Allocation

Within each state, the number of cases allocated to the in-person sample depended on telephone coverage for

[8] All states had automated, centralized files. Some states had systems with complete and current information and were able to provide telephone numbers for 50% to 60% of the cases they sampled. A few states did not have telephone numbers at all or only for a small percentage of cases. In these states, telephone numbers were collected from clients but retained only at the county office level.

[9] In "Medicare/SSI" households, all adult members were covered by Medicare or SSI; in "insured" non-Medicare households, all members were covered by private health insurance or Medicare, none were covered by Medicaid, and at least one member was not a Medicare beneficiary; in "uninsured" households, at least one member was not covered by Medicaid, Medicare, or private insurance; in "Medicaid" households, no members were uninsured and at least one member was covered by Medicaid and was not also covered by Medicare or SSI. In the state that did not provide a Medicaid list sampling frame, the definitions of "Medicaid" and "uninsured" were changed. In that state, "Medicaid" households could include uninsured persons, while "uninsured" households could not include persons on Medicaid.

[10] The selection at this stage was controlled by the computer-assisted telephone interviewing (CATI) program. In most instances, at a given point in time, all households in a category were either accepted or rejected. However, some households were randomly subsampled at this point—Medicare/SSI households were always subsampled at half the rate of insured households.

Medicaid recipients and for the uninsured, since we wished to minimize the design effects for these groups. To allocate the sample, we first grouped the 10 states by the proportion of Medicaid and uninsured households estimated to be without telephones (based on 1991 and 1992 CPS). Six of the 10 states were defined as high-coverage states, 2 as medium coverage, and 3 as low-coverage states. We allocated 50 completed family interviews per PSU. The number of PSUs assigned to a state depend on the number of in-person interviews needed in a state and the expected cost of adding a PSU. In high–telephone coverage states, where the number of in-person interviews would be smallest, we selected two PSUs, while in medium coverage states, we selected four. In low–telephone coverage states, we originally planned to select five PSUs, but the very high cost of in-person interviewing and of adding a PSU in these states led us to reduce the allocation to four PSUs. Thus, the in-person allocation for high-coverage states was 100 family interviews in two PSUs and for other states, 200 family interviews in four PSUs.

Selection of PSUs

For each state, we selected one PSU per stratum. Within each PSU, we selected secondary selection units (SSUs) and then listing areas. Our goal was to have listing areas in a PSU close enough together so that one interviewer could cover them, but distant enough to ensure heterogeneity in the characteristics of their residents. To keep interviewing costs reasonable, we excluded Census block groups with telephone coverage of 95% or more and those with vacancy rates of over 50%. In the high–telephone coverage areas, screening costs would be extremely high, and we would find only a small proportion of the domain of primary interest (uninsured) without phones. Areas with high vacancy rates contain a high proportion of vacation homes.

Selection of PSUs was made with probability proportional to size (PPS),[11] where size is the estimated number of nontelephone families. In two states, we made one certainty selection[12] and the remainder of the selections were made with PPS. Within each PSU, we selected one or more (usually two) SSUs with PPS. Within SSUs, listing areas were selected in one of two ways: In the first three states in which we conducted the survey, selection of listing areas was with PPS; in the remaining states, we formed the replicate subsamples of blocks selected two to six at random.[13] The reason for the change was that we found that PPS selection of listing areas had not worked as well as hoped in the first three states—we had to screen many more households than anticipated.[14]

Listing

As with the selection of listing areas, we used one set of procedures for listing and subsampling households within listing areas for the first three states and another set of procedures for the remainder. In the first three states, we listed all units in a listing area, recording full address and name of occupant, where available. We then compared the addresses found by the lister with a list provided by Donnelley of addresses having published telephone numbers. We also matched our lists with lists provided by the state Medicaid offices. Addresses that could be linked to the Donnelley list and identified as having telephones[15] were not sampled for interviewing. Addresses that matched the Medicaid lists were oversampled, unless it was clear that the Medicaid recipient had a telephone.

We found that the procedure used for the first three states was inefficient. The procedure did not eliminate a large proportion of telephone households, and because we had to screen such a large portion of all households listed, matching our lists with Medicaid lists improved our effectiveness in identifying members of the Medicaid

[11] In PPS selection, each PSU is assigned a measure of size (MOS_{ah}). The probability of selection for a PSU within a stratum is

$$p(a_h) = \frac{a\ MOS_{ah}}{\sum_h MOS_{ah}}$$

where a is the number of PSUs selected.

[12] If a PSU in a stratum would have had a probability of selection greater than 0.80, we selected that PSU with certainty and selected the remainder with PPS. Because there was only one selection per stratum, the remainder of the stratum having a certainty selection was combined with another stratum. In New York, the certainty selection included New York City and surrounding counties. Because this area includes 75% of the nontelephone population, we selected two other PSUs, with one SSU (rather than two) assigned to the two upstate PSUs.

[13] The number selected depended on the expected yield of nonphone households and the number of selections that would minimize the impact on weighting, given prior probabilities and our listing procedures.

[14] Using the notation introduced in footnote 10, if MOS_{ah}, MOS_{bah}, and MOS_{cbah} are the measures of size for a PSU, SSU, and listing area, and a, b, c, and d are the number of PSUs, SSUs, listing areas, and households to be sampled in each stratum, then the probability of a household being selected for screening—P(HH)—is

$$P(HH) = \frac{a\ MOS_{ah}}{\sum_h MOS_{ah}} \cdot \frac{b\ MOS_{bah}}{MOS_{ah}} \cdot \frac{c\ MOS_{cbah}}{MOS_{bah}} \cdot \frac{d}{MOS_{cbah}} = \frac{a \cdot b \cdot c \cdot d}{\sum_h MOS_{ah}},$$

which is equal for all households within a stratum (a desirable result because sample weights will vary less).

However, this result holds true only if we control the final stage of selection. For example, to maintain equal probabilities, we must vary d if the actual number of units in the listing area differs from MOS_{cbah}. If we cannot control the rate of selection in the listing area, unequal probabilities of selection will result. In the first three states, low yields forced us to take all or nearly all units in many listing areas, even if taking fewer was desirable. This resulted in greater inequality of selection probabilities and hence more variable weights.

In the remaining states, our selection was

$$P(HH) = \frac{a\ MOS_{ah}}{\sum_h MOS_{ah}} \cdot \frac{b\ MOS_{bah}}{MOS_{ah}} \cdot \frac{c'}{10} \cdot 1 = \frac{ab\ (MOS_{bah})}{\sum_h MOS_{ah}} \cdot \frac{c'}{10}$$

where c' is the number of replicate subsamples of blocks selected to form the listing area. As discussed below, in states where we used this design, we intentionally listed and attempted to interview all households in a listing area. The result was smaller variation in the probabilities of selection.

[15] If we found an exact address match (including apartment number, if relevant) or a name-and-address match, we assumed the address had a telephone.

domain only marginally. For these reasons, we changed our approach for the remaining seven states. In these states, we listed all addresses in the replicate subsamples (see second paragraph of "Selection of PSUs," above) selected as the listing area.[16] The listers attempted to screen each address to determine whether it had a telephone. We then formed replicate subsamples of addresses that had not been determined to have a telephone and selected a portion for initial release. Because of lower-than-expected yield in the early replicates, we released all replicates for interviewing. Thus, all listed households were contacted for interviewing except those the lister had determined had a telephone.

Unlike the telephone sample, we attempted interviews in all nontelephone households that we contacted. We had originally planned to undersample insured households but found that nontelephone insured households were not as prevalent in these areas as we had anticipated.

Concluding Remarks and Observations

The study's multiple objectives required a sample design that was quite complicated. Designing a sample to interview the general population about health insurance (without the need to oversample) would be much simpler, as would a study that focused on only one of the domains (e.g., the uninsured or Medicaid). One simplification to the design presented that we would suggest in retrospect would be to eliminate the separate frames for published and unpublished telephone numbers. Our targeting of the uninsured was more effective within the unpublished frame than it was in the published.

We found that it is possible to effectively oversample the uninsured population in a telephone survey. For most states, the Medicaid group is so rare that it is very expensive to interview this group without the use of state-provided lists of recipients. In the one state where we had to rely on general population screening for this group, we were able to achieve an 11% hit rate compared with an estimated prevalence of 7%. The uninsured are more prevalent than Medicaid beneficiaries, and in any case, no list of uninsured persons is available to use as a sampling frame.

Design effects were modest given the complexity of the design. Estimated square roots of design effects ranged from a low of 1.2 (for uninsured and insured persons in several states) to a high of 2.9 for persons covered by Medicaid in one state where the state-provided Medicaid file was problematic.[17] Square roots of design effects for most groups were between 1.3 and 1.6. Most of the increase in variance appears to be due to the variability of probabilities of selection. We expected that precision for some groups would suffer because of the oversampling. While we recognized that probabilities of selection and hence sample weights would vary across domains, we had hoped to have small variation within the domains. However, the oversampling adversely affected estimates in ways we did not anticipate. The reasons for this are first, because many households had chances of selection from both the Medicaid and general population frames, and second, insurance coverage within households was much more heterogeneous than we expected.

The use of in-person interviewing contributed less than expected to the estimated design effects. In fact, for most states, design effects for many subgroups are slightly lower when the in-person sample is included. However, the small number of PSUs in each state may make the estimates of design effects unreliable. Nonetheless, in-person interviewing of nontelephone households is quite expensive, and we are currently investigating the bias from relying exclusively on telephone interviewing for some or all of the domains of interest.

[16] In some cases, the selected replicates included a very large block that was either the only block or, because of its size, comprised several replicates. In such cases, we segmented the large block, creating segments of roughly equal size, and selected one or more at random. These subsampled segments, perhaps along with other selected blocks, comprised the listing area and were listed in their entirety.

[17] A design effect is the ratio of the sample variance of the actual sample to that of a simple random sample of the same size. In the samples we report on, there are potential sources of increased variance: geographic clustering of the in-person sample, clustering of persons within families, weights to correct for oversampling, and weights to correct for nonresponse.

References

Kish, L. (1965). Survey sampling. New York: Wiley.

U.S. Bureau of the Census. (1992). Household and family characteristics: March 1991 (Current Population Reports P-20 No. 458). Washington, DC: U.S. Government Printing Office.

Comparisons of Two Sampling Frames for Surveys of the Oldest Old

Willard Rodgers

Introduction

Samples of the U.S. household population are generally obtained using a multistage clustered design that involves listing of housing units at each of a selection of small areas within larger sampling areas or through random-digit dialing (RDD) for samples of the population that can be associated with telephone numbers. Both of these methods require a large amount of screening to be done to obtain samples of rare populations, making the use of list frames appealing when such frames exist. A list frame that is often used for studies of elderly populations is that of current Medicare enrollees. This file, referred to as the Enrollment Data Base (EDB), is a special subset of the Social Security Administration's Master Beneficiary Record that was created for the Health Care Financing Administration (HCFA) and is available to governmental agencies for use as a sampling frame in surveys of elderly or disabled populations. Examples of studies that have used this frame include the National Long-Term Care Study (Manton, 1988) and the Medicare Current Beneficiary Survey (MCBS; Apodaca, Judkins, Lo, & Skellan, 1992) as well as numerous epidemiological studies and studies of health care utilization.

Concerns have been expressed about possible deficiencies in both types of frame: Area probability and RDD telephone samples may (it is feared) underrepresent certain parts of the population, such as those living with their children or with others who, for whatever reason, fail to report them when they enumerate household members during the screening operation, while the Medicare frame may underrepresent, for example, those who have not worked in jobs covered by Social Security. There have not, however, been any systematic comparisons of these sampling frames. The issue is an important one for studies of elderly populations, since a substantial saving in screening costs can be achieved by using the Medicare frame.

Past Research

Assessments of the proportion of the elderly population that is covered by Medicare have been made by Waldo and Lazenby (1984), who estimate an overall coverage rate of 97%, and by Hatten (1980), who estimates the overall coverage rate to be somewhat lower (95% to 96%). A more recent and more detailed analysis of the coverage was completed by Fisher, Baron, Malenka, Barrett, and Bubolz (1990), who compared estimates of the number of people in various age ranges as obtained from the U.S. Census with estimates based on a 5% sample of the Medicare files (the Health Insurance Skeleton Eligibility Write-off file) as of July 1985. Their estimate is that the Medicare population is approximately 96% of the total U.S. population aged 65 or older but that there is considerable range in coverage across subgroups defined by age, gender, and race. Coverage generally increases with age and is somewhat higher for females than for males and for whites than for blacks or for those with race designated as "other."

Apodaca et al. (1992) evaluated the sample for the MCBS obtained from the HCFA frame. One difficulty they encountered is with respect to introducing clustering into the sample. Without clustering of sample addresses, the costs of implementing a survey using face-to-face interviews would be prohibitive because of the travel expenses and wages associated with repeated trips to widely dispersed locations. In about 3% of the records, information about the county of residence was missing or the county codes were invalid. Moreover, it was not feasible to cluster the sample into geographic units smaller than zip code areas.

Another concern is the quality of the addresses in the EDB file. Apodaca et al. (1992) report that 95.5% of advance letters sent to the sample of more than 15,000 were apparently delivered as addressed, while 2.1% were delivered to a new address with change of address forms returned to the survey organization and 2.5% were returned unable to be delivered as addressed. Interviewers sometimes had to undertake extensive tracking to locate individuals who had moved from the address on the EDB or for whom the EDB included only a box number or other form of address that did not identify a place of residence, but only 1.2% of the sampled cases ended up as unlocatable. Other investigators, with more limited resources, have been less successful in tracking; for example, Kelsey, O'Brien, Grisso, and Hoffman (1989) report that they were unable to locate 14% of their sample in a case control study of hip fractures.

Apodaca et al. (1992) also found that 5.7% of the sampled individuals had died before they could be

Willard Rodgers is at the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

The Asset and Health Dynamics of the Oldest Old (AHEAD) project is supported by a cooperative agreement of the National Institute on Aging with the University of Michigan, Grant No. U01 AG09740-0586.

interviewed and that 6.8% of the remaining sample were living in institutions of various types. The EDB is not very useful for drawing samples from populations limited to the institutionalized or noninstitutionalized, since this cannot be reliably ascertained from information on the records.

Methods

The AHEAD Study

Asset and Health Dynamics of the Oldest Old (AHEAD) is a national survey of people aged 70 or older. The first of what is intended as a series of biennial data collections was completed in 1993–94, using a mixed-mode strategy that relied primarily on telephone interviews with those under age 80 and face-to-face interviews with those aged 80 or older. A total of 8,222 computer-assisted interviews were conducted. AHEAD is designed to provide a public use data file that will be useful for addressing issues related to the dynamics of health among the oldest old and how changes in health are related to economic status and changes, to patterns of intergenerational transfers, and to use of health care services.

Area Probability Sample

To obtain a probability sample of people aged 51 through 61 for the Health and Retirement Survey (HRS), interviewers from the Survey Research Center at the University of Michigan screened a national area probability sample of housing units using the Survey Research Center National Sample Design with oversampling in certain areas to achieve larger numbers of African Americans, Hispanics, and residents of Florida.

Survey Research Center interviewers visited households selected from these areas to determine whether or not they were occupied by households. If so, the interviewers screened them to identify those with any resident born between 1931 and 1941, the birth cohorts to be included in the HRS. At the same time, the interviewers also screened these households to identify those that contained anyone born in 1923 or before. Information was collected from each such individual to facilitate their recontact more than a year later, when the first wave of data collection for the AHEAD study would begin: their full name, address, and telephone number; the address of any other place of residence used during the year; and the names, addresses, and telephone numbers of two close friends or relatives who would know how to get in touch with the individual.

These households were the starting point for the selection of the area probability sample that was actually used in the AHEAD study. About one-sixth of the households were deleted from the frame for AHEAD. Some of the deletions were made to accommodate a cut in the funds for AHEAD. The remaining deletions were restricted to households from which at least one person born in 1913 or before had been selected and in the core sample. This part of the age 80 or older sample was replaced by the sample from the second frame, as described in the next section.

In households that contained more than one person born in 1923 or before who were not married to each other, one person or couple was selected for the sample. If a designated individual was married at the time of the first wave of date collection, his or her spouse was also asked for an interview, regardless of year of birth; those spouses born in 1923 or before are treated as part of the sample in their own right, while those born after 1923 provided information about the situation of the older spouse but are not part of the sample.

EDB Sample

HCFA provided the Survey Research Center Sampling Section with 10 tapes that contained information about all Medicare enrollees in all 274 counties from which the HRS area probability sample was selected (a total of 4,373,198 records). Using these tapes, a three-stage probability sample of Medicare enrollees born in 1913 or before was selected using a procedure that maximized the correspondence between this sample and the core sample from the area probability frame. As with the area probability sample, if a selected individual was married at the time of the first data collection, his or her spouse was also asked for an interview; if the spouse was born in 1923 or before, he or she was part of the overall sample.

Analysis Strategy

Sampling weights were developed to reflect differences in the probability of selection. These included factors for the oversamples, for the deletion of one-third of the primary sampling units from the HRS frame, and for the selection of individuals or couples from households with multiple cohort-eligible individuals. The analyses described in this paper are restricted to individuals born in 1913 or before and the spouses of such individuals, regardless of their year of birth, since it is only this subset of the population that was selected from both frames.

Findings

Sampling Efficiency Considerations

Of the 69,377 housing units selected for the HRS screening, 59,918 were found to be occupied by a household unit. All but 214 of these households (99.6%) were screened sufficiently to establish their eligibility for the HRS and/or AHEAD studies. Of these screened households, 9,474 (15.9%) were found to include at least one individual eligible for AHEAD. That is, 13.7% of the original 69,377 housing units were ascertained to have at least one resident thought to be eligible for AHEAD, meaning that 7.3

housing units were screened (or ascertained to be unoccupied) to obtain one that included one or more eligible individuals. Based on the subset of segments that were included in the AHEAD sample, we estimate that 26.5% of the 9,474 households with at least one person aged 70 or older had at least one person aged 80 or older (born in 1913 or before), implying that 27.6 housing units were screened for every household with such a person. This provides a graphic illustration of the dramatic savings in the cost of screening that can be achieved by using a list frame such as the HCFA EDB file for samples of rare populations. This saving would be tempered if, as in the present study, certain parts of the population were to be sampled at a higher rate than the remaining parts. If the distinguishing characteristic(s) were known for the list entries, this would not be an issue; for the EDB, this would apply to oversamples based on year of birth and gender and place some limitations on oversamples based on geography.

For the HCFA sample, 1,700 names of people born in 1913 or before were selected from the EDB tapes supplied by HCFA. For approximately 18% (307) of those individuals, the addresses in the file were potentially problematic. For 10.5%, there was no street address to which an interviewer could go to try to talk to the individual; for most of these, there was only a post office box number. For another 5.9% of the cases, the address was that of another individual or of a bank or other institution. And for 1.7%, the address seemed to be that of a nursing home or other long-term care institution.

Of the individuals who were selected from the HCFA frame, 20.3% were determined to be ineligible for the sample, compared with 13.0% of the 2,439 individuals selected from the HRS area probability frame. This difference is accounted for by the fact that the HRS frame was limited to those living in households at the time of the screening, whereas it was not practical to eliminate the institutionalized from the HCFA sample in advance. The interviewers determined that 15.6% of the HCFA selections were permanently in long-term care facilities, compared with 4.6% of the HRS selections (the latter being limited to those who were not considered institutionalized when the screening was done a year or so earlier). A somewhat higher proportion of the HRS selections were found to have died (7.5%, compared with 4.7% of the HCFA selections), reflecting that the EDB tapes were updated for deaths more recently than the HRS screening. The proportion who died or were institutionalized increased with age (over 40% of the HCFA selections who were born in 1903 or before were found to be ineligible, compared with 12% of those born from 1909 through 1913), but the differences between the two frames are quite consistent across the age groups.

Accuracy of Screening Information

The usefulness of a sampling frame depends on the accuracy of information used to identify individuals who are members of the target population. Errors may be either those of false inclusion or false exclusion. Errors of inclusion can generally be corrected when interviewers call upon the selected individuals and determine that they do not meet the criteria (albeit at a cost), but the frequency of such errors may be related to the frequency of errors of exclusion, which are less easily detected and remedied.

At the time of the HRS screening, interviewers collected information about the year of birth, gender, and marital status of each individual born in 1923 or before. Year of birth was the only criterion for eligibility, but marital status was used to determine whether one or two people were to be interviewed. In addition, the gender of each eligible person was recorded.

The genders of 5 of the 1,427 of the respondents who were selected from the HRS frame were misidentified at the time of the screening, 0.6% of the males and 0.2% of the females. Only two years of birth were incorrectly identified at the screening (at least relative to the answers given in the interviews), and those were only 1-year discrepancies. Almost 3% of the respondents were in a different marital status when interviewed than what they reported at the screening, but most of the discrepancies apparently reflect real changes, in particular, the death or institutionalization of some spouses in the year or more that elapsed between the screening and the interview.

The EDB file contains information on the year of birth, gender, and race of each Medicare enrollee. (The race variable is trichotomous: "white," "black," and "other.") All 992 respondents who were sampled from the HCFA list were correctly classified by gender in the enrollment file. Thirty-six of them said that they were born in a year different from what was recorded in the enrollment file. Most of the discrepancies were small: Of the 36, 23 differed by only 1 year, and only 3 differed by more than 3 years.

The race recorded in the enrollment file is apparently less accurate than gender and year of birth. As shown in Table 1, for 24 of the 992 respondents (and for 52 of the 1,699 total selections), race was not recorded in the EDB. For 21 respondents, the race recorded on the EDB differed from what the respondent reported in the interview. If the EDB were to be used to select a sample of elderly African Americans, we estimate from these data that 8.4% of those who identify themselves as such would not be represented: Race was missing for 4.8% of the 86 respondents who

Table 1. Age distribution of eligible selections from the HRS and HCFA frames compared with CPS estimates (percentages)

Age	HRS (n = 1,730)	HCFA (n = 1,534)	CPS
80–84	62.2	60.1	60.4
85–89	27.1	27.5	29.0
90+	10.8	12.4	10.6
Total	100.0	100.0	100.0

identified themselves as African American, and 3.6% were classified as "white" or "other."

Coverage of the Target Population

An indirect evaluation of the relative coverage provided by the two frames is provided by comparing the demographic composition of the samples obtained from them with each other and with outside estimates of the target population. Table 1 shows the distribution of the eligible samples obtained from the two frames across three age groups. Both samples have age distributions that are close to the distributions estimated from the Current Population Survey (CPS), although the HCFA sample may slightly overrepresent those aged 90 or older.

The HCFA sample appears to underrepresent men slightly: Of the HCFA eligible selections, 32.6% are male, compared with 35.0% according to the CPS and 35.1% of the HRS sample. In particular, the HCFA sample may underrepresent the oldest males: Just 5.7% of the HCFA males are aged 90 or older, compared with 12.0% according to the CPS and 8.9% of the HRS sampled males. The opposite pattern appears for females: Of the HCFA females, 15.5% are aged 90 or older, compared with 12.0% according to the CPS and 11.8% of the HRS females.

With respect to marital status, it appears that the HCFA sample may underrepresent those who are currently married and living with their spouses, while the HRS sample may overrepresent such people. Table 2 shows the proportion of eligible selections who were categorized as married in each of the two samples and in the CPS, both for the entire age range and for subgroups defined by age and gender. The pattern is generally consistent across these subgroups.

Tracking Considerations

The screening operation for the HRS frame was conducted from April through September 1992, but the data collection for AHEAD did not begin until late October 1993, so some effort was required to track individuals who moved, and some individuals died or were institutionalized in the interim. Tracking, however, was relatively easy to implement because of information collected during the screening visit.

Table 2. Percentages of currently married eligible selections in different age and gender groups from the HRS and HCFA frames compared with CPS estimates

	HRS	HCFA	CPS
Total sample	37.2	30.3	33.5
Gender			
Male	68.1	58.1	62.8
Female	20.4	16.9	17.8
Age			
80–84	42.1	38.4	40.2
85–89	32.1	22.4	27.4
90+	21.3	8.9	12.8

Tracking was a more difficult as well as a more frequent exercise for the HCFA sample. It has already been noted that about 18% of the addresses on the EDB file for the individuals sampled from that frame were considered potentially problematic. In addition, when letters were sent in advance to those people or when the interviewers tried to contact them, it was discovered that some of the apparently adequate addresses were out-of-date. The consequence of these inadequate or out-of-date addresses is that the interviewers needed to track a substantial proportion of the respondents. Using a wide variety of tracking procedures, they were eventually able to locate all but 3.5% of the individuals sampled from the HCFA frame and all but 0.5% of those sampled from the HRS frame.

Response Rate Comparisons

Among those in the two samples who were not determined to be ineligible (by reason of death, institutionalization, living outside of the United States, or having been born after 1913 and not married to a person born in 1913 or earlier), the response rates were 83.3% for those from the HRS frame, 72.9% for those from the HCFA frame. This is a substantial, and statistically highly significant, difference.

A possible explanation for this difference lies in the field schedule for the two samples. The field period for the AHEAD data collection began in late October 1993 and continued into June 1994. The long duration was intended, in part, to give interviewers the discretion to postpone an interview if the selected individual was temporarily ill or traveling. The entire sample was not released in October, however. Part of the sample from the HRS screens was not sent to the field until January 1994 in order to have better control over the costs of the study. The HCFA sample was not sent to the field until February 1994 because of delays in getting the EDB tapes from HCFA to the Survey Research Center. This meant that the period during which the interviews for the HCFA sample were collected was shorter than the period available for the HRS sample.

A partial test for this explanation can be made by comparing the two releases of the HRS sample with one another, since the second release of the HRS sample was about 10 weeks after the first release and only about 4 weeks before the release of the HCFA sample. The response rate for the two HRS releases were 83.4% and 82.1%, respectively, in the expected direction, but even the second release achieved a considerably higher rate than the 72.9% obtained for the HCFA sample. Moreover, there is little evidence that the interviewers were able to put less effort into recruiting and interviewing the HCFA sample as compared with the HRS sample: The average number of calls (either by telephone or in person) to all eligible

individuals was 5.42 for each actual interview in the HRS sample, which is only slightly more than the 5.35 calls for each actual interview in the HCFA sample. It may be that a higher proportion of the calls recorded for the HCFA sample were really attempts to locate the sampled individual rather than actual contacts. Moreover, it may not be the sheer number of calls that an interviewer makes that determines the response rate in an elderly population. The opportunity for the interviewer to delay a return to a household because of circumstances such as illness or family visits may be important in achieving an adequate response rate in surveys of the elderly (compare Rodgers & Herzog, 1992).

Comparisons of Respondents From the Two Frames

The ultimate criterion for assessing the adequacy of a sampling frame is whether respondents obtained from that frame accurately reflect the target population. For many of the questions that were asked on the AHEAD survey, there is no outside source of information that would permit the accuracy of the response distributions to be assessed. As an alternative, we have compared the distributions of responses obtained from the two sampling frames with one another. Differences between the two distributions would point to biases in at least one of the frames, but outside considerations would be necessary to assess which frame is more likely to be at fault.

An overview of the findings is useful before we delve into the specific differences. I made a total of 80 comparisons between the respondents from the two frames based on proportions in particular categories or means of intervally scaled variables. These included demographic characteristics, health conditions, health care behaviors and costs, housing characteristics, and economic conditions. Across those 80 comparisons, the two groups of respondents were significantly different from each other at the 5% level on 9, more than the 4 or so that would be expected in the absence of any real differences, but not a lot more, especially if the lack of independence of the tests is taken into account. The analysis also emphasized the importance of taking the sample designs into account. If I were to have ignored the complex sample design by using standard test statistics (which assume simple random samples), I would have concluded that there were 18—twice as many—statistically significant differences between the two groups. The average design effect across those 18 comparisons was 2.3 but with a considerable range (1.01 to 6.81), indicating the importance of taking account of the design with respect to the specific variable under study.

With respect to the differences that were significant at the 5% level, I first note that a smaller proportion of respondents from the HCFA frame than from the HRS frame identified themselves as black (or African American): 8.4% versus 11.4%. The proportion of HCFA respondents is closer to the proportion of the household population aged 80 or older that is black (the 1993 CPS data yield an estimate of 7.8%), so it seems to be that the HRS frame yields too high a proportion of blacks (even after weighting the data to compensate for oversampling).

Consistent with the earlier finding that the HCFA sample may underrepresent married people, a higher proportion of the respondents from the HRS frame were married (42.5%) than were those from the HCFA frame (37.6%). Correspondingly, a lower proportion of the HRS (48.6%) than of the HCFA respondents (52.8%) were widowed.

I found little support for the concern that area probability frames may fail to identify some proportion of households with elderly respondents who have severe health problems. The AHEAD respondents were asked about a series of health conditions, such as arthritis, strokes, and high blood pressure. Only 3 out of 17 differences in proportions reporting such conditions were statistically significant. One of those is with respect to a general question about heart disease, in response to which 37.3% of the HCFA respondents reported such a condition compared with only 31.9% of the HRS respondents. A second difference is that a higher proportion of the HCFA respondents reported having fractured their hip (9.1% vs. 6.1%), and the third is that more HCFA respondents reported having had cataract surgery (43.3% vs. 36.6%). It is true that consistent with the concern about missing the most unhealthy in household screens, all three of these differences are in the same direction—the HCFA respondents were more likely to report the condition than were the HRS respondents.

Out of eight comparisons with respect to ratings of overall health and of vision and hearing, depression, the number of days they spent in bed in the last month, cognitive ability, and so on, only one was statistically significant: On a 5-point scale ("excellent" to "poor" but treated as if intervally scaled), the HCFA respondents rated their vision as poorer than did the HRS respondents. Moreover, there was not a significant difference with respect to the amount of out-of-pocket expenditures for medical care, and perhaps most relevant to the hypothesis that the severely disabled are sometimes missed in household screens, the differences in the average number of activities of daily living (ADLs) and instrumental activities of daily living (IADLs) with respect to which they reported any limitations did not even approach statistical significance.

A concern about the HCFA frame is that it may underrepresent segments of the population that are less likely to be covered by Medicare, perhaps including recent immigrants and others who have not had a history of employment in jobs covered by Social Security. The HCFA respondents, however, were not significantly less likely than the HRS respondents to say that they were born outside of the United States or to report that they had not worked for 10 or more years. There also was neither a statistically significant difference with respect to peak earnings while working nor with respect to the income of the individual, couple, or household.

The final variable for which a statistically significant difference was found between the samples is intriguing. This is with respect to one of a series of questions that were

asked about the housing characteristics of the respondents. There are neither statistically significant differences in the type of place in which they live nor in the probability that they live in a building or community specially for elderly residents, but among the 11% or so who do live in such a place, the HCFA respondents were more likely to say that the place offered them group meals (48% vs. 28%). We cannot confirm this hypothesis, but it is possible that this difference reflects confusion on the part of some of the interviewers about whether or not people living in certain types of housing circumstances were eligible for the study. For example, the HRS interviewers who did the screening may have decided that some facilities that provided board and care for the residents were not households and so screened them out. On the other hand, the AHEAD interviewers, who were given more extensive training about the types of facilities to classify as households or as long-term care facilities, may have classified the same housing units as eligible, especially if they were searching for a specific named individual from the HCFA frame. This emphasizes the need for careful training of interviewers with respect to the eligibility rules for determining whether or not an individual is living in a household rather than in group quarters.

Summary and Conclusions

Comparisons of a sample of the U.S. population aged 80 or older from an area probability frame with a sample from the HCFA list of Medicare enrollees indicate that both frames provide information that is almost always accurate (or at least consistent with respondent reports) with respect to gender and year of birth. The race indicator is sometimes missing from the HCFA file and sometimes inaccurate, perhaps especially for nonwhites. The marital status obtained from the household screening conducted more than a year before the start of interviewing was often inaccurate because of changes during the intervening months.

A substantial proportion of addresses on the HCFA file are not street addresses, and others are out-of-date, so interviewers had to do considerable tracking to locate those selected from the HCFA frame and were unable to do so for about 1 out of 30 selections. Moreover, about a fifth of those who were located turned out to be ineligible for the AHEAD study, primarily because of death or institutionalization. The HCFA frame apparently underrepresented males, those over age 90, and married people, while the area probability sample overrepresented married people.

The response rate for eligible selections from the HCFA frame was about 10 percentage points lower than that for the area probability frame. About a quarter of this difference can be attributed to inability to locate some of those selected from the HCFA frame, and much if not all of the remaining difference may be due to the shorter field period available for the HCFA sample, but the possibility of a less idiosyncratic explanation cannot yet be ruled out.

Comparisons of the respondents selected from the two frames show that the null hypothesis of no difference cannot be rejected at conventional levels for most characteristics, and given the number of comparisons made, the small number that are nominally significant should be regarded with caution. In particular, there is no convincing evidence that the two samples differ with respect to their health, thus offering at least some assurance that area probability samples do not suffer from a substantial bias due to failure to identify households with the most elderly or least healthy. Similarly, there is little evidence in the present data to indicate that the HCFA frame introduces bias due to underrepresentation of groups such as recent immigrants or others who have not achieved rights to Social Security benefits.

Given the lack of evidence of substantial coverage or nonresponse biases in the two frames relative to one another or to outside evidence about the target population, cost considerations become more important, and the balance clearly shifts in favor of the HCFA frame. To obtain the area probability sample, more than 25 housing units had to be screened to identify a single household with someone aged 80 or older. The costs of screening for such a rare population would be a major proportion of the overall field costs. In the present case, the cost of screening almost 70,000 households to obtain a sample of 1,600 to 1,700 people aged 80 or older would have been a substantial multiple of the costs of the actual data collection if the screening had been done exclusively for finding individuals in that age range. The cost of obtaining the sample from the HCFA list is trivial by comparison. Based on these considerations, it is likely that if new cohorts are added to the AHEAD sample in subsequent years, the samples of those cohorts will be drawn from the HCFA frame.

References

Apodaca, R., Judkins, D., Lo, A., & Skellan, K. (1992). Sampling from HCFA lists. American Statistical Association 1992 Proceedings of the Section on Survey Research Methods, 250–255.

Fisher, E. S., Baron, J. A., Malenka, D. J., Barrett, J., & Bubolz, T. A. (1990). Overcoming potential pitfalls in the use of Medicare data for epidemiologic research. American Journal of Public Health, 80, 1487–1490.

Hatten, J. (1980). Medicare's common denominator: The covered population. Health Care Financing Review, 2, 53–63.

Kelsey, J. L., O'Brien, L. A., Grisso, J. A., & Hoffman, S. (1989). Issues in carrying out epidemiologic research in the elderly. American Journal of Epidemiology, 130, 857–866.

Manton, K. G. (1988). A longitudinal study of functional change and mortality in the United States. Journal of Gerontology, 88, S153–161.

Rodgers, W. L., & Herzog, A. R. (1992). Collecting data about the oldest old: Problems and procedures. In R. M. Suzman, D. P. Willis, & K. G. Manton (Eds.), The oldest old (pp. 135–156). New York: Oxford University Press.

Waldo, D. R., & Lazenby, H. C. (1984). Demographic characteristics and health care use and expenditures by the aged in the United States: 1977–1984. Health Care Financing Review, 6, 1–29.

Comparison of Varying Consent Methodologies in a Follow-up Study of Hospital Inpatients and Outpatients

Christina H. Park and Catharine W. Burt

Introduction

The U.S. health care system has undergone dramatic changes in the past two decades and is expected to change even more. The pressure to control the escalating costs of health care as well as to make the care more accessible to all people is forcing the health care system to be restructured even in the absence of national health care reform. The impact of these changes includes a greater diversity in health insurance and benefit programs, development and growth in new and alternative settings of health care, and changes in the medical care received by patients and in the use of medical care technology (Division of Health Care Statistics, National Center for Health Statistics [NCHS], 1992). As a result, there is increasing demand for information on various aspects of health care to allow researchers and policy makers to adequately assess the effects of these changes. An important current concern in health policy research is the evaluation of the effect of various types and patterns of treatment and medical care on patient well-being. Measuring effectiveness of medical practice on a wide range of patients is being recognized as an important research area that can provide useful information to many players in and users of the health care market.

Recognizing the need to be responsive to the changes occurring in the health care system, NCHS recently initiated restructuring and expansion of its existing surveys of health care providers and service settings into an integrated system called the NHCS. One component of the NHCS that is under development is the provider-based patient follow-up component. Starting with the sample event—that is, visit, discharge, or admission—the patients are to be followed up on periodically to collect information beyond what can be obtained from providers. By collecting information on the patients' health and treatment experience from the patients or their family members, especially over time, issues such as the effectiveness of medical treatments, episodes of care for specific diseases, patient satisfaction and quality of care, and the dynamics of health care use and health status can be addressed.

One major operational problem in the event-based follow-up design, however, is the need to obtain consent from patients for their health provider to release to NCHS the identifying information necessary in contacting the patients. This consent requirement is doubly difficult because permission and cooperation are needed from both the provider and the patient for follow-up. To address some of the methodological, legal, and feasibility issues surrounding the consent requirement and to investigate other issues related to the conduct of patient follow-up studies, a feasibility study of following up hospital patients was initiated.

Patient follow-up as defined in this study involved collecting a variety of health-related information from patients (or proxies) and linking this information with their medical records data from the sampled health care event. It would also be possible for further linkage with Medicare data or the National Death Index, if desired. Because confidentiality is a major issue, the primary purpose of this feasibility study was to test different methods of obtaining hospital and patient consent in order to determine which method produces the highest participation at both the hospital and the patient level. The current paper addresses this objective only and does not address other issues that the feasibility study investigated, such as sampling, appropriateness of the questionnaire, success of proxy interviews, and so forth.

Methods

The feasibility study was conducted under a contract with a private survey research firm, Westat, Inc., and the Johns Hopkins School of Hygiene and Public Health from September 1991 to December 1994. During the conceptual development phase of the feasibility study, decisions were made to follow hospital events as opposed to other provider-based events. The patient population included adults who were 18 years old or older who had either an inpatient stay (called "inpatient") or an ambulatory care visit to the emergency or outpatient department of a hospital (called "outpatient"). These patients were then sampled for specific diagnoses to be included in the study.

Christina H. Park and Catharine W. Burt are with the Division of Health Care Statistics, National Center for Health Statistics, Hyattsville, Maryland.

The authors would like to thank the staffs at Westat, the Johns Hopkins School of Hygiene and Public Health, and the National Center for Health Statistics, who contributed to the development and conduct of the study.

A review of state laws conducted during the early phase of the study revealed that most state laws prohibit the release of medical records information containing patient identifying information without prior formal authorization of the patient. Thus, a study design was developed in which participating hospitals were asked to obtain patient consent for the hospital to release the patient's identifying information to the study team (see Figure 1). A test of three different methods of obtaining patient consent was designed: prospective, retrospective passive (retro-passive), and retrospective active (retro-active). In the prospective design, participating hospitals were asked to obtain consent from all patients at the time of the admission or at the time of the visit. In the retrospective designs, hospitals were asked to obtain consent by mail after patients had been discharged from the hospital and had been sampled for specific conditions. In the retro-active design, patients were asked to mail in a reply card if they consented; in the retro-passive design, patients were asked to send in a reply card if they did not consent.

Eighty hospitals were sampled from all regions of the United States, about 20 from each of the Census's four regions—Northeast, South, Midwest, and West. Hospitals were selected from both urban and rural areas. Hospitals were equally divided and randomly designated as inpatient hospitals (i.e., inpatient samples were provided) or outpatient hospitals (i.e., outpatient samples were provided). Each group of 40 inpatient or outpatient hospitals were then equally and randomly assigned to either the prospective or retro-passive consent methodology. The retro-active consent method was not offered initially, but offered only when the hospital refused to agree to either of the other methods.

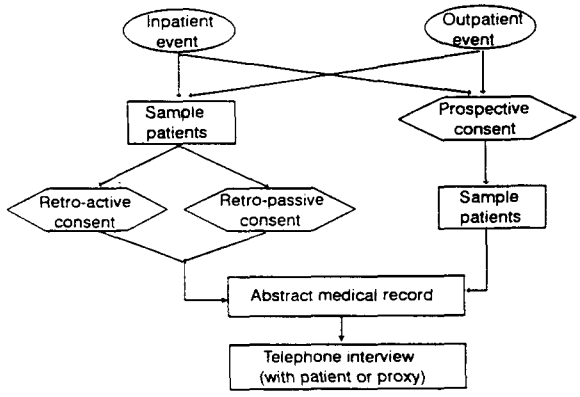
The hospital enrollment procedure was carried out by specially trained field representatives. The sampled hospitals were first contacted by telephone to screen for eligibility and, if determined eligible, to set up an appointment for an induction interview. An induction interview was conducted in person with a hospital administrator (or administrators) to collect hospital information as well as to explain the study methodology. Hospital enrollment became a lengthy process in some hospitals because they required additional review and approval processes beyond those of the hospital administrator. An extensive refusal conversion effort was also carried out offering up to $500 of payment to defray any costs to the hospital.

Once hospital approval for study participation was obtained, the field representatives worked with the hospitals to select patients and to obtain their consent. In the prospective design, all patients being admitted were asked for consent within a specified period (8 weeks for inpatient hospitals and 4 weeks for outpatient hospitals). Once this specified period ended, patients with specific conditions of interest were sampled from all patients seen at that hospital, based on the discharge list for the 4-week sampling period (the last 4 weeks of the 8-week consent period for inpatient hospitals). In the retrospective consent designs, patients were first sampled from a specified 4-week sampling period based on the discharge diagnosis. After patients had been discharged, they were then mailed consent cards to which they replied actively or passively.

Upon receipt of consent from the patients, the field representatives proceeded with medical records abstraction. The abstract forms were the same as those used in the National Hospital Discharge Survey and the National Hospital Ambulatory Medical Care Survey, which collect only a small amount of information. The identifying information of the consenting patient was also abstracted from the medical records and was sent to the telephone center for follow-up interviews, which were to be conducted approximately 2 months after the hospital stay or visit.

The consenting patients (or their proxies if the patients were deceased or too ill) were first mailed a letter explaining the purpose and importance of the telephone follow-up interview. They were then called for the interview, which averaged about 40 minutes in length. The inpatient and outpatient questionnaires differed slightly but covered similar topic areas, such as episode of care, pretreatment health status, current health status, insurance coverage, and demographics.

Figure 1. Diagram of the study design



Results

Among the 80 sampled hospitals, 75 were eligible for participation in the study. The remaining 5 were found to be ineligible during the hospital enrollment period because they had either closed or merged with other hospitals. The enrollment experiences of the inpatient and outpatient hospitals were similar and thus are shown together in Table 1. Combined across all consent methods, a total of 51 hospitals participated in the study, giving a response rate of 68%. Many hospitals assigned to the prospective consent method switched over to a different method, making their participation rate very low—23%. The hospital participation rate for the retro-passive method as originally assigned was 47%. The response rate for the retro-active method could not be determined since none of the hospitals were originally assigned to this method.

Table 1. Hospital participation by consent method

Hospitals	Retro-passive No.	%	Retro-active No.	Prospective No.	%	Total No.	%
Originally assigned	36		0	39		75	
Participated by final method	27		14	10		51	68
Participated by assigned method	17	47	—[a]	9	23	26	35

[a]Not applicable because no hospital was originally assigned to this method.

Small hospitals in rural areas were the most likely to accept the prospective consent method. On the other hand, the retro-passive consent method was most acceptable to the hospitals in the Northeast and least acceptable to the hospitals in the West (see Figure 2). Hospitals in the Midwest refused the most often (53%). Active refusal conversion efforts were instrumental in converting 10 of the initial 34 refusals. The main reason for refusal, cited by 13 hospitals, was that the hospital had insufficient staff resources to support the study. Four of these were specifically due to the process of hospital reorganization. Seven hospitals took the remuneration of $500 for participation in the study.

From each of the 51 participating hospitals, up to 40 patients were sampled, resulting in a total of 1,465 sampled patients. Fifty-one percent of the 833 sampled inpatients consented, and 66% of the 632 sampled outpatients consented (see Table 2). The combined patient consent rates for both inpatients and outpatients for the retro-passive, retro-active, and prospective methods were 81%, 30%, and 20%, respectively. The consent rates did not differ much by patient type, except that the consent rate was higher for outpatients than inpatients on the retro-passive method (85% vs. 76%), and the consent rate was extremely low among outpatients in comparison with inpatients on the prospective method (6% vs. 27%).

Only the patients who consented retrospectively (n = 794) were contacted for telephone follow-up interviews, which occurred 2 to 8 months after the hospital stay or visit. A total of 592 interviews were completed, giving a response rate of 75%. The interview completion rate was higher for inpatients than for outpatients—81% and 68%, respectively (see Table 3). However, this difference was not due to the higher refusal rates for outpatients. It was mostly due to the higher rates of nonlocatables among the outpatients in comparison with the inpatients (15% and 6%, respectively). The retro-active consent method resulted in higher interview completion rates in comparison with the retro-passive consent method. This was expected because the individuals who responded to the retro-active method had already consented by mailing in the reply card.

Table 4 presents the overall response rates obtained by combining consent rates at the various stages of the study. The retro-passive method produced the highest combined response rates for both inpatients and outpatients. For the inpatient group, the rate was 30% when computed according to the assigned method and 41% when computed using the 70% hospital participation rate across all consent methods. For the outpatient group, the rate was 26% according to the assigned method and 38% considering the

Figure 2. Hospital participation by region and consent method

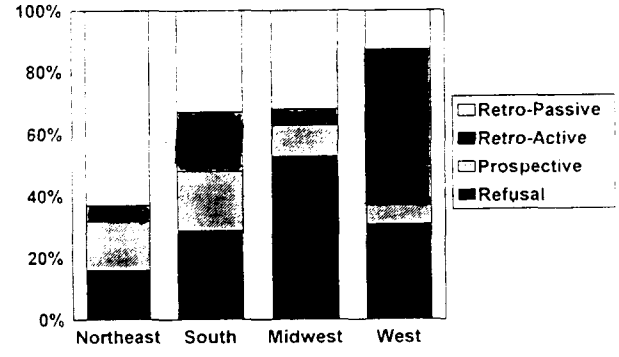


Table 2. Patient consent rates by patient type and consent method (percentages)

	Retro-passive	Retro-active	Prospective	Total
Inpatients	76	30	27	51
Outpatients	85	32	6	66
Total	81	30	20	58

Table 3. Distribution of telephone interview results by patient type and consent method (percentages)

	Inpatients			Outpatients		
	Total	Retro-passive	Retro-active	Total	Retro-passive	Retro-active
Completed interview	81	78	92	68	67	84
Refusal	8	10	1	9	10	3
Nonlocatable	6	7	4	15	16	9
Other	5	5	3	7	7	4
Total	100	100	100	100	100	100

Table 4. Response rates at various stages of the study and combined response rates by patient type and consent method (percentages)

	Inpatients			Outpatients		
	Retro-passive	Retro-active	Prospective	Retro-passive	Retro-active	Prospective
Hospital participation as assigned	50	—	24	45	—	22
Patient consent	76	30	27	85	32	6
Telephone interview	78	92	—	67	84	—
Combined response rate						
By assigned method	30	—	—	26	—	—
By any method[a]	41	19	—	38	18	—

NOTE: — indicates response rate not computed because either the corresponding task or the component parts not completed.

[a]Combined response rates by any method were computed using the hospital participation rates (by any method) of 70% and 66% for inpatients and outpatients, respectively.

66% hospital participation rate across all consent methods. The prospective consent method elicited such low consent rates at the hospital consent and patient consent levels that patients assigned to this method were not contacted for the telephone interview.

Discussion

This feasibility study, with its complex methodological design, revealed several important and interesting findings. Hospital enrollment proved to be the most difficult component of the study in that it required considerably more time and effort at every stage of the process than had been originally anticipated. Sixteen of the 75 eligible hospitals (21%) required approval beyond that of the administrator. These formal approval processes included internal review boards, medical staff reviews, legal counsel reviews, and others. Hospitals in urban areas in the West had the most stringent approval processes, with 73% of hospitals requiring one or more of these formal approvals. It is interesting to note that the West is the region where the retro-active consent method was overwhelmingly preferred over the retro-passive consent method by hospitals, contrary to the experience in other regions.

Many hospitals were reluctant to participate in the prospective consent design clearly because it required considerable work on the part of the hospital staff. The hospital staff had to administer the consent form to every patient being admitted during the consent period, which required hospital resources and commitment. In fact, the low patient consent rate (20%) achieved for this method, particularly among outpatients, may be a reflection of some failure of the hospitals to fully administer the prospective consent procedures to their patients.

Getting hospital approval is absolutely pivotal to the success or failure of any future hospital-based patient follow-up study. Because of this, the role of field representatives who conduct negotiations with hospital administrators is critical. In this feasibility study, each of the eight field representatives worked in different regions of the country; thus, it was not possible to separate their effects from regional effects. However, in general, the field representatives who were the most successful were those who emphasized the sponsorship of the study and the ease with which the study could be completed at the facility.

At the patient level, the retro-passive consent method was the most successful in obtaining consent, as was expected. However, it was still surprising to find that 19% of the patients sent in the reply card to indicate that they did not consent to the release of their identifying information. Higher consent rates had been achieved for passive consent methods in other studies, albeit for different purposes (Ellickson & Hawes, 1989). In comparing inpatients with outpatients, the follow-up interview completion rate was significantly lower for the outpatients, mainly because a larger proportion of these patients could not be located. This is probably a major reason for the seemingly higher consent rate obtained from the outpatients than the inpatients under the retro-passive consent method. This finding implies that many in this outpatient group are either highly mobile or that their identifying information provided to the hospital is less accurate than that provided by the inpatients.

Considering the consent rates at all levels, the retro-passive consent method yielded the highest response rates; however, they were still unacceptably low. Flexibility in consent method appears to be important in gaining hospital approval. Some hospitals are so sensitive to the patient confidentiality issue that limiting future studies to the retro-passive method alone would be unacceptable to such hospitals.

Because of the final low response rates obtained in this study and because of other remaining technical issues, more work is needed in these specific areas before a national patient follow-up study can be launched. Certainly, improvements in the hospital participation rate could be made by expending more resources on hospital induction. In addition to overcoming operational hurdles, many key methodological issues in the patient outcomes research as identified by groups of researchers (Fowler, Cleary, Magaziner, Patrick, & Benjamin, 1994; Maklan, Greene, & Cummings, 1994) need to be further studied and re

solved. These include issues related to characterizing diseases or conditions, controlling for variations in practice patterns and patient factors, and interpreting the significance of outcomes data, to name a few. Although a number of patient follow-up or outcomes studies have been successfully conducted in the past, these have been limited to specific patient population groups, to certain geographic areas, or to narrowly defined research topic areas. Setting up of a patient follow-up study at a national level with a broader goal of assessing the medical effectiveness on a wide range of patients requires more careful planning and further methodological development.

References

Division of Health Care Statistics, NCHS. (1992). NCHS Plan for a national health care survey. In G. S. Wunderlich (Ed.), Toward a national health care survey. Washington, DC: National Academy Press.

Ellickson, P. L., & Hawes, J. A. (1989). An assessment of active versus passive methods for obtaining parental consent. Evaluation Review, 13, 45–55.

Fowler, F. J., Cleary, P. D., Magaziner, J., Patrick, D. L., & Benjamin, K. L. (1994). Methodological issues in measuring patient-reported outcomes: The agenda of the work group on outcomes assessment. Medical Care, 32(Suppl. 7) JS65–JS76.

Maklan, C. W., Greene, R., & Cummings, M. A. (1994). Methodological challenges and innovations in patient outcomes research. Medical Care, 32(Suppl. 7), JS13–JS21.

Optimizing the Trade-off Between Cost and Quality

Seymour Sudman

Although this session is labeled "Sampling and Cooperation," these are not two distinct topics but rather simply two aspects of the same issue—what sample data collection method should be selected to optimize the trade-off between costs and quality. There are two major themes running through this set of papers. One is the use of dual or multiple frames. The other is the sampling of rare and difficult populations. The populations sampled include respondents over 80 years of age, the medically uninsured, drug users, and prostitutes. Obviously, these populations present real challenges that are met by the methods discussed in these papers.

The Park and Burt paper, which compares various methods of obtaining consent for a follow-up study of hospital patients, is a graphic illustration of this point. Suppose one is interested in patients who have had either an inpatient stay or an ambulatory care visit to the emergency or outpatient department of a hospital and who had a specific diagnosis. There are essentially two ways that one could obtain such a population. One could start with a sample of the population in households and screen this population to obtain eligible respondents, possibly using some of the methods that are suggested for sampling rare populations. Alternatively, one could start with a sample of hospital records and from these records, track down the desired population.

For anyone who has not had experience with these alternatives, the solution would seem clear. The household screening would be expensive and difficult since one would be searching for a rare population. The fraction of households with a member who has had a hospital visit in, say, the past year is around 10%, but the fraction with a hospital visit for a specific diagnosis would be a small fraction of that, say, 1% or 2% at most. Starting from hospital records would not require any screening, so the records would certainly be more efficient and, thus, might appear to be the preferred method.

As anyone who has ever tried to obtain hospital records knows, and as the Park and Burt paper makes abundantly clear, however, it is generally a mistake to start from hospital records. Because of issues of cooperation from both the hospitals and patients, the apparently costly screening gives much better cooperation rates and may, in the long run, actually cost less or no more. This is because of the extensive professional time spent with individual hospitals attempting to obtain cooperation. It would be helpful if any available cost or time data were included in the Park and Burt paper.

Some issues that are not raised in the paper need discussion. There is often a limited time schedule for any health survey so that data will be available for policy decisions. Obtaining cooperation from a number of hospitals often extends the period of the study for many months as hospital human subjects committees and boards deliberate about granting permission.

Extending the time period also increases staff time and costs. Again, it would be helpful if data about the length of the field period could be included in the Park and Burt paper.

Turning to the quality of the sample obtained, it is evident that regardless of the method used, cooperation is a major problem. While 51 hospitals, or 68% of all hospitals, ultimately agreed to participate, 14 of these required retroactive consent from patients. The retroactive response rates from patients were approximately 30%, a rate that is very typical of such methods. If these 14 hospitals are excluded, the cooperation rate drops to 49% (37/75).

Ten hospitals (23% of those invited) participated in prospective recruiting, and here the cooperation rate among patients was only 20%. It seems likely that this very low patient cooperation rate may well be caused by busy hospital staff making only minimal efforts, if any, to obtain cooperation. Clearly, prospective recruiting does not work.

Under the best treatment, retro-passive, the overall cooperation rate is (.81) (.49) or about 40%. This evidence, along with the other problems, clearly suggests that national patient surveys using hospitals to generate the sample are not the way to go.

Is there ever a place for hospital-based samples? I think there is when only a single local hospital or a small number of hospitals in a few locations for program evaluation or to test a new procedure are needed. Then, especially if members of the medical staff of the hospital are involved as investigators in the research, it is possible to get hospital participation and agreement on a consent method, such as passive consent, that does not exclude most of the sample.

Seymour Sudman is at the University of Illinois at Urbana-Champaign in the Department of Business Administration and at the Survey Research Laboratory.

To sum up, Park and Burt present some useful data that warn us about assuming that it is easy to get national samples from hospital records. While Park and Burt continue to believe that hospital cooperation rates can be improved by expending more resources, my conclusion is more pessimistic. Remember, however, the alternative of household screening.

The paper by O'Brien, Murray, Rahimian, and Wiebel raises a different issue related to data collection mode—whether telephone methods are adequate for obtaining estimates of sexual risk behaviors and crack cocaine use. Overall, in the United States, about 94% of all households have access to a telephone, but this percentage may be significantly lower in some rural and low-income areas. In the two low-income Chicago neighborhoods studied, O'Brien et al. estimate that telephone coverage was only 70%. The magnitude of biases that result from individuals in households without telephones is a function of the differences in behavior between those with and without phones, as well as the size of the group without phones.

Typically, what is done in studies that measure phone biases is to compare the results that are obtained from samples of phone and nonphone households with those obtained from the total sample. I would have liked to have seen such comparisons in their Tables 5 through 7. In the absence of such comparisons, I made my own comparisons by combining results from the various tables. My Table 1 summarizes these comparisons.

There are clearly differences between phone and nonphone households for each of these variables, but the differences vary in magnitude and importance. Thus, there is a large difference in crack cocaine use between phone and nonphone households that would result in a substantial underestimate (6.8% as compared to the actual 10%) if only a phone survey had been conducted. On the other hand, smaller differences are found between the total sample and phone sample for injection drug use and incidence of syphilis or gonorrhea, and there is only a difference of .1 between the estimate for the total sample of 4.4% HIV seroprevalence and the estimate of 4.3% for the phone household sample.

Table 1. Comparison of telephone and nontelephone household variables (percentages)

	All households (n = 264)	Phone households (n = 176)	Nonphone households (n = 78)
Below poverty	48.7	44.3	65.8
Receive welfare	25.0[a]	20.5	38.5
Employed	60.0[a]	68.8	47.4
Spouse present	50.0[a]	58.0	38.5
Risk behavior			
Sexually active	86.5	84.4	91.0
Gay male	8.6	9.1	7.7
IDU[b]	6.0	4.8	9.0
Crack cocaine in household	10.0	6.8	18.4
Diseases			
Had syphilis	5.0	3.8	7.7
Had gonorrhea	12.0	11.3	14.1
HIV seropositive	4.4	4.3	5.1

[a]Rounded estimate based on discussion in text.

[b]"IDU" stands for injection drug user.

What does one conclude from this for studies of individuals engaged in high-risk behaviors? Are telephone surveys adequate, or are the more expensive face-to-face interviews necessary? It really depends on the level of accuracy required for making public policy decisions, as well as the geographic area being covered. For a local study of a low-income area, the costs of face-to-face interviewing decline if there is a large sample, so face-to-face interviewing is the best alternative if high accuracy is needed. For national studies, especially if resources are limited, the alternatives would seem to be a pure telephone survey, recognizing that some sample biases may be present, versus a telephone survey with a subsample of face-to-face interviews in households without telephones to adjust for the phone sample biases. It must also be recognized that the questions asked in surveys like the one reported by O'Brien et al. are threatening to many respondents, so that response errors may swamp any of the potential sample biases.

I turn last to the paper by Rodgers, which compares sample results from a list sample of the elderly obtained from Health Care Financing Administration (HCFA) and a sample obtained using standard area probability sampling methods and screening for the desired age groups. As the paper points out, and as is obvious in any comparison of the use of lists with screening, list samples are generally much cheaper, with the ratio of costs between list samples and screening depending on the rarity of the population.

Lists have problems that need to be evaluated before they can be used. The HCFA list was problematic for about 18% of the names: Street addresses were missing, the addresses of individuals or institutions acting for the person listed were used, as were the addresses of nursing homes and other long-term care institutions for some recipients. Of course, some of these problems would actually be benefits if the study were to include institutionalized persons. Lists typically have some time delay, so that people who move between the time the list is compiled and the time of interview are missed. There are also people who move after the area sample screening has been conducted, but this is a small number if, as is typical, the interviewing follows shortly after the screening.

A major difference in this study is that cooperation from the area sample screening frame was 83% on the final study compared with 73% on the HCFA frame. The reasons for this difference are not really clear. There are always two major reasons for noncooperation—nonlocation of respondents and actual refusal of respondents who are contacted. It would be useful to have separate information for these two components. A priori, one might speculate that nonlocation of HCFA respondents is the primary cause, for reasons I mentioned a bit earlier.

Most crucial, of course, is the actual comparison of responses from the two surveys. On the primary health variables of interest, there are really no important differences. Some of the differences that appear to be statistically significant may well be the consequence of multiple significance tests being conducted. There are some demographic differences observed, but these may well be a function of differential cooperation or of the weighting schemes used.

I agree with Rodgers on which frame one would use for surveys of the oldest old. The HCFA frame is much cheaper than screening, and its limitations are relatively small. Primarily, these limitations relate to the loss of respondents who move and cannot be located. Clearly, the HCFA is the frame of choice if it is available. If one needed very precise estimates and substantial resources were available, a dual frame design could be used with most of the sample coming from the HCFA frame and a small percentage from an area sample screening frame.

My discussion focuses on only three of the papers, but the same comments relate to the other three papers in this session. In each of them, the same trade-off issues surface. For example, in the fascinating paper by Berry, Duan, and Kanouse, I would conclude that prison samples of street prostitutes are pretty good, since it appears that being arrested is almost like being chosen for a random sample. Clearly, samples selected from clinics are not representative of the population.

If I have a bone to pick with the papers in this session, it is that cost concerns are an important, if unarticulated, part of each of them. I would have liked to have a more explicit recognition of this point and some cost comparisons of the alternatives, such as those given in the Rodgers paper.

Discussion: Sampling and Cooperation

James M. Lepkowski

Some difficult survey design problems concern the development and manipulation of the set of materials used to select the sample, the sampling frame. The issues of frame construction are made all the more challenging when combined with considerations of sampling rare or elusive populations and obtaining adequate levels of cooperation with survey subjects who may not want to be found, let alone interviewed. The papers in this session by Sandra Berry, Naihua Duan, and David Kanouse; Robert Bray, Sara Wheeless, and Larry Kroutil; and John Hall illustrate a number of the problems and difficulties encountered in both frame construction and eliciting adequate levels of cooperation when attempting to select a sample from groups that are difficult to study.

There are several common themes across these papers about which it is useful to comment before turning to a few aspects of each paper. Three themes common to all of these papers—probability sampling, surveying difficult-to-study populations, and the use of multiple frames—are addressed first before turning to a few comments on each paper.

Three Common Design Themes

All three papers employ probability sampling methods to select samples for quite diverse problems. Probability sampling requires that at least in principle, every element in the population has a known and nonzero chance of selection into the sample. It suggests that the population is finite or at least countable and that one could conduct a census to collect the data of interest using a list of all eligible members of the population. This finite population and chance selection framework establishes a statistically sound basis for inference from sample to population.

The enumerability requirement poses the greatest challenge to probability sampling. The assumption that a list of the entire population, or at least a set of materials that could in principle be used to construct such a list, is necessary for selection makes the sample selection and survey design potentially more expensive than convenience, purposive, or quota sampling methods. All three papers assume that probability sampling methods are necessary for the populations of interest. The Berry et al. paper directly addresses the issue of using a competing strategy, convenience sampling. The problem is whether convenience sampling methods, which would probably be substantially less expensive to implement for the population of interest, provide adequate information for inference for the kinds of characteristics being studied.

A second theme in each of the papers, although handled quite distinctively in each paper, is a topic about which much has been written in the last decade: sampling rare or elusive populations. Policy and academic interest about the homeless, the uninsured, those engaged in risky health behaviors, those receiving certain types of transfer or in-kind income, and various groups in institutions, has stimulated development of sampling methods to deal with difficult-to-study groups. Concern has been about how to design a cost-efficient sample for a population that is traditionally difficult to find and how to elicit cooperation from those selected into the sample.

Berry et al. have one of the most difficult-to-study populations imaginable, street prostitutes in a U.S. metropolitan area. Bray et al. supplement a traditional household frame with samples of institutionalized populations and of homeless persons. These are groups for which frames have been developed for censuses and probability sampling, but they are now combined with the household population in one sample design. Hall has an oversampling of Medicaid and uninsured populations with which to contend. While these populations are not necessarily rare in the general population, the mode requirements imposed to reduce data collection costs create rare populations, such as Medicaid recipients or uninsured families without telephones.

These studies illustrate to varying degrees the four key issues that emerge as central problems when dealing with these difficult-to-study groups: determining who is and is not a member of the population, developing a list for sample selection, obtaining access to the population members, and gaining their cooperation.

All three papers deal with the determination of who is in the population or at least in a portion of the population that must be sampled. Berry et al. developed rules to establish who is a street prostitute, ultimately relying on self-reports to identify street prostitutes. Bray et al. had to determine who was homeless. The definition of homelessness is not straightforward, since some persons are at imminent risk of

James M. Lepkowski is a Senior Study Director at the Institute for Social Research and an Associate Professor in the Department of Biostatistics at the University of Michigan, Ann Arbor.

homelessness, even though they may currently be residing in a home. Hall had to determine the insurance status of families, a task complicated by the variation in coverage among members of the same family.

Even if the difficult issues of creating an operational population definition are resolved, some type of frame or listing must be developed. Berry et al. had perhaps the most difficult listing problem for street prostitutes, but the listing problem for the other two studies was substantial as well. No official or unofficial listing of street prostitutes exists. The listing operation was further complicated by the need to sample across space and time simultaneously. Their sampling frame was a set of materials including locations where street prostitution occurs, times of day and days of the week when it occurs, and the list of women approached at sample locations and times who identified themselves in screening questions as street prostitutes. Bray et al. had to obtain lists of institutions and persons within them for the institutionalized portion of their sample. For the homeless, they had to identify multiple potential locations where the homeless could be found, such as vacant buildings, city parks, cars, or even the streets, and list eligible persons at those locations. Hall was able to obtain lists of current Medicaid recipients in all but one of the 10 states in his study, but the uninsured can only be identified through personal interview.

The third issue in dealing with these difficult-to-study populations is obtaining access to the members. Berry et al. illustrate the problems inherent in trying to reach an elusive group that shuns contact with official sources. They contended with the difficulties of approaching persons possibly engaged in illegal activities on the street in high-crime areas at times of the day that exposed their interviewers to potentially dangerous situations. Bray et al. had potentially to contend with similar issues in approaching the homeless. They also had to deal with obtaining access to the institutionalized population by first gaining the cooperation of an institution. They were successful in a high percentage of the institutional contacts, but no contact could be established with any of the population members in institutions that refused to cooperate. Hall also faced problems obtaining lists of Medicaid recipients in 10 states. In the one state that did not provide lists, special sampling procedures had to be developed to screen for Medicaid recipiency in general telephone and face-to-face interviewing modes.

The last issue that makes studying these kinds of populations so difficult is gaining cooperation of the individual members once they have been contacted and selected for interview. What is remarkable about two of these studies is the high degree of cooperation that they did achieve. Berry et al. were able to gain the cooperation of a significant proportion of those approached for screening and of those who were eligible for the study based on the screening criteria. They do not share in detail the techniques interviewers employed to gain such significant levels of cooperation. Bray et al. also do not share specific techniques on obtaining cooperation, but they do report very high levels, even for the homeless population, for which one would expect substantial difficulty in obtaining cooperation. They demonstrate that not only can the issues of definition, listing, and access be satisfactorily addressed in a study of the homeless, but also high levels of cooperation can be elicited as well.

Finally, obtaining complete coverage of the target population in many studies leads naturally to a consideration of multiple frames. No one frame is suitable for the complete coverage of the entire population demanded in the Bray et al. study of drug abuse. As they note, traditional household frames have been criticized for leaving out institutional and homeless groups, in which drug abuse rates are higher than for the household population. Multiple frames are one solution to an incomplete frame. In this instance, the household frame contained the vast majority of drug users, but the supplemental institutional and homeless surveys covered populations that have higher drug use prevalence. This is a typical multiple frame application in which coverage is the issue. Hall also used multiple frames but for a different purpose, cost efficiency. As in the drug abuse surveys, several frames in the Hall study contained high concentrations of the subgroups of interest. Screening those frames with their higher "hit rates" could have reduced the overall costs of data collection, particularly for smaller subgroups, such as Medicaid recipients.

There are two aspects of multiple frame sampling that could be useful in studies such as these in improving the efficiency of the sample design. Hartley (1962) proposes a post-stratified estimation procedure for multiple frame designs that could lead to more precise estimates. The approach obtains optimum mixing parameters for combining estimates across frames and is an alternative to the weighting schemes examined by Bray et al. In addition, multiple frames may be used in combination with optimum allocation in stratified sampling to obtain estimates with the best precision (see, e.g., Kalton & Anderson, 1986). It would be of some interest to examine the allocation of sample across frames in the study reported by Hall to determine if the cost efficiency per unit variance could be improved further. Hall does not provide detail about the allocation of sample across frames to allow an assessment of whether some further gains are possible.

Specific Features of the Studies

In addition to these common design themes, there are several features of each of these papers that deserve additional comment.

Berry et al. demonstrate that it is feasible to conduct a survey of such an elusive group as prostitutes. They then present a clever use of the data in a quasi-experimental study comparing those from the probability sample with artificially constructed convenience samples drawn from the same probability sample. The results are intriguing and generally in directions that one might expect. They strengthen the case against studies based on convenience

samples or at least emphasize the importance of eventually conducting a study based on a sound probability sample.

Unfortunately, the results of the comparisons they present are susceptible to two threats to validity, nonresponse and unequal chances of selection. The authors present detail about the cooperation interviewers were able to elicit from potential subjects, with a cooperation rate that may be as low as 61%. This cooperation rate does not appear to be a response rate that also takes into account the proportion of women who could not be approached on the street to be screened. The cooperation rate they achieved is quite remarkable, but if inability to approach some subjects is taken into account as well, concerns arise as to whether the response rate could adversely affect the comparisons between the overall sample and each of the convenience subsamples selected from it. One could construct plausible examples in which differential response rates across convenience samples as well as differences between responding and nonresponding individuals could lead to the observed results.

The second threat concerns an issue that is sometimes referred to as "time biased sampling." For the "ever arrested" and "ever convicted" convenience samples, individuals who were incarcerated for longer periods had lower chances of being selected into the probability sample, simply because they may have been in jail or prison at the time of the survey. It is not clear that this is a substantial threat to the validity of the comparisons, but at a minimum, a weight might be constructed to compensate for length of stays in jail or prison for these two convenience samples. The data may not have been collected that would permit such a weight to be constructed, but the authors may want to consider whether there is a threat to the validity of the comparisons represented by this time bias.

Bray et al. present evidence about whether limiting the survey population to the household population causes any substantial bias in estimates. This issue is, of course, much more critical for drug abuse than many other topics. The size of the differences that they report can be interpreted differently depending on one's perspective. Bray et al. argue that the addition of the institutional and homeless surveys did not change the prevalence or population count estimates obtained from the household survey substantially, except for two types of drugs. On the other hand, those interested in formulating policy for combating drug abuse for those types of drugs will find the percentage increase in prevalence or population estimates troubling. For example, in their Table 2, the household population estimates across the illicit drug use categories are increased from 1.6% for inhalants to 31.5% for crack cocaine. The relative increases for crack cocaine and heroin are substantial and bound to be of concern for some in policy areas. One must keep in mind, however, that the relative standard errors for prevalence and population estimates shown in their Table 2 are the largest for these two groups. The relative increases for these two categories are themselves subject to significant sampling error as well. One must also keep in mind that these large relative increases in estimates are only obtained at a substantial added cost to data collection. Bray et al. did not say anything about the increase in cost of data collection to conduct multiple surveys routinely.

There is one minor technical matter that is not clear in the presentation on the weights. It appears that the multiplicity weight for persons who were identified as being in two different frames was adjusted by dividing by two. This is sensible if the probabilities of selection applied to the two frames are equal or nearly so. However, both the institutional and homeless populations were sampled at very high rates relative to the household population. The multiplicity weighting should be based on the inverse of the joint probability of selection, which will result in multiplicity weights that are dominated by the smaller probability of selection. It is unlikely that the effect of such a correction would have any impact on results presented in the paper.

Hall presents one of the most complicated multiple frame sampling operations that might be encountered in practice. It is useful to know that this many frames can be handled successfully in practice. Readers might find it of interest to know the extent to which the use of these multiple frames actually led to improved cost efficiency. A useful methodological extension of the present work would be a comparison of costs under the multiple frame design to costs under the assumptions of a single frame screening survey.

Finally, the use of unequal probabilities of selection can lead to inefficient estimates in survey results. Unequal probabilities of selection are often employed for a specific purpose, such as to increase the numbers of smaller subgroups' members in the sample. Comparison of subgroups becomes analytically more efficient. Compensating weights are required, though, when subgroups are combined for other analyses. The weights, as Hall observes, can be quite variable when groups that were sampled at the same rate are combined. The weighting contributes to an increase in variance for survey estimates that can be substantial. There are many examples of the adverse impact of this type of oversampling and compensatory weighting on estimates that combine across under- and oversampled groups, including the National Health and Nutrition Examination Survey (NHANES). Survey sample design must often balance competing objectives that can have negative effects on each other if the design is optimized for one instead of the other objective.

References

Hartley, H. O. (1962). Multiple frame sampling. American Statistical Association Proceedings of the Social Statistics Section 1962, 203–206.

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. Journal of the Royal Statistical Society, Series A, 149(Part 1), 65–82.

Discussion Themes From Session 3

Michael Hilton, Rapporteur, and Lorraine Midanik, Chair

As noted in Sudman's discussion paper, although this session is labeled "Sampling and Cooperation," these really are two aspects of the problem of design and cost and quality. The discussion following presentation of the papers reflected this observation. Three themes seemed to emerge from this discussion: considerations of frame definition, how the selection of a frame affects data quality, and the problems of access to sample when it is obtained from lists.

Considerations of Frame Definition

The most common observation was that selection of the frame needs to be made in the context of the degree of precision that is required, the amount of data available in the frame about the respondents (particularly in a list frame), and costs/efficiency. Since the topic of costs pervaded all the discussion, it will be considered separately. The Bray, Wheeless, and Kroutil paper illustrates a situation in which population estimates of drug use are required; hence, probability methods were necessary. In other instances, illustrated by the Berry, Duan, and Kanouse and O'Brien, Murray, Rahimian, and Wiebel papers, information about illicit or socially undesirable behaviors is needed, and the frames that are used have to allow some access to populations that would report such behaviors.

It was also noted that when issues such as respondent availability control the selection of the sampling frame, then there is the risk of overselection of subjects who more frequently exhibit the behavior or are more accessible at the sampling points. Hence, estimates must adjust for these factors whenever possible. Related to that point is the issue of who is actually eligible. Are individuals who for some reason are temporarily in the frame eligible in the same manner as those who are "residents"?

The Influence of Frame Selection on Data Quality

The major quality issue related to the selection of the sampling frame is the issue of obtaining access to respondents. While concerns about efficiency often drive the selection of lists, most lists are biased, and samples selected from lists frequently present problems of lack of cooperation due to inability to contact respondents. For example, Rodgers noted that one-third of the nonresponse in the Medicare list portion of his study resulted from failure to contact potential respondents. On the other hand, a recent study conducted by Westat used a list sample and obtained a very high cooperation rate because the interviewers had a name when they attempted the interview.

Another issue that arose in the discussion of the effect of frame selection on data quality was the effects of correlation between key respondent attributes and response rate. It was observed that the nonresponse rates in the Berry et al. and O'Brien et al. studies appear high, but it has been observed in the seroprevalence study that there is a high correlation between seroprevalence and nonresponse. Hence, analysis of nonresponse bias in studies such as these is a critical issue. One solution suggested was to increase incentives to improve response rates and then analyze the differences between the easy-to-get and hard-to-get respondents. In the discussion, it was noted that Laumann, Gagnon, Michael, and Michaels (1994) did not find correlations between nonresponse and the behaviors being studied, which were found in the seroprevalence studies. In the O'Brien et al. study, higher-income individuals were more difficult to locate than lower-income respondents. Overall, there was no consistent pattern, and it is also noted that incentive negotiations in street studies may be risky for the interviewers' safety.

Problems of Access to Sample

For sampling frames that involve institutionalized or other populations that are not directly accessible, access can be a considerable problem. This issue is raised in Session 1, "Measuring Medical Care and Health Status," in the paper by Harris, Tierney, and Weinberger, and it is a significant issue in this session. The most problematic case is that discussed by Park and Burt, in which the sampling frame was a list of hospital patients derived from a national sample of hospitals. In this instance, accessing the sample required institutional permission and then consent of the respondent. In most instances, the respondent's consent had to be obtained prospectively, which means before the

Michael Hilton is at the National Institute on Alcohol Abuse and Alcoholism in Bethesda, Maryland. Lorraine Midanik is at the School of Social Welfare at the University of California, Berkeley.

individual could be contacted for interview. Thus, the initial approach to the respondent was outside the control of the interviewers. Regardless of the strategy used, the results are disappointing.

Several observations were made about this problem. First, it is necessary to devote significant time and resources to establishing relationships with key gatekeepers to obtain access to such institutional data. In some instances, a federal agency can facilitate access to such records. Several instances were cited in which this approach was effective. However, in most instances, there was general agreement that a local contact at each institution is also an important factor because of the need for local sponsorship by hospital institutional review boards. An additional suggestion was to sample fewer institutions in these circumstances to allow for more effective allocation of resources.

Another strategy worth considering is to work through networks of managed care companies or networks of primary care providers that routinely collect follow-up data and through which access to records may be acquired through a central part of the company. It is, of course, uncertain what biases such strategies would introduce to the overall study design. Certainly, the uninsured would not be represented in samples such as these, although with the coming of managed care to Medicaid, this issue might be addressed. However, access to Medicaid data would still require negotiation with each state in the sample and would probably require some clearance at the institutional level and perhaps at the patient level.

For some hard-to-reach groups, such as the homeless or gays and bisexuals, local advocacy groups may enhance the likelihood of completing interviews. However, sponsorship carries its own set of problems, particularly if there is more than one group claiming sponsorship.

Similar difficulties have been encountered in surveys of professionals, in which competing specialty organizations claim to speak for a professional group.

Issues of Cost

As noted in the opening paragraph of this session summary, sampling design is always a trade-off between cost and quality. Several of the investigators who presented papers in this session noted that particularly where the populations of interest are hard to locate or present specific challenges to obtaining cooperation, the selection of sampling frames may ultimately be a matter of cost versus quality, and some error in estimation may be inevitable. A missing feature in most of the papers presented in this session was data on cost. This would have been a helpful addition to the discussion.

One way of addressing costs is to use several frames and then estimate the extent of overlap. The paper by Bray et al. illustrates this strategy. There was some discussion about the procedure for weighting such samples. James Chromy described the method that was used in Bray et al.'s study, which was to divide each sample weight by 2 and sum the resulting quotients. This method will provide the correct weight for an individual who might be picked up in more than one frame.

It was noted that a major cost factor in list samples is the cost of obtaining access to respondents. Lists may be out-of-date, special consent procedures may be required (as illustrated by the Harris et al. and Park and Burt papers), or there may be a need for tracking respondents who have changed address since the list was compiled. In each instance, the final decision is how much to spend to eliminate bias in the results.

Themes to Be Pursued in Future Research

1. Probably the most important unanswered question that emerged from the discussion on this topic is the question of cost. Clearly, more studies that employ multiple frames are needed to provide data on the cost-quality trade-offs. This should clearly be a topic for discussion at the next conference.

2. A related question has to do with the use of lists. When are they useful? One paper in this session (Hall) indicates that when sampling rates are high, Medicaid lists are more efficient than random sampling. However, both Hall and Rodgers note that oversampling of minorities or other groups from lists may not be efficient because minorities are often poorly identified. Rodgers also notes in his paper that one-third of the nonresponse from the list sample was due to failure to locate the respondent.

3. The issue of screening techniques for rare populations has been discussed at previous Conferences on Health Survey Research Methods. It appears that the trade-offs, especially the cost-efficiency trade-offs, would be improved by further discussion.

4. Matching overlapping frames, especially those that include different modes, is becoming an issue as efforts are made to find more efficient ways to collect data. The issue of duplication requires more discussion. Methodological issues remain unaddressed.

5. Finally, there was general agreement that more needs to said about gatekeeping in list and institutional samples. Strategies for overcoming resistance, gaining access to subjects, and, of course, the costs in professional time and reimbursement are areas that appear to require further consideration.

Reference

Laumann, E. O., Gagnon, J. H., Michael, R. T., & Michaels, S. (1994). The social organization of sexuality: Sexual practices in the United States. Chicago: University of Chicago Press.

Special Populations and Sensitive Issues

The papers in this session focus on measurement error due to characteristics of the sampled populations or the topics being addressed in the surveys in question. Three of the papers in this section (Horm, Cynamon, & Thornberry; Duffer, Lessler, Weeks, & Mosher; and Turner, Ku, Sonenstein, & Pleck) describe approaches to measurement that seek to minimize error due to respondent editing by enhancing the respondents' sense of privacy or security as they respond to questions about the highly sensitive topics of sexual behavior and substance abuse. Two of the remaining papers (B. Kahana, Kercher, E. Kahana, Namazi, & Stange and Hendershot, Rogers, Thornberry, Miller, & Turner) address respondent characteristics that create barriers to response validity and consider ways of approaching these issues technologically or by other means. Finally, the paper by Shepherd, Hill, Bristor, and Montalvan describes the impact on response validity as mode of interview shifts from paper-and-pencil interviews (PAPI) to an audio computer-assisted self-interviewing (ACASI) format. All but the paper by B. Kahana et al. consider the issues of measurement error in the context of the ACASI format, and taken together, the session provides some insight into the potential effects and limitations of that mode of interviewing.

The Influence of Parental Presence on the Reporting of Sensitive Behaviors by Youth

John Horm, Marcie Cynamon, and Owen Thornberry

Introduction

A continuing challenge facing survey researchers is administering questions on respondent-perceived sensitive behaviors in a manner that minimizes response bias. Negative personal behaviors may be frequently underreported or inaccurately reported because of fear of disclosure, including inadvertent disclosure to household members. Telephone and self-administered interviews, which are often used as solutions to this problem, have serious shortcomings (Aquilino, 1994; Schwarz, Strack, Hippler, & Bishop, 1991; Johnson, Hougland, & Clayton, 1989; Gfroerer, 1985; Groves, 1990). Telephone coverage is incomplete, and respondents may not be comfortable answering questions about sensitive behaviors when there is a concern that another household member may overhear or listen in on the conversation. Further, the absence of a social relationship between the interviewer and the respondent during telephone interviews is believed to reduce the respondent's willingness to reveal personal information. Although generally producing higher reported levels of sensitive behaviors, a self-administered document has the major shortfall of requiring the respondent to have a sophistication and skill in both reading and filling out complex questionnaires. The recently developed computer-assisted self-interviewing (CASI) technique, with audio, overcomes some of the problems of literacy and privacy but is not always feasible due to budgetary constraints (O'Reilly, Hubbard, Lessler, Biemer, & Turner, 1994; Mosher, Pratt, & Duffer, 1994). This paper discusses an alternative mode that is inexpensive and offers privacy for household surveys on sensitive topics.

Methods

In 1992, the National Center for Health Statistics (NCHS) fielded the Youth Risk Behavior Survey (YRBS) as a follow-up to the National Health Interview Survey (NHIS). The YRBS is one component of a surveillance system conducted by the Centers for Disease Control and Prevention consisting of continuous national, state, and local school-based surveys and a one-time household-based survey (YRBS) using a national probability sample. This report discusses data from the household survey, in which approximately 11,000 youths aged 12 through 21 years were asked about risk behaviors most commonly engaged in by this age group. The behaviors included use of alcohol, drugs, and tobacco; physical inactivity; dietary practices; sexual behavior that could lead to pregnancy or increased risk of sexually transmitted diseases; and transportation safety practices (seat belt use, driving while intoxicated, helmet use). One in-school and up to two out-of-school youths were selected from each sampled household.

Because of concerns about accurate and complete reporting for the sensitive subjects included on the 1992 YRBS, the use of innovative methods to minimize risk of disclosure was explored. Preliminary testing in the NCHS questionnaire design research laboratory and focus groups confirmed that this age group would not be comfortable answering sensitive questions when asked at home using conventional personal or telephone methods and that they would rather lie than refuse to answer questions or participate in the survey. In order to ensure the privacy of the respondents so that participation and honesty could be maximized, the interviews were administered using a portable audiocassette tape player with headphones (PACTAPH), with the questions and answer categories recorded on tape. The answers were entered by the youths on an answer sheet that in no way revealed the topics on the tape. If parents asked to see the questionnaire, the interviewer showed them a booklet with the questions in a different order from that on the tape. This methodology proved to be highly successful and has been described elsewhere (Camburn, Cynamon, & Harel, 1991).

Preliminary testing indicated that the interviewers, whose usual role in the task was greatly diminished, felt their expertise was being underutilized. Children were uncomfortable when the interviewer did not appear to be occupied, and interviewers felt awkward and useless with nothing to do. This presented an opportunity to develop a questionnaire for the interviewer to fill out that detailed the dynamics in the household while the youth was listening to the tape. It provided information about the circumstances surrounding the interview as well as giving interviewers

The authors are with the Division of Health Interview Statistics, National Center for Health Statistics, Hyattsville, Maryland.

something to do so the youths did not feel as though they were being watched.

Interviewers completed an observation questionnaire for each youth in the survey. Questions were asked about interference from and proximity of parents and other persons present, the level of cooperation from the selected youth, household distractions, and the youth's ability to use the tape player and headphones.

The focus of this report is an examination of the influence of parental presence on response patterns for YRBS respondents between the ages of 12 and 17 years. The observation form contains several questions pertaining to the location of the parent(s) at various times during the interview. The one selected for this report describes where the parent was when the child began listening to the tape: close enough to see the answers, in the same area but not close enough to see the answers, not in the same area but still able to see the child, at home but unable to see the child, or not at home. If it appeared that the parent might pose a problem, the interviewer was instructed to suggest that the child take the tape player to a quieter room so they could concentrate better or to actively engage the parent in conversation as a distraction from the child. In this latter event, the interviewer filled out the observation form upon leaving the household.

Broad assurances of confidentiality and nondisclosure, while important in gaining the cooperation of respondents, may not have the same influence on young persons whose biggest perceived threat is the consequences of disclosure to parents. If the PACTAPH administration has the desired effect, levels of reporting of sensitive behaviors should be similar regardless of the proximity of the parent to the respondent during the interview. In addition, reports of risk behaviors should be higher (assuming higher reporting means more honest reporting) than in those surveys using modes that afford less privacy.

Findings

For eligible respondents screened in from the NHIS sample, the response rate to the 1992 YRBS household survey was 77.2%. This analysis is restricted to 6,300 respondents, 12 through 17 years of age. Two-year age groups were used to ensure adequate sample sizes and to remain sensitive to rapid changes in behavior throughout the age range. Table 1 shows that when a parent was at home during the interview, the youngest children were more likely to have them close by. For children 12 and 13 years of age, 10% of the parents were situated where they could see the child's answers, and an additional 32% were in view of the child. Parental proximity steadily declined as the age of the youths increased. The older groups (14 and 15 years and 16 and 17 years) were given progressively more privacy in the interview setting, with 52% and 62% of the parents either out of the room or not home.

Table 1. Number and percentage of respondents by location of parent and age of child: 1992 YRBS

	Ages 12–13		Ages 14–15		Ages 16–17	
	No.	%	No.	%	No.	%
Close	216	9.8	142	6.7	86	4.4
In area	697	31.8	567	26.6	425	21.6
Not in area	339	15.4	309	14.5	227	11.6
Can't see youth	693	31.6	753	35.3	668	34.0
Not home	250	11.4	364	17.0	559	28.4
Total	2,195	100.0	2,135	100.0	1,965	100.0

To calibrate parental influence, distributions for questions that should pose no threat to the respondents were examined. Physical activity, having eaten fruit yesterday, and receiving AIDS education in school show no distinct patterns of variation based on location of the parents, regardless of the child's age (see Table 2).

If the PACTAPH is effective in allaying fear of disclosure, the distribution for sensitive behaviors should be similar to that for nonsensitive behaviors. Table 2 shows the distributions for selected sensitive behaviors: having smoked in the past 30 days; ever having tried marijuana or alcohol; ever having had sexual intercourse; and, for the oldest group, having driven while intoxicated. Although most levels of reporting appeared to increase slightly when parents were out of the room but still at home, there were a few significant differences within age groups. Reporting on other behaviors not included in this report had similar patterns. This could represent evidence that the PACTAPH technique removed the influence of the presence of parents on reporting. On the other hand, the PACTAPH technique may have had no additional effect beyond gains from the general assurances of confidentiality and nondisclosure.

Comparisons were also made of the levels of reporting sensitive behaviors when the parents were home and when they were not home. Table 2 shows higher prevalence rates of cigarette smoking in the past 30 days, ever having drunk alcohol, ever having tried marijuana, and having had sexual intercourse for the children in the two older age groups whose parents were not home. This indicates that parental presence may have an overall negative impact on the reporting of sensitive behaviors but still does not provide evidence that the PACTAPH mode improves reporting levels. These findings may also reflect the possibility that children whose parents are not at home at the time of the interview have the least supervision and might therefore be more likely to engage in risk behaviors. The demographics for this group of parents show that they have attained higher income and educational levels than those parents who are present during the interview. Not only do these children have the opportunity to engage in these behaviors, but they also are more likely to have the means.

Table 2. Percentage of youths with selected characteristics by age and location of parents: 1992 YRBS

	Close	In area	Not in area	Can't see youth	Not home
Ages 12–13					
Family income < $20,000	33.7	34.8	28.6	25.0	30.3
Highest education of parent < 12 years	15.8	15.7	12.6	11.4	12.6
Ate fruit yesterday	52.7	64.4	66.6	63.0	59.9
Taught about AIDS in school	86.9	83.0	88.1	86.8	83.8
Exercised in past week	15.5	12.3	15.7	10.7	16.5
Smoked cigarettes in past 30 days	5.4	7.6	8.8	7.8	8.8
Ever tried marijuana	0.4	3.9	4.1	2.8	4.4
Ever drank alcohol	21.6	26.9	34.5	27.1	30.4
Ever had sexual intercourse	na	na	na	na	na
Drove a car after drinking	na	na	na	na	na
Ages 14–15					
Family income < $20,000	36.6	31.1	22.8	26.0	25.1
Highest education of parent < 12 years	22.3	14.9	11.5	15.0	10.0
Ate fruit yesterday	61.7	55.8	53.9	59.9	60.4
Taught about AIDS in school	94.4	91.7	92.5	90.7	91.0*
Exercised in past week	13.1	17.5	12.3	12.8	13.7
Smoked cigarettes in past 30 days	22.3	19.4	20.1	19.4	26.6
Ever tried marijuana	11.5	11.0	13.5	9.8	17.0
Ever drank alcohol	57.0	53.3	54.8	56.0	63.6
Ever had sexual intercourse	24.7	27.3	29.4	31.1	31.6*
Drove a car after drinking	na	na	na	na	na
Ages 16–17					
Family income < $20,000	35.9	32.4	24.6	25.4	29.6
Highest education of parent < 12 years	27.6	15.4	13.6	10.2	15.4
Ate fruit yesterday	47.5	51.7	52.8	54.3	51.6
Taught about AIDS in school	92.1	93.0	93.8	90.3	91.3
Exercised in past week	18.3	29.3	17.4	21.5	21.6
Smoked cigarettes in past 30 days	28.7	27.7	32.4	27.1	35.2
Ever tried marijuana	21.7	26.4	29.9	25.4	35.0
Ever drank alcohol	64.6	73.8	79.1	71.3	80.6
Ever had sexual intercourse	47.6	56.5	57.4	55.5	59.8
Drove a car after drinking	0.8	6.2	11.8	10.6	10.2

NOTE: "na" indicates that questions were not asked for youths of that age. All estimates are weighted.

*Indicates that the trends across location of parents is statistically significant at the $p < .05$ level.

There were no YRBS household interviews conducted using another mode with which to compare the PACTAPH methodology. However, relevant data were available from the 1992 National Household Survey on Drug Abuse (NHSDA) conducted by the Substance Abuse and Mental Health Services Administration (Office of Applied Statistics, 1994). The NHSDA and YRBS have similar sampling methodologies, both are weighted to produce national estimates, and there are several questions that are comparable. The exact questions administered by the YRBS and the NHSDA are in the Appendix. The NHSDA used the promise of anonymity to encourage participation, and except for responses to the cigarette questions, which were recorded by the interviewer, sensitive questions were addressed using a self-administered, paper-and-pencil format in which the interviewer read the questions and the respondent marked an answer sheet. Like the YRBS, information on the location of the parent during the interview was collected in the NHSDA.

Except for marijuana use in the youngest age group, there is significantly higher reporting on the YRBS than the NHSDA for all behaviors for all age groups (see Table 3). The differences are greatest for reports of cigarette smoking, which may reflect the format requiring a verbal response recorded by the interviewer. Detail on these behaviors by the proximity of the parents during the interview is provided in Table 4. Both surveys demonstrate consistently higher reporting when parents are not present (parent at home but unable to see youth or not at home). However, the relative differences are much greater in the NHSDA than in the YRBS. This further supports the premise that the primary concern to youths is the immediate threat of disclosure to parents rather than anonymity or confidentiality.

Conclusions

This paper provides evidence consistent with findings of previous research of a relationship between degree of privacy and the frequency of positive responses to sensitive questions by youths. The evidence suggests that the PACTAPH interviewing technique provides a greater degree of privacy from parental disclosure than do interviewer- and/or self-administered approaches. Youths clearly are concerned about parents or other household members learning of their responses and require the most secure setting and interview mode to respond honestly to sensitive questions. The use of the PACTAPH approach appears to provide the level of privacy necessary for maximum disclosure of sensitive behaviors. The findings of this research suggest that broad promises of anonymity or confidentiality are perhaps less important to honest reporting by youths than are assurances of privacy from the immediate threat of disclosure to parents.

Table 3. Percentage of youths reporting selected risk behaviors by age: 1992 YRBS and 1992 NHSDA

Risk behavior	Ages 12–13	Ages 14–15	Ages 16–17
Smoked cigarettes in past 30 days			
YRBS	7.8	21.0	30.1
NHSDA	1.9	9.4	18.1
Ever tried marijuana			
YRBS	3.3	12.0	28.6
NHSDA	2.0	9.9	20.9
Ever drank alcohol			
YRBS	28.0	56.6	75.0
NHSDA	15.7	39.2	65.6

NOTE: All differences in estimates between the YRBS and the NHSDA are statistically significant at the $p < .05$ level as determined by a Z test.

The PACTAPH interviewing technique offers a relatively inexpensive approach for measuring participation in negative personal behaviors in the household setting. The major disadvantage of the approach is that the use of complex skip patterns is not possible. An alternative technique with considerable promise is audio computer-assisted self-interviewing (ACASI). The ACASI approach offers the potential of the same level of household privacy as the PACTAPH, and in addition, the administration of complex questionnaires is possible. The primary disadvantages relate to costs and the potential for a lengthy developmental period.

Appendix

The YRBS and NHSDA used different forms of questions on sensitive behaviors. The response categories from the two surveys are recoded for comparability as indicated below.

Smoked Cigarettes in Past 30 Days

YRBS During the past 30 days, on how many days did you smoke cigarettes?
- 0 days.
- 1+ days.

NHSDA When was the most recent time you smoked a cigarette?
- Never—skipped from earlier question.
- > 30 days ago.
- ≤ 30 days.

Ever Tried Marijuana

YRBS How old were you when you tried marijuana for the first time?
- Age given.
- Never done this.

Table 4. Percentage of youths reporting selected risk behaviors by age and location of parents: 1992 YRBS and 1992 NHSDA

Risk behavior	Ages 12–13		Ages 14–15		Ages 16–17	
	Parent present	Parent not present	Parent present	Parent not present	Parent present	Parent not present
Smoked cigarettes in past 30 days						
YRBS	7.5	8.2 ns	19.8	22.0 ns	29.4	30.6 ns
NHSDA	1.1	2.6	6.4	11.2	16.7	17.9 ns
Ever tried marijuana						
YRBS	3.3	3.3 ns	11.6	12.4 ns	27.1	29.5 ns
NHSDA	1.4	1.9* ns	4.7	12.3*	17.1	21.8 ns
Ever drank alcohol						
YRBS	27.8	28.2 ns	54.5	58.5 ns	74.6	75.2 ns
NHSDA	12.3	16.2 ns	32.3	42.7	57.6	67.6

NOTE: "Parent present" means youth was in view of parent during at least part of the interview. "Parent not present" means parent either was home but could not see youth or parent was not home. Differences in estimates are statistically significant at the $p < .05$ level unless otherwise indicated; "ns" means not statistically significant at the $p = .05$ level.

*Differences between YRBS and the NHSDA are not statistically significant.

NHSDA About how old were you the first time you actually used marijuana or hash, even once?
Age given.
Never used.

Ever Drank Alcohol

YRBS How old were you when you had your first drink of alcohol other than a few sips?
Age given.
Never done this.

NHSDA About how old were you the first time you had a glass of beer or wine or a drink of liquor, such as whiskey, gin, scotch, etc.?
Age given.
Never.

References

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use. Public Opinion Quarterly, 58, 210–240.

Camburn, D., Cynamon, M., & Harel, Y. (1991). The use of audio tapes and written questionnaires to ask sensitive questions during household interviews. Presented at the National Field Directors/Field Technologies Conference, San Diego, CA.

Gfroerer, J. (1985). Influence of privacy on self-reported drug use by youths. In B. Rouse, N. Kozel, & L. Richards (Eds.), Self-report methods of estimating drug use (NIDA Research Monograph No. 57).

Groves, R. (1990). Theories and methods of telephone surveys. Annual Review of Sociology, 16, 221–240.

Johnson, T. P., Hougland, J. G., & Clayton., R. R. (1989). Obtaining reports of sensitive behavior: A comparison of substance use reports from telephone and face-to-face interviews. Social Science Quarterly, 70, 174–183.

Mosher, W. D., Pratt, W. F., & Duffer, A. P. Jr. (1994). CAPI, event histories, and incentives in the NSFG Cycle 5 pretest. Presented at the meeting of the American Statistical Association, Toronto, Canada.

Office of Applied Statistics, Substance Abuse and Mental Health Services Administration. (1994). National Household Survey on Drug Abuse: Main findings 1992 (DHHS Publication No. [SMA]94-3012). Rockville, MD: Public Health Service.

O'Reilly, J. M., Hubbard, M. L., Lessler, J. T., Biemer, P. P., & Turner, C. F. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. Journal of Official Statistics, 10, 197–214.

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. Applied Cognitive Psychology, 5, 193–212.

Impact of Incentives and Interviewing Modes: Results From the National Survey of Family Growth Cycle V Pretest

Allen P. Duffer, Judith T. Lessler, Michael F. Weeks, and William D. Mosher

Study Background

The National Survey of Family Growth (NSFG) is administered by the National Center for Health Statistics (NCHS) to provide national estimates of behaviors related to the birthrate as well as those related to maternal and infant health. The survey collects data on sexual activity, marriage, cohabitation, contraception, sterilization, infertility, breast-feeding, miscarriages, stillbirths, and induced abortions. Important independent variables such as work histories, education histories, and living arrangements are also collected.

There have been four previous cycles of the survey, conducted approximately every 5 years, beginning in 1973. In October of 1992, the Research Triangle Institute began work on Cycle V of the survey. A pretest was conducted from October 11 to December 20, 1993, in six sites located in New York, Texas, and North Carolina. Eight hundred and three women between the ages of 15 and 44 were selected for the pretest from the rosters of households that had participated in the 1991 National Health Interview Survey (NHIS), also conducted by NCHS.

Pretest Design

In Cycle V, significant changes were made in the study. The pretest was designed to test a number of innovations, including the use of the following: (a) computer-assisted personal interviewing (CAPI), (b) audio computer-assisted self-interviewing (ACASI), (c) neutral sites in which to interview sampled women, and (d) incentives.

To test the effect of these innovations on response rates, the reporting of sensitive behaviors, and the cost of the data collection effort, sampled women were assigned to one of the following experimental treatment groups: (a) in-home CAPI administration, no incentive; (b) in-home CAPI administration, $20 incentive; (c) in-home CAPI plus an ACASI component, no incentive; (d) in-home CAPI plus an ACASI component, $20 incentive; and (e) CAPI administration (no ACASI) at a neutral site, $40 incentive.

Allen P. Duffer, Judith T. Lessler, and Michael F. Weeks are with the Research Triangle Institute, Research Triangle Park, North Carolina. William D. Mosher is with the National Center for Health Statistics, Hyattsville, Maryland.

Those women receiving the in-home CAPI treatment were administered the entire questionnaire at their homes by experienced field interviewers. Those who received the ACASI component as part of the CAPI interview were asked near the end of the interview to listen to a recorded version of questions and answers played by the computer over headphones. The respondent then keyed her answers into the computer; the computer recorded the response and played the next appropriate question. The ACASI component included a question on whether or not the woman had ever had an abortion, a question that had been previously asked by the interviewer in Section B of the questionnaire. The respondent was asked to reconsider her previous response, given the more private and confidential setting created by the use of ACASI.

The women assigned to the neutral site treatment were asked to meet an interviewer in a predetermined location outside the respondent's home and to complete the interview. The neutral sites were situated in areas geographically convenient for the sampled women and the interviewers. The types of locations varied by primary sampling unit (PSU) and included meeting rooms in libraries, hotels, universities, and office buildings. During the interview, these rooms were occupied by only the interviewer and the respondent.

For the purposes of testing the effects of incentive payments, there were three groups: in-home, no incentive; in-home, $20 incentive; and neutral site, $40 incentive. Originally, we had planned to use only two conditions—in-home, no incentive and neutral site, $40 incentive—and to randomly assign the treatments within PSU. We were concerned, however, that if the neutral site group had a better response rate or improved data quality, we would not know whether the improvement was due to the nonhome site or the $40 incentive. We were also concerned that the $40 incentive plus the cost of setting up and renting neutral sites for 6 months of fieldwork in 150 to 200 PSUs across the country might be too expensive, not feasible, or both. In order to see whether a more modest incentive would be effective, we introduced a third group in the pretest: those who would receive a $20 incentive for an in-home interview. We thought that this would help us determine whether the incentive or the neutral site produced any effects we might note and that it would help to control costs. Consequently, within PSUs, sampled women were randomly

assigned to either a neutral site or in-home treatment. Then, in three of the six PSUs (one in each pretest state), we tested the $20 incentive; in the other three PSUs, the women assigned to the in-home treatment were not offered an incentive.

Response Rate Results

Table 1 summarizes the response rates by characteristics of the respondent and by incentive; these rates are for those sampled women who were located and had an opportunity to be affected by the incentive treatment. Overall, the two groups of women who were offered an incentive had a response rate about 7% higher than those who were not offered an incentive. The pattern of differences between those offered and not offered an incentive varied by characteristic of the sampled women. Of particular note is the effect the payment of incentives had on the participation of blacks, low-income women, and generally reluctant respondents (those who did not report their income in the NHIS). In this case, the incentive payment brought into the survey groups of women who have higher abortion rates and appear to have more sexual partners than others (Henshaw & Silverman, 1988; Leigh, Temple, & Trocki, 1993).

Interviewing Mode Effects

We also looked at the effect of the mode of interview on the reporting of abortions and other sensitive behaviors. In the following tables, we present the differences in reporting by site (in-home vs. neutral site), by incentive ($0 vs. $20 vs. $40), and by mode (in-home CAPI vs. in-home ACASI vs. neutral site CAPI).

Table 2 compares the reporting of sensitive behaviors by women by site of interview and incentive. In general, there is little difference in the reporting under the interviewing conditions. There are, however, three notable exceptions: (a) reported number of lifetime sex partners, (b) the lifetime incidence of being forced by a man to have sex, and (c) the reported lifetime incidence of abortion.

Those women who received a $20 incentive had somewhat higher reports than those who received none. It should be noted that the increase in reports is greater in the neutral site interviews. While these results suggest that payment of an incentive increases the reporting of sensitive behaviors, those reports are further enhanced by conducting the interview in a private setting.

Two primary strategies were used in the pretest to improve abortion reporting and to investigate the quality of these reports. First, the ACASI component of the interview explained to women that prior surveys had revealed that women were reluctant to talk about abortions and asked them to answer the abortion questions again now that they could do so in complete privacy. Second, the interviews at a neutral site were thought to provide additional privacy for women who might be reluctant to talk about sensitive issues in their homes. Table 3 shows that in both the in-home and ACASI treatments, the percentage of women reporting abortions was greater for those who received a $20 incen-

Table 1. Response rates of sampled women who were eligible for the pretest by characteristics

Eligible for pretest	Overall	Hispanic	Black, non-Hispanic	Nonblack, non-Hispanic	< 18	18+	Income < $20,000	Income $20,000+	Income unknown
All treatments									
Located, eligible	645	74	165	406	41	604	139	414	92
Located eligibles responding	500	59	135	306	31	469	111	329	60
Proportion responding	0.78	0.80	0.82	0.75	0.76	0.78	0.80	0.79	0.65
In-home $0									
Located, eligible	269	20	56	193	14	252	42	185	42
Located eligibles responding	196	13	42	141	11	185	31	142	23
Proportion responding	0.73	0.65	0.75	0.73	0.79	0.73	0.74	0.77	0.55
In-home $20									
Located, eligible	188	32	68	88	12	176	56	105	27
Located eligibles responding	153	29	58	66	12	141	48	85	20
Proportion responding	0.81	0.91	0.85	0.75	1.00	0.80	0.86	0.81	0.74
Neutral site $40									
Located, eligible	188	22	41	125	12	176	41	124	23
Located eligibles responding	151	17	35	99	8	143	32	102	17
Proportion responding	0.80	0.77	0.85	0.79	0.67	0.81	0.78	0.82	0.74

Table 2. Effect of incentives and site of interview on reporting of sensitive behaviors

Characteristic	In-home $0 Average	In-home $0 n	In-home $20 Average	In-home $20 n	Neutral site $40 Average	Neutral site $40 n
Sex partners in last 12 months	1.2	182	1.1	142	1.2	134
Sex partners since January 1, 1989	1.8	182	1.8	142	1.9	134
Sex partners in lifetime	5.1	180	5.4	142	6.6	133
Sex partners before marriage	4.4	88	5.3	68	5.2	66
No. cigarettes/day (current)	14.7	55	14.6	49	13.1	35
No. cigarettes/day (past)	13.3	24	12.4	27	13.6	27
Age at first sex	17.7	183	17.3	143	17.4	134
	Proportion (n = 194)		Proportion (n = 152)		Proportion (n = 147)	
Grades in high school C or less	0.06		0.05		0.06	
Parents not living together at R's birth	0.04		0.06		0.05	
Smoked at least 100 cigarettes in life	0.59		0.50		0.58	
Ever forced by a man to have sex	0.15		0.18		0.22	
Ever had an abortion	0.17		0.25		0.29	

Table 3. Distribution of the number of abortions reported in Section B by treatment and incentive

		In-home CAPI							In-home ACASI						Neutral site CAPI	
			$0		$20		Total		$0		$20		Total		$40/Total	
	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	No.	%
No. abortions																
0	380	77.1	83	86.5	56	77.8	139	82.7	78	79.6	58	72.5	136	76.4	105	71.4
1	79	16.0	10	10.4	12	16.7	22	13.1	16	16.3	11	13.8	27	15.2	30	20.4
2	24	4.9	1	1.0	4	5.6	5	3.0	3	3.1	9	11.3	12	6.7	7	4.8
3	9	1.8	2	2.1	0	0	2	1.2	1	1.0	1	1.3	2	1.1	5	3.4
4	1	0.2	0	0	0	0	0	0	0	0	1	1.3	1	0.6	0	0
Total	493	100.0	96	100.0	72	100.0	168	100.0	98	100.0	80	100.0	178	100.0	147	100.0
1+ abortions reported in Section B			13	13.5	16	22.2	29	17.3	20	20.4	22	27.5	42	23.6	42	28.6
1+ abortions reported in ACASI									24	24.5	24	30.4	48	27.1		

tive. The table also indicates, as noted above, that the neutral site interviews resulted in increased reports of abortions. When a comparison of the percentage of abortions reported for the in-home, no incentive cases (13.5%) is made with that of the ACASI, $20 incentive cases (30.4%), it suggests that the combination of the incentive plus the private setting has a substantial impact on reporting. A significance test indicates that the difference is significant at alpha less than .01.

Results From Prior Research

Much of the research on the payment of incentives points to its effectiveness in increasing response rates (Kulka, 1992; Willimack, Petrella, Beebe, & Welk, 1992). While the research on the use of incentives in personal interview surveys is limited, the effectiveness of incentives seems to be related to the extent of the demands placed on the respondent, the requirement for cooperation over time, or

the lack of other motivating factors for the respondent (Cannell & Fowler, 1977). On the National Adult Literacy Survey, a 2,000-case field test found that payment of both $20 and $35 incentives increased response rates by 8% and 10%, respectively, over those cases receiving no incentive (Berlin et al., 1992). Also consistent with our findings, they found the incentives to be more effective with disadvantaged and minority populations.

With regard to data quality, most evidence indicates that a monetary incentive results in greater commitment on the part of respondents, resulting in better data quality (e.g., Berk, Mathiowetz, Ward, & White, 1987; Sudman & Ferber, 1974; Goetz, Tyler, & Cook, 1984; James & Bolstein, 1990). There is some contrary evidence suggesting that incentives encourage respondents to provide interviewers with answers the respondent may think are most correct or desirable (Cannell & Henson, 1974; Weiss, 1975). Our findings, which indicate more complete reporting of sensitive behaviors by those who receive an incentive, tend to support the theory that the payment of an incentive results in a greater commitment to the survey task.

Historically, the underreporting of abortions has been a significant problem on the NSFG (Jones & Forrest, 1992). Other studies, such as the National Household Survey on Drug Abuse (NHSDA), have also identified underreporting of sensitive behaviors in interviewer-administered questionnaires (Turner, Lessler, & Devore, 1992). The use of a self-administered questionnaire (SAQ) in Cycle IV of the NSFG raised the reports of abortions from 35% of an external count of abortions (based on counts from abortion providers) to 71%. The use of an SAQ on the NHSDA resulted in increased reporting the more sensitive the drug (cocaine vs. marijuana) and the more recent the use (last 30 days vs. lifetime).

Certainly, there are limits to the effectiveness of traditional SAQs. As Mangione, Hingson, & Barrett (1982) have demonstrated, SAQs do not necessarily result in more valid data because the data is often incomplete. This can be attributed to the sensitivity of the material being requested (Medina-Mora, Castro, Campillo-Serrano, & Gomez-Mont, 1981) and to comprehension problems. Illiteracy and other language problems may prevent the respondent from completing the questionnaire without assistance.

That reports of abortions in the pretest were comparable in both the ACASI and the neutral site treatments suggests that privacy is a key factor in eliciting reports of sensitive behaviors. ACASI appears to be an effective and efficient means of creating a private setting in which to ask sensitive questions. In addition, the ACASI technology has essentially eliminated concerns about illiteracy, and because it allows programmed routing and consistency checks, it should result in fewer missing and inconsistent data in a self-administered format.

We were somewhat surprised by the impact of the neutral site on increased reports of abortions. We assumed that the fact that the questionnaire was interviewer administered would be a deterrent to increased reporting. While previous studies have shown that lack of privacy has a negative effect on the reporting of drug use, particularly among 12 through 17 year olds (Turner et al., 1992), other studies have shown that privacy does not have an effect on the reporting of sensitive behaviors (Schober, Caces, Pergamit, & Branden, 1992; Tourangeau, Jobe, Pratt, & Rasinski, 1994). It is possible that some of the effect we experienced was the result of a large incentive ($40) as well as the payment of transportation and child care costs as required. In addition, the NSFG questionnaire was long and required considerable up-front work in completing a life history calendar and numerous event histories prior to the asking of any sensitive questions. This situation may have enhanced the rapport between the interviewer and respondent and resulted in greater respondent comfort in reporting sensitive behaviors. Finally, the neutral sites were completely private (no family members and no passersby), which may have also increased the respondent's comfort level.

Conclusions

The pretest was designed to test the effect of cash incentives on response rate and the effect of interview mode on data quality. Both the $20 and $40 incentive treatments significantly increased the response rate and may have also had some effect on the increased reports of abortions. Based on the pretest results, we estimate that either of these treatments would increase the response rate about 7% over the no incentive treatment. There was no significant difference in the response rate for the $20 and $40 treatments, probably because the respondent had to leave her home for the $40 incentive, increasing the response burden.

The data also indicate that the use of neutral sites and the ACASI methodology for questionnaire administration improved the reporting of sensitive behaviors. Both the ACASI and the neutral site interview modes produced a significant increase in the number of women who reported ever having had an abortion, compared with the CAPI-only mode. There was no significant difference between the ACASI and neutral site modes in either the proportion of women reporting that they had ever had an abortion or in the number of abortions reported.

There is some evidence in the data that the neutral site mode also yielded higher reports of other sensitive information related to sexual behavior. (These questions were not asked as part of the ACASI component, so no effect of ACASI on these behaviors was determined.) In a limited examination of other variables, however, we found no difference on less sensitive items, such as reported smoking behavior, dates of first intercourse, and high school grades.

References

Berk, M. L., Mathiowetz, N. A., Ward, E. P., & White, A. A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. Journal of Official Statistics, 3, 449–457.

Berlin, M., Mohadjer, L., Waksberg, J., Kolstad, A., Kirsch, I., Rock, D., & Yamamoto, K. (1992). An experiment in monetary

incentives. American Statistical Association 1992 Proceedings of the Survey Research Methods Section, 393–398.

Cannell, C. F., & Fowler, F. J. (1977). Interviewers and interviewing techniques. In National Center for Health Services Research, Advances in health survey research methods: Proceedings of a national invitational conference (NCHSR Research Proceedings Series, DHEW Publication No. [HRA] 77-3154, pp. 13–23). Rockville, MD: National Center for Health Services Research.

Cannell, C. F., & Henson, R. (1974). Incentives, motives, and response bias. Annals of Economic and Social Measurement, 3, 307–317.

Goetz, E. G., Tyler, T. R., & Cook, F. L. (1984). Promised incentives in media research: A look at data quality, sample representativeness, and response rate. Journal of Marketing Research, 21, 148–154.

Henshaw, S., & Silverman, J. (1988). The characteristics and prior contraceptive use of U.S. abortion patients. Family Planning Perspective, 20(4) 158–168.

James, J. M., & Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. Public Opinion Quarterly, 54, 346–361.

Jones, E. F., & Forrest, J. D. (1992). Underreporting of abortion in the surveys of U.S. women: 1976 to 1988. Demography, 29, 113–125.

Kulka, R. A. (1992, October). A brief review of the use of monetary incentives in federal statistical surveys. Paper prepared for the COPAFS/OMB Symposium on Providing Incentives to Survey Respondents, Cambridge, MA.

Leigh, B. C., Temple, M., & Trocki, K. (1993, October). The sexual behavior of U.S. adults: Results from a national survey. American Journal of Public Health, 83, 1400–1407.

Mangione, T. W., Hingson, R., & Barrett, J. (1982). Collecting sensitive data: A comparison of three survey strategies. Sociological Methods and Research, 10, 337–345.

Medina-Mora, M. E., Castro, S., Campillo-Serrano, C., & Gomez-Mont, F. A. (1981). Validity and reliability of a high school drug use questionnaire among Mexican students. Bulletin on Narcotics, 33, 67–75.

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

Sudman, S., & Ferber, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. Journal of Marketing Research, 11, 128–135.

Tourangeau, R., Jobe, J., Pratt, W., & Rasinski, K. (1994). Design and results of the Women's Health Study. Paper presented at the annual meeting of the American Statistical Association, Toronto, Canada.

Turner, C., Lessler, J., & Devore, J. (1992). Effects of mode of administration and wording on reporting of drug use. In C. Turner, J. Lessler, & J. Gfroerer, (Eds.), Survey measurement of drug use: Methodological studies (pp. 177–219). Rockville, MD: National Institute on Drug Abuse.

Weiss, C. H. (1975). Interviewing in evaluation research. In E. I. Struening & M. Guttentag (Eds.), Handbook of evaluation research (pp. 355–395). Beverly Hills, CA: Sage.

Willimack, D. K., Petrella, M., Beebe, T., & Welk, M. (1992, August). The use of incentives in surveys: Annotated bibliography. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Mild Cognitive Impairment and Accuracy of Survey Responses of the Old Old

Boaz Kahana, Kyle Kercher, Eva Kahana, Kevan Namazi, and Kurt Stange

Introduction

Interest in cognitive functioning in late life has for a long time been restricted to the work of cognitive psychologists concerned with basic processes of learning and thinking and clinicians focusing on cognitive deficits that interface with everyday functioning. Accordingly, there are extensive literatures on the nature of cognitive declines among the healthy aged on the one hand (Schaie, 1983) and on diagnosing and treating older persons suffering from dementia on the other (Eisdorfer, 1977; Butler, Lewis, & Sunderland, 1991). However, little attention has been paid to the problems posed by mild cognitive deficits in relation to aspects of daily functioning and particularly in relation to the ability to respond to social science health surveys.

In gerontological research, concerns have been noted over a potential increase in nonresponse rates among the old old and about greater inaccuracy in their survey responses (Lawton & Herzog, 1989; Rodgers, Herzog, & Andrews, 1988). However, studies of response effects among the old old have been relatively rare (Carsjo, Thorslund, & Warneryd, 1994). Whereas the basis for expecting differential survey responses by the old old is seldom specified, cognitive deficits seem to be the most likely cause.

Gerontological researchers and psychologists have recently directed their attention to a condition termed "mild cognitive impairment," or MCI (Powell, 1994). This term describes a broad range of conditions, including limited dementia (Schaie, 1983), mild dementia (Salthouse, 1991), age-related memory impairment (Crook & Larrabee, 1988), and benign senescent forgetfulness (Doppolt & Wallace, 1955). Although these conditions are not treated as separate diagnostic entities in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV; American Psychiatric Association, 1994), they do involve performance deficits that represent reduced ability to perform on a continuum of memory tasks used to diagnose dementia. Regardless of the etiology and classification systems utilized, the existence of mild cognitive deficits is a real and fairly prevalent phenomenon among the old old (i.e., those aged 75 or older) and to a lesser extent, among the young old (i.e., those aged 65–74). The implication of these cognitive deficits among older adults for recalling and for providing valid information relevant to social research has not been previously explored. As the old old are increasingly becoming the focus of health research, it is critical to explore whether there are methodological problems presented by mild, age-associated memory impairments. In addition, there are other cognitive losses which may influence survey responses of the old old (Salthouse, 1991).

In health surveys, there are few research-based guidelines about the use of information provided by the cognitively frail elderly. Threats to validity of data provided by elderly persons with cognitive deficits are generally handled by exclusion criteria. Persons with obvious cognitive or physical deficits often refuse participation, based on self-assessed or caregiver-assessed inability to respond. Prior screening using formal mental status questionnaires is generally not practical in surveys of older adults living in community. Such screening is costly and may threaten participation by those found to be sufficiently intact to participate. Thus, the prevalent approach to addressing inclusion of cognitively impaired elders in health surveys has been that of inclusion in the survey and later exclusion from data analyses of those with multiple errors on mental status questionnaires.

Decisions about exclusion criteria and/or cutoff points in screening tests for exclusion are typically accorded a one-line explanation in methods sections of health surveys. Yet the underlying assumptions bear examination, as they may have major implications for results. Problems may arise either based on undue restrictions of the sample or due to inappropriate inclusion of persons with MCI whose responses may suffer from poor reliability and limited validity. Furthermore, since cognitive impairment is a major dimension of ill health among the old old, exclusion of cognitively impaired respondents may bias our understanding of health care needs or health behaviors.

In longitudinal studies, such exclusions add to the existing problems of attrition due to death, refusals, and severe physical incapacity.

Rationale

In order to explore the role of MCI in affecting responses to health surveys, we examined the relationship

Boaz Kahana is at Cleveland State University, Cleveland, Ohio. Kyle Kercher, Eva Kahana, Kevan Namazi, and Kurt Stange are at Case Western Reserve University, Cleveland, Ohio.

between four alternative indicators of possible cognitive deficits and response patterns to our health survey.

We have recently completed the fifth wave of a longitudinal study of adaptation to frailty among an initial sample of 1,000 healthy elders living in three Florida retirement communities (E. Kahana & B. Kahana, 1995). During initial waves of the study, there were very few participants with serious health or cognitive limitations. However, in later waves, there has been increasing evidence of cognitive deficits among respondents. We originally established exclusion criteria whereby those individuals with four or more errors on the 10-item Pfeiffer Short Portable Mental Status Questionnaire (SPMSQ; Pfeiffer, 1975) would be excluded from data analyses. We undertook this research in part to help establish appropriate and empirically based exclusion criteria based on a threshold at which inadequate responses are likely to pose threats to survey validity.

Rather than conceptualizing threats to response validity as a unitary concept, we propose that multiple domains of responses may be affected. Our approach is illustrative of the variety of ways in which MCI might influence responses to health surveys. Our study presents data on the association between alternative cognitive impairment criteria and several dimensions of survey responses of participants in our longitudinal study.

If mild cognitive deficit leads to reduced quality of survey responses to health items, then we should observe the following results: Each of the four alternative measures of cognitive impairment, as described below, should show an association with the following outcome variables reflecting accuracy and reliability and completeness of survey responses: (a) the accuracy of height and weight estimates, (b) the reliability of physical health and psychological well-being measures, and (c) the frequency with which respondents fail to answer questions as part of physical health and psychological well-being measures.

Measures

Predictor Variables: Cognitive Deficits

We operationalized cognitive deficits through four separate measures (see Table 1 for univariate statistics) reflecting alternative methods: (a) Our self-report indicator of memory problems (metamemory) was a 5-point Likert scale that asked respondents, "During the past year, how much trouble did you have with your memory?" (b) Our interviewer-rated measure of mental confusion included a 5-point Likert scale that asked the interviewer to rate the respondent on his/her degree of mental confusion, from "not at all" to "very much." (c) The digit-symbol subtest of the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1981) constituted our performance-based measure of fluid intelligence. The digit-symbol test involves learning the association between meaningless symbols and meaningful semantic units (digits or numbers). This is a complex task that involves many aspects of cognitive functioning, including perceptual integration of complex visual stimuli, storage of information about the digit-symbol association, and retrieval of this information. It also measures speed in copying symbols. To the extent that respondents learn the associations between digits and symbols, their performance speed increases dramatically. In the current study, the performance of our respondents on the scale displays a mean of 34.8, a standard deviation of 10.2, and a range of 5 to 66. (d) The SPMSQ (Pfeiffer, 1975) constituted our performance-based measure of primarily crystallized intelligence. It measures memory for both recent and past events, memory for well-rehearsed and non-well-rehearsed information, orientation in place and time, and simple arithmetical abilities. Pfeiffer (1975) suggests that the number of errors roughly predicts different levels of competence behavior. Thus, five or more errors indicate that the person is in all probability unable to reliably take medications at prescribed times. Seven or more errors indicate the person cannot be left alone and is in need of a protective environment. In the current study, the performance of our respondents in this scale displays a mean of 0.7, a standard deviation of 1.3, and a range of 0 to 10.

Table 1. Means, standard deviations, and ranges for variables in the study

	M	SD	Range
Predictor variables			
SPMSQ	0.7	1.3	0–10
Digit-symbol	34.8	10.2	5–66
Self-reported memory	2.0	0.9	1–5
Interviewer's assessed memory	1.1	0.5	1–5
Outcome variables			
Height in inches			
Subjective	65.0	3.8	54–78
Objective	64.1	3.8	51–76
Absolute	1.2	1.8	0–13
Weight in pounds			
Subjective	145.1	27.0	68–234
Objective	145.9	28.0	67–245
Absolute	3.5	4.9	0–52
CES-D	18.5	5.4	10–43
SOPH	6.7	2.1	3–12
PANAS-PA	15.6	3.9	5–25
PANAS-NA	8.8	3.4	5–24

Outcome Variables: Reliability, Accuracy, and Completeness of Survey Responses

Outcome variables were assessed by three distinct measures (see Table 1 for univariate statistics): (a) how accurately respondents estimated their height and weight relative to performance-based measures (via ruler and scale), (b) how reliably respondents reported their physical health and psychological well-being, (c) how often respon

dents failed to answer questions regarding their physical health and psychological well-being.

Reliability of Responses to Selected Physical Health and Psychological Well-Being Scales

To assess whether the reliability of physical health and psychological well-being measures varied by degree of cognitive deficit, we examined the Cronbach's alpha for each of our six health and well-being composite scales, broken down by 10 categories (deciles) of the digit-symbol test and 4 categories of the SPMSQ (0, 1, 2, or 3 or more errors).

As one set of outcome variables, our study included a measure of physical health (i.e., subjectively rated overall physical health [SOPH]) and three measures of psychological well-being: a shortened version of the Center for Epidemiologic Studies Depression scale (CES-D) and the Positive Affect (PA) subscale and Negative Affect (NA) subscale of the Positive Affect/Negative Affect Scale (PANAS). SOPH consisted of three 5-point Likert scales asking respondents to evaluate their health from "poor" to "excellent." In the current study, these three items showed a clear-cut unidimensional factor structure and an overall alpha of .87. The composite scale has a mean of 6.7, a standard deviation of 2.1, and range of 3 to 12.

Our measure of depression consisted of 10 of the original 20 items in the CES-D (Radloff, 1977) with an alpha reliability of .83, a mean of 18.5, a standard deviation of 5.4, and a range of 10 to 43. PA and NA each included five items (rated on 5-point Likert scales) comprising these two subscales of the PANAS (Watson, Clark, & Tellegen, 1988). In the current study, the PA and NA subscales have respective overall alphas of .78 and .83, means of 15.6 and 8.8, standard deviations of 3.9 and 3.4, and ranges of 5 to 25 and 5 to 24, respectively.

Accuracy of Height and Weight Self-reports

We also assessed how accurately respondents estimated their height and weight compared with objective measures of height and weight. As a measure of accuracy, we subtracted the absolute value of the self-report measure from the absolute objective measure. The resulting height discrepancy measure had a mean of 1.2 inches, a standard deviation of 1.8, and a range of 0 to 13 inches. The resulting weight discrepancy measure had a mean of 3.5 pounds, a standard deviation of 4.9, and a range of 0 to 52. The correlation between objective and self-reported height was .86, and between objective and self-reported weight was .97. As an alternative measure of discrepancy between self-reported and objective measures of height and weight (more suitable for cross-tabular analyses), we also divided height and weight evaluation of respondents into categories of (a) accurate response, (b) mild over- or underestimation, and (c) extreme over- or underestimation.

Failure to Respond

As a final set of outcome variables, we measured the number of times respondents failed to respond to each or replied with "I don't know" to the preceding outcome variables: six physical health and psychological well-being scales and two measures of accuracy of height and weight self-reports. This single overall measure of nonresponse displayed only eight cases with at least one nonresponse. Accordingly, given the lack of variability in the measure, we dropped it from our originally planned analyses. We elaborate on this finding in the "Results" section.

Sample

The sample consisted of 598 residents of three Florida retirement communities. These respondents participated in the fourth annual wave of a longitudinal study dealing with adaptation to frailty. The mean age at Wave 4 was 83, with a range of 76 through 104. There were 148 respondents over the age of 85. Approximately 65% of the sample were women, 44% were currently married, 50% were widowed, 2% were currently divorced, and 4% were never married. Average education was 13.4 years.

Results

Descriptive Data

Table 1 provides the means, standard deviations, and ranges for each of our predictor and outcome variables. Another set of descriptive statistics on predictor variables includes frequencies (not shown in Table 1). With regard to the SPMSQ, 69.7% of respondents made 0 errors, 17.3% made 1 error, 7% made 2 errors, 2.7% made 3 errors, 1.3% made 4 errors, and less than 0.5% made 5, 6, 7, 8, 9, or 10 errors. Regarding self-reported problems with memory (metamemory), it is noteworthy that only 29.3% reported no trouble at all with their memory, 41.4% reported a little trouble, 25.8% reported having some trouble, but only 3.5% reported much or very much trouble. Furthermore, interviewers' ratings of confusion present a more positive portrait than respondents' self-reports of their memory: Rated as being not at all confused are 91.4% of respondents. Conversely, only 1% were rated as being "much" or "very much confused." Finally, digit-symbol responses, which reflect a complex set of cognitive demands, showed greater response variability than any other cognitive measure. A broad range of responses, from 5 to 66 correct, was observed.

Table 2 presents the intercorrelations of the four measures of cognitive impairment. The SPMSQ showed a high and significant correlation with interviewers' ratings ($r = .43$). This high correlation may in part be due, however, to the interviewers' opportunity to observe test responses to the SPMSQ. Correlations between the remaining indices of

Table 2. Zero-order correlations among measures of cognitive impairment (N = 560)

	1	2	3
SPMSQ errors	—		
Interviewer's rating of confusion	.43*** (.219)	—	
Self-reported memory problems	.16*** (.041)	.18*** (.061)	—
Digit-symbol correct	−.20***	−.31***	−.12**
	(.045)	(.096)	(.032)

NOTE: Numbers in parenthesis represent R^2 for cubic equation.
**p < .01. ***p < .0001.

cognitive deficits were moderate, ranging from −.12 between digit-symbol results and self-reported memory to −.31 between digit-symbol results and interviewers' ratings. These data suggest that the two performance-based measures, the observer-rated measure and the self-reported measure, may each tap different aspects of cognitive functioning.

Tests for nonlinearity (cubic equation) were also examined to ensure that the weak correlations observed are not based on lack of linear relationships among variables. Results showed little evidence of nonlinear associations (see Table 2).

Association Between Cognitive Deficits and Quality of Survey Data

Weight and Height Estimate

We examined the relationship between the accuracy of weight and height estimates and our four different measures of cognitive deficits. We first present cross-tabulated associations and subsequently linear and nonlinear parametric relations.

SPMSQ

Weight. When SPMSQ errors were cross-tabulated with the accuracy of weight estimates, no statistically significant association was found ($X^2 = 0.62$). There appeared to be little difference in the accuracy of weight estimates by those making 0, 1, 2, or 3 or more errors on the SPMSQ. Even the comparison of extreme groups (those with 0 SPMSQ errors and those with 3 or more) yielded comparable proportions of severe over- or underestimations of weight (17% or 15%).

Height. Similar non–statistically significant findings were obtained when the accuracy of height estimates and SPMSQ errors were cross-tabulated using five categories relevant to the accuracy of estimates. However, when categories of over- and underestimation were collapsed into three response options to handle inadequate cell sizes, evidence of significant association was obtained ($X^2 = 14.5$, $p < .03$). It appears that only 16% of respondents with 3 or more SPMSQ errors were able to give their actual height. In contrast, 29% of those making 2 SPMSQ errors, 38% with 1 error and 36% of those with 0 errors gave accurate height appraisals.

Digit-Symbol

Weight. Digit-symbol responses were divided into four quartiles based on the accuracy of responses and cross-tabulated with five levels measuring the accuracy of weight estimates. Once again, there were no significant associations observed ($X^2 = 14$, $p > .05$).

Height. A significant association was observed between digit-symbol responses and a three-category measure of the accuracy of reported height ($X^2 = 22.4$, $p < .03$). It appears that individuals in the two higher performance groupings were least likely to provide severe under- or overestimations of height. Interestingly, the likelihood of reporting one's exact height did not appear to be influenced by digit-symbol performance.

Linear and Nonlinear Parametric Associations Between SPMSQ and Weight/Height Estimates

Table 3 presents the correlations between our four continuous (i.e., noncollapsed) measures of cognitive deficits and two outcome measures of the absolute discrepancy between the subjective and objective estimates of height and the subjective and objective estimates of weight. The relationships were consistently weak with only two obtaining statistical significance at $p < .05$. Statistically significant but weak relationships were observed between digit-symbol accuracy and the accuracy of height estimates and between self-reported memory and the accuracy of weight estimates.

The small linear correlations observed could, however, mask larger, nonlinear relationships—for example, if only

Table 3. Zero-order correlations among four measures of cognitive impairment and accuracy of height and weight (N = 560)

	Digit-symbol	SPMSQ	Interviewer assessment	Meta-memory
Height discrepancy	−.12** (.021)	.08 (.019)	.07 (.017)	.00 (.003)
Weight discrepancy	−.06 (.007)	.05 (.013)	.05 (.005)	−.09* (.007)

NOTE: Numbers in parenthesis represent R^2 for cubic equation.
*p < .05. **p < .01.

the worst scores on the digit-symbol test were associated with height or weight discrepancies. Nonlinear tests based on cubic functions did not, however, reveal any substantial nonlinearities (see Table 3).

Cognitive Deficits and the Reliability of Survey Responses

We sought to determine whether there is a relationship between indices of cognitive impairment and consistency (reliability) of responses to health and well-being scales. Reliabilities were computed for our four criterion scales by 10 decile groupings of performance on the digit-symbol test and on four groupings of SPMSQ performance.

Table 4 reports reliabilities for digit-symbol-based groupings. As shown in Table 4, there were no consistent differences in reliabilities of scales (Cronbach's alpha) for respondents who performed at different levels of digit-symbol accuracy. Regardless of digit-symbol performance (from most to least correct), respondents provided reliable data on health scales. Even when comparing the lowest decile group with other groupings, there were no discernible patterns of lowered reliability.

When reliabilities of scales were compared for respondents with differing SPMSQ accuracy, reliability remained high for each subgroup (see Table 5). Once again, no consistent differences in reliability based on SPMSQ errors were discernible.

Table 4. Reliability coefficients (alpha) for CES-D, PANAS-PA, PANAS-NA, and SOPH with deciles of the digit-symbol

Digit-symbol deciles	n	CES-D	PANAS-NA	PANAS-PA	SOPH
1st	62	.797	.823	.722	.904
2nd	59	.708	.686	.728	.852
3rd	55	.791	.711	.803	.886
4th	58	.854	.837	.783	.885
5th	68	.750	.835	.559	.800
6th	56	.852	.837	.703	.882

Table 5. Reliability coefficients (alpha) for CES-D, PANAS-PA, PANAS-NA, and SOPH measures with SPMSQ scores

SPMSQ	CES-D	PANAS-NA	PANAS-PA	SOPH
0 errors	.828	.843	.763	.868
1 error	.807	.799	.717	.863
2 errors	.823	.720	.752	.818
3+ errors	.756	.811	.820	.886

Cognitive Deficits and Failure of Survey Measures

The degree to which memory deficits result in increased rates of "I don't know" responses or in missing data may be tested by comparing rates of such responses for study participants with different levels of cognitive deficit. In attempting to answer these research questions based on our data set, we discovered that there were too few "I don't know" responses or missing data to provide sufficient variability for an empirical exploration. Specifically, there was a maximum of eight instances of missing data or "I don't know" responses for any of the variables considered.

This finding in and of itself suggests that "I don't know" responses or missing data do not pose a major problem for face-to-face interviews involving older adults with MCI.

Discussion

Findings of our study provide consistent indications of reliable and generally accurate survey responses of old old persons even when they exhibit MCI. Structured surveys may help elicit from the elderly accurate and valid responses by providing a meaningful context for information retrieval. Under such conditions, the elderly tend to show minimal deficits (Craik & Jennings, 1992). It is thus clear from the data that mild memory loss does not automatically translate into incapacitating confusion.

To the extent that memory deficits result in problems with information retrieval, older adults with short-term memory deficits may have difficulty in accurately recalling information. Such recall problems may pose particular threats to reporting on health care utilization data, including the number of physician visits, timing of prior hospitalizations, or outpatient surgeries. In our health survey, there was no opportunity to obtain independent health care utilization data that would permit examining accuracy of recall. Respondents' memory problems, when self-perceived, may be reflected in increased "I don't know" responses or missing data. Alternatively, to the extent that respondents are unaware of memory deficits, inaccurate information on items requiring recall may be generated. Respondents may also deal with deficits by refusing to answer more demanding portions of the survey. While our data provide preliminary evidence about survey response accuracy and validity, future studies should explore in greater depth the accuracy in reporting health care utilization and other free recall data.

In considering our findings, it is noteworthy that they support Carp's (1989) conclusions that older adults do not require special methods of data collection and are able to perform on a variety of complex tests and questionnaires. Our data do not support suggestions that older persons show lower internal consistency in handling complex scales, for example, those in which negatively and positively worded items are balanced (Rodgers, Herzog, & Andrews, 1988).

In spite of general evidence of the ability of older adults with MCI to give consistent and reliable responses to

survey data, some words of caution about response accuracy are warranted. There were some indications that respondents with poor SPMSQ or digit-symbol performance were slightly more prone to inaccurate height (but not weight) estimations. This finding was surprising, since information about height generally represents a long-term memory that one continues to rehearse throughout life, whereas accuracy of weight reporting may relate more to short-term memory (due to greater fluctuations in weight). Inaccuracies in reporting weight or height were common, even among those with no evidence of cognitive impairment. Thus, motivational or other influences independent of cognitive deficits may also impact on accuracy of survey reporting.

In considering implications of our study for inclusion of older adults with cognitive impairment in health surveys, it is important to note that the vast majority of respondents showed only mild cognitive deficits. Although our findings underscore the robustness of survey responses, research on the cognitively more impaired elderly is needed.

Additionally, our findings also reveal that self-assessed memory has little correspondence with performance-based cognitive indicators (see Table 2). Thus, memory loss captured in performance-based measures does not automatically translate into self-reports of confusion. Recent dementia research has documented such discrepancies between the two methods (Zanetti, Bianchetti, & Trabucci, 1995).

Conversely, our findings indicate a substantial correlation between interviewers' ratings of confusion and the performance-based measure comprising the SPMSQ. In evaluating this correlation, it should be noted that interviewers may base their ratings on their observation of errors made by study participants in response to the SPMSQ. Thus, the two methods do not necessarily constitute independent assessments. Based on our findings, we would not recommend substitution of interviewers' ratings of confusion for performance-based assessments such as the SPMSQ.

In evaluating implications of this study, it should be noted that in this research, as in other community-based surveys, there is a selection factor operating that excludes individuals who do not feel sufficiently cognitively or physically intact to participate in research. Accordingly, we may generalize our findings only to those segments of older adults with MCI who are able to compensate for their cognitive deficits sufficiently to maintain independent lifestyles and who volunteer to participate and continue in survey research. Of course, it is this group that generally participates in health surveys, and thus the findings have important and reassuring implications for survey researchers.

Conclusion

Our results, based on a convergent set of indicators, show that survey response accuracy generally remains unaffected by mild cognitive deficit among the aged living in community. The findings thus add reassuring information to other recent studies regarding the limited evidence of nonresponse rates among the old old. It thus appears that concerns about conducting reliable and valid social research among the old old may have been exaggerated.

References

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders: DSM-IV (4th ed.). Washington, DC: American Psychiatric Association.

Butler, R. N., Lewis, M. I., & Sunderland, T. (1991). Aging and mental health: Positive psychosocial and biomedical approaches (4th ed.). New York: MacMillan.

Carp, F. (1989). Maximizing data quality in community studies of older adults. In M. P. Lawton & A. R. Herzberg (Eds.), Special research methods for gerontology (pp. 93–122). Amityville, NY: Baywood.

Carsjo, K., Thorslund, M., & Warneryd, B. (1994). The validity of survey data on utilization of health and social services among the very old. Journals of Gerontology, 49(3), S156–S164.

Craik, F., & Jennings, J. (1992). Human memory. In F. Craik & T. Salthouse (Eds.), The handbook of aging and cognition. Hillsdale, NJ: Lawrence Erlbaum.

Crook, T., & Larrabee, G. (1988). Age associated member impairment: Diagnostic criteria and treatment strategies. Psychopharmacology Bulletin, 24, 509–514.

Doppolt, J., & Wallace, W. (1955). Standardization of the Wechsler Adult Intelligence Scale for older persons. Journal of Abnormal and Social Psychology, 51, 312–330.

Eisdorfer, C. (1977). Stress, disease and cognitive change in the aged. In C. Eisdorfer & R. Friedel (Eds.), Cognitive and emotional disturbances in the elderly. Chicago: Year Book Medical.

Kahana, E., & Kahana, B. (1995). A preventative model of successful aging. In V. Bengtson (Ed.), Continuities and discontinuities in the lifespan. New York: Springer.

Lawton, M., & Herzog, A. (1989). Special research methods for gerontology. Amityville, NY: Baywood.

Pfeiffer, E. (1975). A short portable mental status questionnaire. Journal of the American Geriatric Society, 23, 433–441.

Powell, D. (1994). Profiles in cognitive aging. Cambridge, MA: Harvard University Press.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1, 385–401.

Rodgers, W. L., Herzog, A. R., & Andrews, F. M. (1988). Interviewing older adults: Validity of self-reports of satisfaction. Psychology and Aging, 3, 264–272.

Salthouse, T. (1991). Theoretical perspectives on cognitive aging. Hillsdale, NJ: Erlbaum.

Schaie, W. (1983). Longitudinal studies of adult psychological development. New York: Guilford Press.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–1070.

Wechsler, D. (1981). Wechsler Adult Intelligence Scale—revised. New York: Psychological Corp.

Zanetti, D., Bianchetti, A., & Trabucci, M. (1995). The puzzle of functional status in mild and moderate Alzheimer's Disease. The Gerontologist, 35(2), 148.

Converting an Ongoing Health Study to CAPI: Findings From the National Health and Nutrition Examination Survey III

Jane Shepherd, David Hill, Joel Bristor, and Pat Montalvan

For many years, the National Health and Nutrition Examination Survey (NHANES) has been a prominent source of information on the health and nutritional status and disease and risk factors of the U.S. population. NHANES III was conducted from 1988 through 1994, involving over 33,000 comprehensive household interviews along with follow-up examinations in mobile examination centers (MECs) or in homes.

During the 6-year data collection period, survey operations were in progress simultaneously in three locations, referred to as "stands." A total of 89 stands were operational during the 6-year period. In each stand, a field office was set up, and on average, 380 sampled persons were interviewed in households and examined in the MEC over a 6- to 7-week period. The field staff consisted of 15 persons organized into three administrative teams, about 27 full-time household interviewers, 10 backup staff, and about 32 examination center staff organized into two teams. All field staff were on full-time travel status. Each field office was equipped with a micro VAX computer and CRT terminals to operate the NHANES Automated Field Office Management System (AFOMS) software developed by the National Center for Health Statistics (NCHS). This hardware platform was utilized further during computer-assisted personal interviewing (CAPI) activity to download case assignments to portable laptop computers. Field office equipment also included modems for on-line electronic data connections to off-site offices and support facilities.

At each sampled household, field interviewers conducted screening interviews to determine eligible persons. The interviewer then completed a family questionnaire for each family having a sampled person, conducted an adult or youth medical history questionnaire for each selected person, took each adult sampled person's blood pressure, and arranged MEC examination appointments. If a sampled person was unable to come to the MEC, a qualified examiner visited the household to administer an abbreviated, 45-minute home exam.

For the fourth year of NHANES III, a major conversion effort was undertaken to change all household interviewing from hard copy questionnaires to CAPI. The conversion to CAPI was completed at the conclusion of the fourth year, and CAPI data collection was implemented for the fifth and sixth years of the survey.

Most of the development work for the CAPI conversion was accomplished within a 6-month period. Following the developmental work, a 5-day pilot test was conducted to evaluate the software and operational procedures. The major aspects of the conversion included instrument redesign from hard copy to CAPI, introduction of CAPI features to support survey operations, interviewer and field office staff training, data editing and postprocessing, and data quality assurance.

Instrument Conversion and Design

In converting from hard copy to CAPI, a number of issues that were basic to the design and operation of the survey had to be addressed. They included the following:

1. The screener questionnaire needed to remain as hard copy to provide flexibility in its administration. The screener was a brief instrument, typically completed on the doorstep with a respondent or neighbor, to collect basic demographic information needed to complete the household subsampling procedures. It was felt that converting this instrument to CAPI and requiring that interviewers use computers might make it more difficult to complete and have an adverse impact on the screening response rate. Therefore, the CAPI questionnaire had to be designed to allow information from the screening instrument to be entered for those households with sampled persons.

2. Since NHANES involved multiple instruments that were administered with different persons within the household, the CAPI setup had to allow the interviewer to select which person to interview and which instrument to use at any point in time. Availability of the respondents dictated the sequence of the interviews within a household, and the CAPI system had to accommodate this need.

3. Within the conduct of the interview, information might be provided by a respondent that was out of sequence for the interview. For example, a parent reporting for

Jane Shepherd, David Hill, and Pat Montalvan are with Westat, Inc., Rockville, Maryland. Joel Bristor is with the National Center for Health Statistics, Hyattsville, Maryland.

one child would report about the vitamins taken for other sampled children within the household. CAPI needed to capture this information and provide a method for inserting the information in its proper place.

4. Due to the oversampling of Hispanics in the survey, many interviews were conducted in Spanish. Interviewers needed the ability to move between English and Spanish during an interview or between respondents within the same household.

Features and Capabilities Introduced With CAPI

Basic features introduced with CAPI included automatic skip and flow routing, range checks, customized word fills, context dependent help, and automatic status code assignment for completed cases. More advanced features, including consistency edits and precoded items, were introduced to ensure data collection quality.

Screen design and data entry conventions were examined as part of the conversion. It was important to maintain the look and feel of the individual questions to facilitate the transition for the interviewers. When possible, the format of questions was maintained between the hard copy and CAPI. Question numbers were shown on the CAPI screens, and words that were underlined in the hard copy were displayed in reverse video to indicate emphasis.

In addition, several special features were added to facilitate the conversion for interviewers. These included

1. creating a main roster to list all required interviews within a household (screener, adult, youth, and family questionnaires) with the name of the respondent and the completion status. Interviewers were able to administer the questionnaires in any sequence, if the screener was completed.

2. developing specialized CAPI features to support the needs of the interviewer in converting from hard copy to CAPI. These features included

   a. capability to interrupt the interview at any point and access a special navigation screen with options to correct screener demographic information (spelling of name, date of birth, and gender); add information about vitamin, mineral, or prescribed medicine usage for another household member whose CAPI interview was already completed; create a note about vitamin, mineral, or prescribed medicine usage for a person not yet interviewed that would be displayed during the appropriate CAPI interview; and view a table of contents indicating location within the current interview or switch the language of the interview between English and Spanish (see Figure 1).

Figure 1. CAPI navigation menu

INTERRUPT MENU

1. Call the Field Office/Schedule SP Appointment
2. Update/Correct Screening Information
3. View Table of Contents
4. Add a Complete Vit/Min, Meds, or Antacids Record
5. Note for Vit/Min, Meds or Antacids
6. Switch Language of Interview

SELECT MENU OPTION: (1)

[PRESS ENTER KEY TO RETURN TO INTERVIEW]

NOTE: "SP" stands for sampled person.

   b. capability to create a comment at any point in the interview for review later by the interviewer or as part of the field office data review.

   c. capability to execute a nondestructive jump back. Interviewers were able to interrupt the interview and access a special window displaying a list of completed items or sections. After interviewers selected a jump back item, the system took them back to that item and displayed the previous response. If the interviewer altered the response, the system would skip to the next appropriate question, based on the new response. After completing any newly identified items as a result of a response change, the interviewer could "fast-forward" back to the original place in the interview and continue.

3. modifying the design of selected items to minimize text entry. The number of response categories was expanded for some items to include responses commonly reported as "other, specify," and specialized displays were included on some CAPI screens. For example, reported medical conditions were displayed during the section collecting information about medications. Also, customized on-line coding screens were developed for frequently reported conditions, vitamins, and antacids. Context-specific help screens were also included to assist the interviewers with difficult questions or terminology.

4. developing a Spanish version of the instrument. Interviewers had the capability to switch languages at any point in the household interview.

Other requirements of the conversion included support for blood pressure recording in the CAPI adult questionnaires and CAPI support for the home examination component, which included spirometry software. The home exam was a separate component administered by a medical technician involving numerous physical measurements. The CAPI system for the home exam had to coexist with the

spirometry software and provide integrated menus to access this software and store the results of the spirometry test.

Interviewer/Field Office Staff Training

The conversion to CAPI interviewing in Year 5 was accomplished while the survey was in operation, without a break in the household interviewing or examination activities.

Interviewer Training

The NHANES interviewers were very experienced in the use of the hard copy instruments and knowledgeable about the questionnaire content. Moreover, they were proficient in the field procedures associated with the administration of the survey questionnaires, experienced in their editing tasks, and well versed in the submission of work to their supervisors. Therefore, the focus of the training was to cover only those interviewer tasks that would be affected by the conversion to CAPI. These included the use of the laptop, the Westat Cheshire CAPI system, modifications and enhancements to the questionnaires due to the conversion, administrative changes implemented because of CAPI, as well as any changes to their postfield activities.

Training was conducted in 3½ days. It included a half-day introduction to the CAPI system and laptop use, 2½ days of instrument specific training, and another half day covering interviewing and reporting procedures affected by the conversion. The training techniques employed included lectures, interactive mock interviews, and dyad role-playing.

Field Office Staff Training

Each NHANES field office had a field manager, three staff members responsible for various field and office support tasks, and one data editor responsible for data quality review.

A 2-day training session was conducted for field office staff and followed by on-site support from home office staff for 2 weeks following the systems conversion.

The training session focused on core tasks affected by the conversion to CAPI. This included case assignment and reassignment, the receipt and monitoring of all types of work, quality control, adjudication, and the closedown of operations, including the delivery of data at the end of each stand.

All interviewers were able to make the conversion to CAPI successfully. Most interviewers preferred the CAPI system to hard copy data collection. The field office staff also made a successful transition to CAPI and automated processing procedures. Within the field office, the job descriptions for the field office manager and the data editor were changed substantially by the introduction of CAPI, and they required the most follow-up training. This was primarily related to changes in case-processing flow within the field office and the need to prepare the CAPI data files for final delivery to NCHS at the end of each stand.

Data Editing and Postprocessing

With the CAPI implementation, the role of the editor changed significantly. The editor's position before CAPI implementation was to review all hard copy instruments for skip errors, reconcile demographic information across the various data collection instruments, review comments and open-ended text responses, and update the hard copy instruments as appropriate, sometimes in consultation with the interviewers. The editor was also responsible for handling the flow of information to and from the MEC. A summary of medical conditions and the questionnaire responses was sent to the MEC prior to each sampled person's appointment for review by the medical staff. Also, key data items identified for data retrieval were forwarded to the MEC following the editor's review. This process was labor-intensive and entirely manual prior to the introduction of CAPI.

Many of these processes were automated with CAPI. The role of the editor now required knowledge of how to update on-line CAPI data files and to perform data editing prior to file delivery at the end of each stand. The review for skip errors and reconciliation of demographic data across survey instruments was eliminated with CAPI. Other manual operations were facilitated by the generation of reports from the CAPI system for review by the data editor.

Editors were trained to update the CAPI database based on comment review and to recode open-ended responses as appropriate. Also, automated reports were generated for data items to be provided to the MEC or those requiring data retrieval. The editors on all three teams were able to make a smooth transition to CAPI and were successful in making the final postcollection updates needed to ensure data delivery following each stand.

Data Organization and Quality

Major concerns in implementing an automated system in midstream included ensuring high-quality, consistent, and well-documented data across the hard copy and CAPI modes of data collection and deciding how to employ quality-enhancing CAPI features such as special probes and on-line edits in ways that enhance data quality.

Database Design

The first issue in designing the CAPI system was to ensure consistency between hard copy and CAPI data. The starting point for the CAPI database design was the existing records layouts from keying operations for the hard copy

questionnaires. These records layouts specified variable names and ranges for all data items in the five hard copy instruments. Documentation produced during the CAPI implementation included hard copy questionnaires annotated with variable names linking to the hard copy data collection years, flow charts of the overall system and each instrument, a codebook, and screen printouts identifying the variables collected.

Demographic Variable Reconciliation

In instances in which the same data items, such as age and gender, were collected via multiple hard copy instruments, different variables were created. The CAPI database was designed to collapse some of these variables into a single variable, such as current age, and execute checks against this current value throughout the various CAPI instruments.

Edits were implemented in CAPI to check for the consistency of key demographic information across the screener, adult or youth interview, and family questionnaire. For example, during the screener, a respondent could answer questions about the ages of other household members as a proxy. Therefore, this information did not always agree with data provided by each household member during the adult or youth interview. CAPI edits were incorporated to display previously reported demographic information and update the information as necessary during the extended interviews. By maintaining a single current value for all demographic data items across the different interviews, extensive postcollection editing was avoided.

In examining data from Phase 2 (Years 4, 5, and 6) of the survey, only a small number of cases required adjudication of demographic data (see Table 1). In Year 4, the data editor reviewed these items at each stand before the questionnaires were keyed in. In Years 5 and 6, the CAPI software was responsible for this adjudication. Few inconsistencies were found in the Phase 2 data among these items.

Review of Skip Patterns

During preliminary review of adult questionnaire data from Phase 2 of the survey, few skip pattern errors were detected. For Year 4 data, the editors did an effective job of scanning the hard copy questionnaires and correcting any skip errors prior to key entry. In Years 5 and 6 of the survey, CAPI successfully replaced the role of the data editors with regard to the elimination of skip errors at the point of data collection. The few skip errors detected in the CAPI data resulted from data updates made during postprocessing.

Table 1. Phase 2 demographic variables requiring adjudication

	N	Age	Gender
Year 4	6,903	< 1%	< 1%
Years 5 and 6	12,515	< 0.1%	< 0.1%

Logical Consistency

Logical consistency edits were introduced into Section B of the Youth Questionnaire to maintain the relationship between data items not accounted for by skip patterns. In Section B, questions were asked about the age when the baby was first fed something other than breast milk, when the baby was first fed formula, and when the baby first started eating solid foods. These ages should have been less than or equal to the present age of the baby in order to maintain the logical consistency among items.

In the hard copy data collection, the interviewer reviewed the question responses to check for logical consistency; in CAPI, logical edits were built into the system to indicate potential discrepancies with age. In Year 4, approximately 2% of the cases had at least one age inconsistency in these items. In Years 5 and 6, only 6 of 3,986 youth interviews (0.15%) were found with age inconsistencies in this section. Again, these inconsistencies resulted from data updates made during postprocessing.

Use of Dependent Probes in CAPI

Several quality-enhancing CAPI features were introduced into the food frequency section of the adult and youth interviews. The food frequency section contained 67 questions about food consumption. The section was divided into questions about each of the major food groups (milk and milk products, main dishes, fruit and fruit juices, vegetables, etc.). Most questions asked about the consumption of a particular item or type of food (such as broccoli or salad) over the past month. If the food item was never consumed, the interviewer entered 0 to indicate "never" and to skip the quantity/unit question for that food item.

In converting this section to CAPI, words that were underlined in the hard copy were displayed in reverse video so that the interviewer could emphasize these words when reading the questions. Reverse video was used to reference the time period for the question and selected food items.

The section was extensive, and interviewer Q × Qs and hand cards were used to assist in identifying food items corresponding to a given question. Special probe boxes were added to the CAPI screens for selected questions to reference additional food items or common brand names as applicable.

In addition, special on-line data dependent probes were introduced to ensure consistency in reporting food items. Sometimes questions about food items typically consumed together were asked at different points in the section. In these instances, it was up to the interviewer to probe or refer back to earlier questions to catch any potential inconsistencies in reporting during the hard copy administration. For example, the question about milk consumption

was asked before the question about cereal consumption, and these items were frequently consumed together.

In CAPI, data dependent probes were introduced to remind the interviewer about previously reported consumption. For the cereal question, a display at the bottom of the screen was included to indicate previously reported milk consumption. Based on prior responses within the section, the display could read, "INTERVIEWER: SP HAD MILK 1 TIME PER DAY" or "SP NEVER HAD MILK."[1] Similar data dependent probes were introduced for other items, including a probe about bread consumption when asking about margarine and butter and about mentions of tossed salad when reporting salad-dressing usage.

Few differences were found in reports of cereal-without-milk consumption between the hard copy (Year 4) and CAPI data collection modes (see Table 2). There were also very similar findings in the percentage of sampled persons reporting tossed salad without dressing and white bread without fats (butter/margarine) between the hard copy and CAPI modes (see Table 3).

On-Line Coding

The last question in the food frequency section asked, "Have I missed any other foods or beverages that you had at least once per week in the past month?" In Year 4, the interviewer listed the items not classified elsewhere. In the CAPI implementation, a special screen appeared when additional food items were reported that listed some items commonly missed, such as egg substitutes, decaffeinated coffee, and pancakes. The interviewer was able to make a selection from this special list, enter another food item, or jump back to a prior question if the item was inadvertently missed.

Table 2. Reports of cereal-without-milk consumption from adult food frequency section (percentages)

	Year 4	Year 5
Bran cereal without milk	0.7	0.7
Wheat cereal without milk	0.4	0.4
Hot cereal without milk	6.1	5.6
Other type of cereal without milk	3.6	3.0

Table 3. Reports of salad and bread consumption from adult food frequency section (percentages)

	Year 4	Year 5
Tossed salad without dressing	9.3	7.3
White bread without butter/margarine	25.9	22.7

[1] "SP" stands for sampled person.

In comparing data from Year 4 with data from the CAPI stands in Year 5, we found that "other" foods or beverages were reported with approximately the same frequency in Years 4 and 5 (see Table 4). However, due to the introduction of the special coding screen in CAPI, a much smaller percentage required manual review and coding.

Other coding screens introduced for the prescribed medicine, condition, and antacid questions were also effective in reducing the amount of postprocessing.

Summary

While there is rapidly growing attention being given to using CAPI as an alternative to hard copy questionnaires in health research, most discussions seem to focus on the creation of new CAPI surveys from the ground up. This leaves unattended a number of important issues faced when converting to CAPI in ongoing studies.

The process of instrument conversion for an ongoing study needs to include consideration of how the instrument is currently being administered in the field. Issues related to how the interviewer uses the hard copy to check other responses or verify information need to be considered in the design of the CAPI system. In NHANES III, special features were added to support these needs, including the capability to jump back to earlier responses, view a table of contents, update key variables, and move between the questionnaires. Also, special data dependent probes and edits were introduced to display the same types of information the interviewer could have accessed by glancing back through the hard copy.

CAPI systems provide many standard features that facilitate data collection, such as automated routing and editing. However, in order to integrate CAPI into a complex, multicomponent study, additional capabilities were required. Several special menus were designed for the NHANES application to provide capabilities for the interviewer to select specific case components, interrupt the interview and complete other tasks, enter comments, and switch the language of the interview. Other special features included the ability to record blood pressure data and link to spirometry software on the same laptop. CAPI systems for surveys such as NHANES need a flexible, modular design that permits the addition of new components and the

Table 4. Reports of other foods or beverages from adult food frequency section (percentages)

	Year 4	Year 5
Cases reporting "other"	7.5	8.8
No. "other" food items reported		
1	6.2	6.7
2	1.3	2.1
3–6	< 1	< 1
Cases requiring manual coding	7.5	2.4

ability to link to other software systems to support more complicated types of data collection.

NHANES has the unique requirement of training both interviewing and field office staff. The training program was designed to focus on the areas of change introduced by the conversion to CAPI. All staff members made the conversion successfully.

Examination of data collected pre-CAPI with data collected using CAPI indicated little difference in overall quality for the items reviewed. The thorough manual review of all questionnaire items and adjudication of key variables across the hard copy questionnaires prior to key entry by the field data editors was successfully automated by CAPI. Also, data editors were trained to update CAPI data files and process the data for final delivery from distributed field locations.

As a leading example of converting a major ongoing health study to CAPI, the NHANES III experience offers important findings about how the capabilities of CAPI can be utilized effectively in health survey research.

Reference

U.S. Department of Health and Human Services (1994). Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–94. Hyattsville, MD: National Center for Health Statistics.

Multilingual ACASI: Using English-Speaking Interviewers to Survey Elderly Members of Korean-Speaking Households

Tabitha P. Hendershot, Susan M. Rogers, Jutta P. Thornberry, Heather G. Miller, and Charles F. Turner

Introduction

Most national surveys presume that respondents speak enough English to communicate adequately with an interviewer or are sufficiently literate in English to read survey instructions and complete a self-administered questionnaire (SAQ). This presumption presents problems for several subpopulations in the U.S. whose written or oral command of English is insufficient to complete an English-language survey. Such individuals are likely to be excluded from most major in-person surveys, due to the logistic and economic difficulties of conducting field interviews in a myriad of foreign languages. Efforts to collect information through proxies are hampered by their incomplete knowledge of the subject or the absence of a bilingual person to serve as a proxy at the time of the interview. Excluding non-English-speaking individuals from U.S. health surveys has left us with an incomplete picture of national health problems and health care needs. This gap may be especially critical among the elderly, a group with varied and sometimes extensive medical problems who may not be as well integrated—either linguistically or culturally—into mainstream American society as younger people.

Scientists at Research Triangle Institute (RTI) have recently developed a new interview technology that may facilitate the inclusion of non-English-speaking populations in national surveys. Audio computer-assisted self-interviewing (ACASI) technology uses a laptop personal computer with a digital audio adapter to administer questionnaires in multiple languages. Subjects hear the questions through headphones and enter responses directly into the computer using the keyboard. ACASI offers several advantages over face-to-face interviews and SAQs. It provides privacy without requiring literacy; every question is asked in exactly the same way and in the same order; the system can be programmed in any spoken language; and it can, if needed, display visual cues, such as pictures of medicines. Because data files are automatically produced as subjects enter responses, the ACASI system eliminates the delay in coding, entering, and cleaning data, and it also eliminates some of the clerical errors inherent in creating data files from paper-and-pencil interviews (PAPI).

With support from the National Institute of Child Health and Human Development and the National Institute on Aging, we are presently conducting a 4-year program of basic research on the application of ACASI technology to survey measurement. An important aspect of the initial stages of this research focuses on the development and testing of field procedures for multilingual field interviewing using ACASI technology and monolingual field interviewers.

This study reports preliminary findings from an ACASI health survey in a small sample of 30 older Koreans who speak limited or no English. This study has two aims: (a) to evaluate whether or not field interviewers who speak no Korean can screen households, obtain consent from eligible subjects, and administer an ACASI interview to Korean respondents who speak little or no English and (b) to assess the consistency of responses obtained from the multilingual ACASI interviews and from telephone follow-up interviews conducted in Korean.

The present study complements earlier research that successfully used multilingual ACASI to survey a younger sample of Spanish-speaking respondents (Turner, Rogers, Hendershot, Miller, & Thornberry, 1995). As will be discussed later, the present research tests what is in many ways a worst-case scenario for multilingual ACASI interviewing.

Methods

Two community service organizations serving the Korean population in the Washington, DC, metropolitan area provided names of potential respondents. Inclusion criteria limited the sample to Korean-speaking adults over 50 years of age who had limited or no written or oral English language capabilities. Lead letters written in Korean were sent to households of all potential respondents prior to contact by a field interviewer. All interviews were conducted by two experienced field interviewers, neither of

The authors are with the Program in Health and Behavior Measurement and with the Center for Survey Research at the Research Triangle Institute, Rockville, Maryland, branch. The research reported in this paper was supported by the National Institute of Child Health and Human Development and the National Institute on Aging of the National Institutes of Health under grant number HD31067 to Charles Turner.

whom could speak Korean. Training for interviewers included information on administration of the ACASI interview as well as Korean customs and translations of introductory phrases, such as "hello" and "thank you." Each field interviewer carried a cellular phone linking them to a Korean-speaking central office interviewer.

As a first step in the interview, field interviewers were instructed to create a roster of all household members 18 years old or older, although the eligible subject was known in advance of contacting the household. The household screening instrument, which was available in English and Korean, collected data on the age, sex, and relationship of all individuals living in the dwelling unit to the head of household. Two approaches were used if the interviewer was unable to successfully conduct the screening or recruit the eligible subject: (a) Interviewers sought assistance of an English-speaking member of the household to explain the survey and to elicit consent, or (b) interviewers contacted the Korean-speaking central office interviewer, using a cellular phone to assist in explaining the study and recruiting the preselected respondent. To assist in the refinement of field procedures, interviewers were instructed to complete a standardized questionnaire eliciting details of their interview experience. Administrative problems or other pertinent events that occurred during the interview process were recorded. Subjects received training materials to illustrate use of the ACASI equipment, including a show card for key function, in Korean.

Of the 68 individuals who were sent lead letters, interviewers were not able to locate 11 (16%). Of the remaining 57 eligible subjects, 13 (23%) were not available and nine (16%) refused to participate. One case is not included in these analyses because of the subject's marked hearing impairment, which resulted in a proxy completing the interview. This preliminary study reports findings from 30 interviews with adult Koreans conducted from September 1994 through April 1995.

The ACASI interview consisted of 42 questions on the respondent's health, demographic characteristics, and use of health services. Most questions were closed-ended and used simple response categories; open-ended questions were not complex and generally required a numeric response, such as age or number of children. In addition, there were several questions concerning the respondent's knowledge and use of English and the respondent's evaluation of the ACASI system. All recruited subjects completed the interview; there were no break offs once the interview started. Each respondent was paid $10 for participating; this payment was not considered to be an incentive, since notification of payment was made at the end of the interview.

To assess the reliability of data obtained using ACASI, a follow-up telephone interview was conducted in Korean by the central office interviewer. Follow-up telephone interviews were conducted within 5 to 38 days (mean of 14.3 days) after the ACASI interview. The follow-up interview consisted of eight questions on health behaviors repeated from the ACASI interview; in addition, the household screening information was elicited again during the follow-up interview. One of the eight health-related questions was dropped from analyses because the ACASI translation differed from that asked at follow-up.

Results

Characteristics of the 30 respondents who completed the ACASI interview are shown in Table 1. The majority of respondents (64%) were female. The age of subjects ranged from 54 to 84 (with a mean age of 71.3 years). Most respondents were married or living together as a married couple (63%), 20% were widowed, 10% were separated, and 7% had never married. All respondents reported having children. Less than one-half (43%) had completed high school, 13% had attended college, and two respondents (7%) had completed college. Although nearly one-half of respondents reported ever working for pay for one or more months, only three respondents were currently employed.

Respondents were asked to evaluate their understanding of English based on how easily they read and speak English and whether they had ever talked on the telephone in English for longer than two minutes. Approximately one-third of the Korean respondents reported they did not speak or did not read English at all. The vast majority (97% and 93%) expressed some difficulty in speaking or reading English (see Table 2). Eighty percent of respondents indicated they had never conducted a telephone conversation in English that lasted two or more minutes.

In order to assess respondents' experience using the ACASI system, subjects were asked how well they under-

Table 1. Selected characteristics of respondents (N = 30)

Mean age (years)	71
Female	64%
Married/living together as married	63%
Completed high school	43%
Completed college	7%
Currently employed	10%

Table 2. Respondents' subjective assessments of their English language skills (percentages; N = 30)

Difficulty speaking English	
Don't speak English	30
Lots of difficulty or worse	57
Some difficulty or worse	97
Difficulty reading English	
Don't read English	27
Lots of difficulty or worse	50
Some difficulty or worse	93

NOTE: These estimates were tabulated from responses to the following two questions: (a) "Now we would like to ask you some questions about your understanding of English. Would you say that you can speak English easily, with some difficulty, with a lot of difficulty, or not at all?" and (b) "Would you say that you can read English easily, with some difficulty, with a lot of difficulty, or not at all?"

stood the audio component and how comfortable they were using the keyboard (see Table 3). The majority (77%) indicated that it was either very easy or somewhat easy to understand the Korean questions through the headphones. Only three respondents (10%) reported problems understanding the questions being asked and also reported difficulty using the keyboard but managed to complete the interview. Of the five respondents who reported using a hearing aid, only one reported problems understanding ACASI questions. The vast majority of respondents (87%) indicated that they were very comfortable or somewhat comfortable using the computer. Time to complete the ACASI interview ranged from 9 to 27 minutes; the mean completion time was 16.4 minutes. On average, time to complete the ACASI interview did not differ by age of the respondent. Time to complete the survey was slightly longer ($p = .05$) among those who expressed more than "some difficulty" reading English, compared with those who read English easily or with some difficulty. To assess the reliability of respondents' answers to the ACASI interview, telephone follow-up interviews were conducted by the Korean-speaking central office interviewer an average of 14.3 days after the initial household interview. Results of these comparisons for 27 respondents are presented in Table 4.

Table 3. Respondents' evaluation of ease of use of ACASI technology (percentages; N = 30)

Ease in use of computer	
Very comfortable	50
Somewhat comfortable	37
Somewhat uncomfortable	13
Very uncomfortable	0
Understanding of questions	
Very easy	27
Somewhat/very easy	50
Somewhat difficult	13
Very difficult	10

NOTE: Tabulated from responses to the following questions: "How easy or difficult was it for you to understand the questions being asked in the tape recording? Was it very easy, somewhat easy, somewhat difficult, or very difficult for you to understand the questions being asked?" and, "How comfortable did you feel typing into the computer? Did you feel very comfortable, somewhat comfortable, somewhat uncomfortable, or very uncomfortable typing your answers into the computer?"

Overall, the proportion of consistent responses is not very impressive and ranges from 70% to 93%. Consistency appears to vary by the retrospective period of report. For example, reports on current smoking were fairly consistent (range of 83% to 100%), as were reports on current use of eyeglasses or contact lenses (range of 78% to 100%) when comparing respondents who had hearing problems (column a), were younger (column b), or had higher levels of education (column c) with all respondents. In contrast, consistency of reports of visiting a dentist in the last year ranged from 60% to 80% and ever having had a cholesterol test ranged from 56% to 100%. It is unclear if inconsistency in reports of cholesterol tests reflects the lifetime retrospective period or a misunderstanding of the question. Although the ACASI and follow-up interviews both used the

Table 4. Consistency of responses in (a) multilingual ACASI interviews and (b) telephone follow-up interviews

	Korean				Spanish
Measurements	All	Exclude hearing problems[a]	< 70[b]	High school graduates[c]	All
No. children	70*	70**	80*	75*	88*
Attended high school	81***	85***	90***	67	92***
Ever worked for pay for 1+ months	81***	85***	90***	83***	85***
Smokes cigarettes now	93***	100***	90†	83‡	100***
Wears eyeglasses or contact lenses[d]	85‡	90†	100***	92†	92***
Visited dentist in past year	74***	80***	60	75‡	81***
Ever had a cholesterol test	70†	80***	90***	67	84***
Base N	27	20	10	12	26

[a]Excludes five respondents with completed follow-ups who reported use of a hearing aid plus two additional respondents who reported problems with volume and speed of ACASI questions.

[b]Excludes 17 respondents aged 70 or older.

[c]Excludes 15 respondents who reported in Korean telephone interview that they had not completed high school.

[d]Note that there was a slight change in the wording of this question between interviews. The ACASI instrument asked respondents if they wore "eyeglasses or lenses," while the follow-up interview asked if they wore "eyeglasses or contact lenses."

*$p \leq .01$ for test of association between measurements of interval level variables at two time points (by Pearson product moment correlation).

**$p \leq .05$ for test of association between measurements of interval level variables at two time points (by Pearson product moment correlation).

***$p \leq .01$ for test of association between measurements of two-category variable at two time points (by likelihood ratio chi-square test of independence model for 2×2 table; $df = 1$).

†$p \leq .05$ for test of association between measurements of two-category variable at two time points (by likelihood ratio chi-square test of independence model for 2×2 table; $df = 1$).

‡$.10 \leq p \leq .05$ for test of association between measurements of two-category variable at two time points (by likelihood ratio chi-square test of independence model for 2×2 table; $df = 1$).

English and Korean words for cholesterol, several respondents did not understand the Korean word for cholesterol but did recognize the English word.

Consistency also appears to vary by certain characteristics of respondents. Compared with all respondents, respondents who did not report hearing problems (column a) have a slightly higher proportion of consistent responses for all questions except number of children, which is essentially the same. Younger respondents (column b) also have higher rates of consistency for all questions except current cigarette use (which is essentially the same) and visiting the dentist in the past year (which is slightly lower). However, the consistency of responses does not appear to improve as a function of increased education (column c). There were gender differences in response consistency, but no clear trend emerges from these data. This may be related to small sample size, especially for male respondents.

Particularly striking is the lack of consistency of reports of number of children and whether or not the respondent had attended high school. Overall, only 70% of respondents reported the same number of children during the ACASI interview and at follow-up. The central office interviewer indicated that this question was problematic for several reasons. Some subjects reported the number of sons and had to be prompted to give the number of daughters; other subjects reported the number of children in the United States and had to be reminded to include the number of children remaining in Korea. Ultimately, the question asked at follow-up was reworded, asking subjects how many children in total they had. In the ACASI interview, two female respondents reported that they were never married and had children. At both the screening and follow-up interviews, both women reported that they were currently single heads of household and that they had children. The central office interviewer indicated that it would be very rare for this birth cohort of Korean women to have had children out of wedlock. It is possible that these respondents misunderstood the ACASI question on marital status, marking "single" instead of "widowed," "divorced," or "separated." It is unclear why the correlations are poor for the question on high school attendance, although the Korean-speaking central office interviewer thought that some people might have interpreted the question to mean graduation from high school.

Conclusions

The small sample size leaves us with suggestive rather than definitive results. Findings from this pretest suggest that monolingual English-speaking field interviewers can successfully administer an ACASI health survey to older Korean-speaking respondents. Use of an automated data collection system prevented problems associated with skip patterns that are often encountered in SAQs. In addition, the Korean language ACASI program coupled with cellular phone access to one Korean-speaking central office interviewer obviated the need for multilingual field interviewers. The Korean-speaking central office interviewer was crucial in the recruitment of subjects. However, after gaining entry into the household, the field interviewers were able to accomplish most of the remaining tasks by themselves. This is remarkable, given the absence of English-speaking proxies in the majority of recruited households.

While these results are promising, this study is not without its share of problems. Some problems could be avoided in future iterations of research; others resist solution. Accessing potential respondents was hindered by physical and cultural barriers. Many apartment buildings in the Washington area use intercom systems to alert residents to visitors at the building's entrance. It was not possible to use the central office interviewer to communicate to residents via the apartment building's intercom system. Moreover, many Washington area residents are hesitant about admitting strangers into their residence, and this tendency may be more pronounced among the elderly. Thus, field interviewers found it difficult to recruit subjects living in secured apartment buildings. Not surprisingly, most respondents were unfamiliar with RTI. It might have been helpful to provide more information on the bona fides and research activities of RTI in the lead letter that was sent to subjects prior to contact by the field interviewer. In addition, the concept of surveying populations was not necessarily familiar to this group of subjects. The contact and recruitment rates may have been increased by providing in the lead letter a more detailed description of this survey research and how the findings will be used.

It is important to note that several subjects in this study were very old and had hearing problems. It was difficult for some respondents to hear the questions asked during the ACASI interview; some noted that the rapid pace of speech and short time interval between questions made responding difficult. While we clearly cannot change the auditory capabilities of subjects, we can change some characteristics of the ACASI system to increase acceptability for elderly respondents. As misconfigured in this study, the ACASI system delivered sound through only one side of the headset at one predetermined volume level. Allowing respondents to vary volume and providing sound through both sides of the headset would be a logical improvement. In addition, several respondents commented on the youthfulness of the recorded voice. Because age and experience are positively valued in Korean culture, using an older woman to record the interview at a slower rate of speech would be preferable.

There may be other operational changes that would make the ACASI system more user-friendly. Several older respondents expressed anxiety about using the computer, and some wanted to quit when they experienced a problem. In addition, after the study was completed, we identified some irregularities due to software problems in the ACASI program that prevented all subjects from hearing the full set of instructions in Korean. In particular, the function and location of the enter key and the role of function keys in

changing answers proved problematic for several respondents. A software bug had caused this part of the Korean instructions to be omitted from some of the interviews. Clearly, more careful testing of programs and equipment could improve matters, but this will require greater involvement of bilingual staff in these tasks. (In contrast to our earlier multilingual ACASI research with Spanish-speaking respondents, none of the authors had any knowledge of Korean.) In addition, we believe that simplifying instructions on how to use the computer and ensuring that all subjects receive adequate instruction are needed. It may also be desirable to design a simplified version of the keyboard, perhaps with a limited number of color-coded keys. While difficulties were reported by some, it should be pointed out that all recruited respondents completed the interview, with very little help provided over the cellular phone by the Korean-speaking central office interviewer.

This study may represent one of the worst-case scenarios for monolingual English-speaking interviewers to conduct ACASI interviews with a non-English-speaking population. The Koreans interviewed for this study were very old, and many had hearing problems; few had English-speaking proxies available in the households to help with recruitment and interview procedures; and surveys were foreign activities for them.

Our results from multilingual ACASI interviewing of Spanish-speaking respondents (Turner et al., 1995) provide a useful perspective on our results from Koreans. We found many fewer problems accessing respondents, and, as shown in the final column of Table 4, we achieved higher rates of consistency in responses. Interviews, which included the same questions asked of the Korean cohort but spoken in Spanish, were completed with 30 of 34 pre-identified Spanish-speaking respondents (2 subjects refused to participate, 1 was unlocatable, and 1 was ineligible). Our Hispanic sample was much younger (mean age of 36.6 years) than the Koreans, and they reported more variability in English speaking and reading skills. We found that virtually all of the Spanish-speaking respondents (97%) reported no difficulties using the ACASI technology. Comparison of responses given during the ACASI interview to those given during a Spanish telephone follow-up interview showed higher rates of consistency than we found for our elderly Korean sample (see Table 4). In the Hispanic sample, there were no differences in consistency of responses between those with "good" English skills and those with "poor" English skills. All respondents who reported that they currently smoked gave the same answer at reinterview. Over 90% of responses were consistent for questions concerning the number of children, educational level, and use of eyeglasses or contact lenses, and more than 80% were consistent for questions on ever having tested for blood cholesterol, working for pay for one or more months, and having visited the dentist within the past year.

Given the exploratory nature of our studies to date and the small sample sizes we have used, our results remain suggestive. However, the evidence to date does suggest that multilingual ACASI surveying is feasible, and in many cases, it can yield data that are quite consistent with that obtained by interviewers who speak the respondents' native languages. It is also clear that at a minimum, hearing loss will pose some problems when using this technology with the elderly. Future investigations using larger and more diverse samples of non-English-speaking respondents will help us better delineate the survey conditions under which it will be most appropriate to use multilingual ACASI interviewing.

Reference

Turner, C. F., Rogers, S. M., Hendershot, T. P., Miller, H. G., & Thornberry, J. P. (1995). Improving representation of linguistic minorities in health surveys: A preliminary test of multilingual audio-CASI. Unpublished manuscript.

Impact of ACASI on Reporting of Male-Male Sexual Contacts: Preliminary Results From the 1995 National Survey of Adolescent Males

Charles F. Turner, Leighton Ku, Freya L. Sonenstein, and Joseph H. Pleck

Overview

Since 1988, the National Survey of Adolescent Males–1 (NSAM-1) has tracked the sexual, contraceptive, and AIDS risk behaviors of a national probability sample of young men who were aged 15 to 19 in 1988. This longitudinal research effort gathered follow-up data from this cohort in 1990–91, and it is conducting a new wave of data collection in 1995. Data from the prior rounds of this survey have provided a unique resource for studying changes in behaviors that are central to our understanding of the transmission of sexually transmitted diseases, HIV risk, and unintended pregnancy among adolescents and young adults in the United States.

Data from prior rounds of the NSAM-1 have also presented perplexing methodological puzzles. Reporting of male-male sexual contacts, for example, has occurred at rates that are considerably lower than would be predicted based upon the retrospective reports of national samples of adult men. Similarly, analyses of the stability of reporting of male-male contacts between 1988 and 1991 yielded evidence of considerable rescission (e.g., respondents reporting some male-male contact in 1988 but reporting no lifetime contact in 1991).

These considerations and our desire to increase the actual and perceived privacy of the interview context have motivated us to adopt and evaluate the impact of audio computer-assisted self-interviewing (ACASI) in the 1995 round of the NSAM. ACASI technology permits respondents to listen on headphones to spoken questions (recorded digitally) and/or to read questions on the computer screen of a laptop personal computer. They respond directly on a computer keyboard. This permits respondents to answer confidential questionnaires even if they have limited reading abilities.

The results presented in this paper are properly termed preliminary. They report the results for approximately the first 45% of the NSAM-2 cases (N = 928). The major focus of our attention is an experiment embedded within the survey. NSAM-2 respondents were randomly assigned to receive the most sensitive sections of the NSAM either in a paper self-administered questionnaire (SAQ) or an ACASI interview.[1] While data from these initial interviews do not provide national estimates of male-male sexual contact, it is possible to use these data to assess whether different response distributions were obtained from those respondents who received ACASI rather than paper SAQs.

In the following pages, we briefly review the history of the NSAM and past problems with NSAM estimates of the prevalence of male-male contact that motivated our decision to use ACASI. We will then provide an overview of the design of the 1995 rounds of NSAM-1 and NSAM-2 and our methodological experiment. We will conclude by presenting some of the preliminary results from this experiment.

1988 and 1991 NSAMs

The 1988 and 1991 rounds of the NSAM were the first surveys of the sexual and HIV-risk-related behaviors of probability samples of young men in the U.S. conducted since 1979. The NSAM surveys were originally designed to complement the National Center for Health Statistics's National Survey of Family Growth (Cycle IV—1988) for women of childbearing ages, although NSAM provides richer data about sexual activity and risk behaviors. Both waves of NSAM covered similar topics, with varying degrees of emphasis and reference time periods. Core topics included demographic characteristics; family background; educational history and aspirations; and a detailed history of sexual, contraceptive, and HIV-related behaviors, including

Preparation of this paper and the research reported herein were supported by grant R01-HD30861 from the National Institutes of Health National Institute of Child Health and Human Development. In preparing this draft, the authors have benefitted from contributions by Harvey Zelon, NSAM Survey Director; Frank Mierzwa, NSAM Regional Supervisor; and James Chromy, Chief Scientist in the Research Triangle Institute's Statistical Research Division. Mr. Mierzwa prepared a report on NSAM training and field operations that we have summarized in this presentation.

Charles F. Turner is Director of the Program in Health and Behavior Measurement at the Research Triangle Institute in Rockville, Maryland. Freya L. Sonenstein is Director of the Population Studies Center, and Leighton Ku is a Senior Research Associate at the Urban Institute, Washington, DC. Joseph H. Pleck is a Professor of Human Development and Family Studies at the University of Illinois at Urbana-Champaign.

[1] Since in-house processing and keying of paper SAQs are slower than for ACASI computer files, we have attempted to ensure that the data reported in this paper reflect equivalent interview periods. ACASI data include all ACASI interviews received at the Research Triangle Institute through Wednesday, May 10, 1995. Paper SAQ data reflect all paper SAQs on hand at the Research Triangle Institute on Thursday, May 11, 1995.

detailed histories of first and last intercourse and information about recent partners; use of alcohol and drugs; attitudes about condom use; gender role attitudes; and knowledge about sex, AIDS, and contraception (Sonenstein, Pleck, & Ku, 1991).

Interviews for the 1988 NSAM were carried out between April and November 1988 with a nationally representative sample of 1,880 never married, noninstitutionalized men 15 to 19 years old, living in households. Between December 1990 and May 1991, 1,676 follow-up personal interviews were conducted. Thanks to a strong tracking and field effort, we reinterviewed 89% of the original respondents (not including 11 respondents who died between 1988 and 1990). In 1991, we found that 1988 respondents who were lost to follow-up tended to be slightly older, but more importantly, there was no attrition bias by race or by behavioral outcomes, such as sexual activity or condom use (Ku & Kershaw, 1991).

The 1988 and 1991 NSAM waves have provided a rich body of data for studying behaviors that involve risk of HIV transmission as well as unintended pregnancy, drug dependency, and other phenomena. (See, for example, work by the present authors included in the references.) There are, however, some perplexing puzzles in these data. As discussed below, the most troubling of these involve prior NSAM measurements of male-male sexual contact—the most common mode of HIV transmission.

Measurements of Male-Male Sexual Behaviors

The 1988 NSAM estimated that only 2.1% of males aged 15 to 19 reported any male-male contact during their lifetime, with 1.4% reporting male-male oral or anal sex. Only 0.3% of the 1988 NSAM sample reported male-male oral or anal sex during the 12 months prior to the survey. Furthermore, longitudinal analyses comparing reports in the 1988 and 1991 NSAMs have revealed considerable inconsistency in the reporting of lifetime contacts between 1988 and 1991 (Ku, Sonenstein, & Pleck, 1992a). Only 11 of the 30 men who indicated any lifetime male-male oral or anal intercourse in the 1988 NSAM acknowledged these contacts in the 1991 follow-up.

Besides these troubling discrepancies over time, the prevalence estimates obtained in the 1988 and 1991 NSAMs are extremely low when viewed in the context of the retrospective reports given by adult men about their adolescent behaviors (see Turner, Danella, & Rogers, 1995). Previous analyses of the 1970 Kinsey data estimated that 20.3% of adult men in 1970 had some reported contact with another male in their lifetime; 8.4% of men only reported having contacts before age 14, while 11.9% reported some contacts after age 14 and 6.7% of men reported some male-male sexual contacts during adulthood (Fay, Turner, Klassen, & Gagnon, 1989; Turner, Miller, & Moses, 1989).

If there were no major changes in the patterns of same gender sexual behaviors between 1970 and 1988, these results would imply that the prevalence observed in the 1988 NSAM should be much higher than 2%. Additional analyses of the 1970 Kinsey Institute data set have provided a more precise indication of the extent of this discrepancy. Turner, Danella, and Rogers (1995) report that in addition to the 8% of men who reported experiences prior to age 14, 81% of males reporting same gender sexual contacts also reported that their first contact occurred before age 19; 52% of males reported that their first contact occurred before age 15. These estimates would suggest that the 20% estimate for male-male contact in the 1970 Kinsey Institute Survey should translate into an estimate of between 10% and 16% for a study that interviewed a sample of 15 to 19 year olds.

The 1992 National Health and Social Life Survey obtained results that are roughly consistent although slightly lower than those reported for the 1970 Kinsey survey. The National Health and Social Life Survey, however, did not ask about male-male sexual contacts before puberty. For contacts after puberty, the investigators found that 9.1% of American men in 1992 reported having male-male contacts after puberty and 4.9% reported such contacts after age 18 (Laumann, Gagnon, Michael, & Michaels, 1994). These results would imply that (a) 4.2% of men had male-male contacts that were restricted to adolescence, (b) some portion of the 4.9% reporting adult contacts began such contacts in adolescence, and (c) an unknown percentage had only prepubertal contacts.

Turner, Danella, and Rogers (1995) speculated that the most plausible hypothesis for the divergent results in the NSAM is that the reporting of same gender experiences is considerably more sensitive for adolescents than for adults, and hence the reporting biases inherent to these measurements will differ. This is plausible for two reasons. First, adolescents will be reporting on relatively recent behaviors, while adults may be providing retrospective reports of behaviors that have become less sensitive with the passage of time. Qualitative research on reporting of sexual behaviors suggests that reporting of very recent sexual events is particularly sensitive (Spencer, Faulkner, & Keegan, 1988). Similarly, a large experimental study of the effects of offering a private interviewing mode on the reporting of illicit drug use found that the advantage of the more private mode of administration is most pronounced for reporting of recent behaviors (Turner, Lessler, & Devore, 1992). A second reason to expect divergences in survey estimates is that adolescents are reporting at a time when their own sexual identities may not be well defined, and hence, they may be more fearful of reporting stigmatized behaviors.

These concerns about anomalies in the 1988 and 1991 NSAM measurements motivated our decision to use ACASI in the 1995 wave of NSAM. Our hope was that this technology would attenuate the apparent underreporting bias in prior waves of NSAM and that we would find greater logical consistency over time in our measurements of male-male sexual behaviors. Below, we briefly describe the ACASI technology used in the NSAM and provide a summary of our research design and preliminary findings.

ACASI Technology

In 1991, scientists at the Research Triangle Institute (RTI) developed and field tested a computer-driven technology that administers survey questionnaires in an audio format and records respondents' answers without the intervention of a survey interviewer (O'Reilly & Turner, 1992; Turner, Lessler, & Gfroerer, 1992, pp. 304–305). This process is entirely private—respondents listen to questions through headphones, and they enter answers by pressing labelled keys on a keypad. Development of this technology was spurred by an initial discussion of the feasibility of ACASI interviewing between the first author (Turner) and David Celentano at a meeting of the steering committee for a multisite evaluation of HIV prevention programs (Project Light, 1991, Feb.; Project Light, 1991, May; Turner, 1991).[2] James O'Reilly and Darren DeLoach developed and programmed RTI's initial systems, and this technology was successfully piloted at RTI during the spring of 1992 (O'Reilly, Hubbard, Lessler, Biemer, & Turner, 1994).

ACASI technology offers several important advantages over the paper SAQ methods that were available for the 1988 and 1990 NSAMs and those currently in use by other investigators (see Turner, Danella, & Rogers, 1995). Most importantly, ACASI

1. can be used with any respondent who can hear and speak—it does not require literacy in any language;
2. permits efficient multilingual administration of surveys without requiring multilingual survey interviewers;
3. offers the traditional advantages of computer-assisted survey technologies (i.e., computer-controlled branching through complex questionnaires, automated consistency and range checking, automatic production of data files, etc.); and
4. provides a completely standardized measurement system in which every respondent (in a given language) hears the same question asked in exactly the same way.

Research Design for 1995 NSAM-1 and NSAM-2

The two previous waves of NSAM-1 were conducted as a longitudinal panel survey of males 15 to 19 years old who were first interviewed in 1988. To obtain the best measures of period, age, and cohort effects on sexual and contraceptive behaviors, we have expanded the NSAM into a "staggered prospective multiple cohort study," using the terminology of Fienberg and Mason (1985). Cohorts (or panels) are followed longitudinally, with new cohorts periodically introduced. This design offers many advantages for cohort analysis. Most importantly, it permits longitudinal analysis of age effects and the use of multiple cohorts to help distinguish period and cohort effects (Glenn, 1977).

In 1995, we are conducting the third round of interviews with the original cohort (NSAM-1), and we have added a new cohort (NSAM-2) of young men who are 15 to 19 years old in 1995.

The 1995 NSAM research program will include

1. the third wave of NSAM-1 data collection. This includes the original cohort of young men, who will be about 22 to 27 years old in 1995. Data will be collected using methods from the first two waves: a personal interview with a written instrument and an SAQ.
2. a new primary panel of 15 to 19 year olds (NSAM-2). The general structure of the data collection is similar to that used before with the following methodological improvements:
   a. The coverage for the sample has been extended to include college dormitories and prisons, and the Hispanic oversample has been expanded.
   b. The most sensitive questions are being asked using RTI's ACASI technology.
3. a methodological experimental panel of 15 to 19 year olds. A randomly selected comparison group will be interviewed with the most sensitive questions asked in a traditional, written SAQ (as was done in the 1988 NSAM-1).

The original 1988 NSAM-1 panel oversampled black youth (and effectively oversampled Hispanics). The new NSAM-2 panel of 15 to 19 year olds will oversample black and Hispanic youth because HIV, STDs, and adolescent pregnancy disproportionately burden these communities. In our original design for the 1995 NSAM research program, we proposed the following sample sizes:

Panel	Black	White and Hispanic	Other	Total
Primary	600	593	800	1,993
Experimental	126	124	168	418
Total	726	717	968	2,411

Funding constraints, the considerable expense of screening over 60,000 households to identify a sample of 2,411 households with a 15- to 19-year-old male, and other problems have caused us to reduce the total sample size. We presently anticipate completing approximately 2,000 total interviews in NSAM-2.

[2] In this discussion at the February 28, 1991, meeting of the Steering Committee for Project Light, it was Dr. Celentano (not the first author [Turner]) who suggested investigating the possibility of developing a "audio-CAPI" (audio computer-assisted personal interviewing) system using voice synthesis. Use of digitized (rather than synthesized) voice was subsequently implemented by O'Reilly at RTI in April and May of 1991. During this same period, G. Johnston implemented a Macintosh-based "audio-CAPI" system using a digitized voice at the University of Michigan. Johnston's development of his system probably antedates Celentano's suggestion by some months.

Preliminary Results

Status of Fieldwork

As of May 6, 1995, 42,282 sample lines had been released for screening in NSAM-2, and screening had been completed on 33,126 of these assignments. Of screening assignments, 3.3% (1,386) were found to contain an eligible adolescent male. As of May 6, interviews had been completed with 927 of these eligible respondents.

Field Experience With ACASI

A total of 123 field interviewers were trained in late January and the first week of February on the use of the ACASI software. The field problems using the ACASI hardware have been minimal, given the number of field interviewers working on the NSAM. There have been a few instances (approximately 6) in which a computer problem that occurred in the field could not be solved over the phone. In these cases, a replacement machine was shipped to the field interviewer via Federal Express. We have also replaced approximately 10 Antex Audio Interfaces and a few fraying cables used to connect the Antex box to the computer (Mierzwa, 1995).

Other than these problems, our 138 computers have held up well in the field. Many of the field interviewers hired for the NSAM were inexperienced in using computers, but supervisors report that they have become quite comfortable with the technology. Reports from the field interviewers indicate that most respondents seem to enjoy using the computer, and they find it to be an interesting aspect of the survey. Our survey staff do not know of any respondent who has refused to use the computer. The major recommendation made by our survey team is that future surveys eliminate the external audio interface and associated cables (Mierzwa, 1995).

RTI's new generation of ACASI software does just that. It will run on laptop computers that have integrated sound chips, such as the TI 4000M and IBM Thinkpad 755 series of laptops. With this new system, the field interviewer plugs headphones directly into a port on the laptop. There are no external boxes or cables other than the power cable (see Cooley, Turner, O'Reilly, Allen, & Paddock, in press).

Expectations for ACASI in NSAM

While we embarked upon the experiment of incorporating ACASI into the NSAM in the hope of reducing the apparent underreporting of male-male sexual contacts, our expectations were tempered by two considerations. First, it is entirely plausible (as noted above) that the underreporting in prior rounds of the NSAM could be due to differences in the sensitivity of the reporting of male-male contact by teenage males who may not be fully confident in their sexual identity. Compared to an adult reporting on his teenage sexual behaviors, a teenage boy is both reporting a much more recent behavior and is more likely to be insecure in his interpretation of how that report fits into his sexual identity.

Secondly, the increment in privacy afforded by switching from a paper-and-pencil SAQ to ACASI is not as substantial as that which has induced 2- to 4-fold increases in reporting of sensitive behaviors in other experiments. Our own recent work comparing telephone ACASI (T-ACASI) reports of sensitive behaviors with those obtained in a standard telephone interview (Turner, Miller et al., 1995) and other experimental comparisons of paper SAQs to interviewer-administered survey measurements (e.g., Aquilino, 1994; Jones & Forrest, 1992; Turner, Lessler, & Devore, 1992) involved a much stronger manipulation of privacy than is involved in the NSAM's comparison of measurements obtained using a paper-and-pencil SAQ versus an ACASI interview. We thus embarked upon these preliminary analyses prepared to find relatively modest differences between the measurements obtained in the two experimental conditions in the 1995 NSAM.

Preliminary Estimates of Male-Male Contact

Table 1 presents the results obtained from the first 928 respondents in NSAM-2. The table shows the percentage of respondents who report engaging in each of six types of male-male sexual contact: masturbating another male, being masturbated by another male, insertive oral sex, receptive oral sex, insertive anal sex, and receptive anal sex. The final line of the table shows the results for a composite measure comparable to that previously discussed for the 1988 NSAM-1. This measure indicates whether the respondent reported at least one type of male-male contact.

It will be seen from Table 1 that there are substantial and statistically reliable differences between the reports given in the ACASI interview and in the paper-and-pencil SAQ. Respondents were more than four times more likely to report some male-male contact in the ACASI interview. Although the odds ratios for the individual behaviors vary somewhat (from 2.1 to 5.4) and several are statistically unreliable with our current incomplete sample size of 928, there seems little reason to doubt that ACASI will reduce the underreporting of male-male sex in the 1995 NSAM-2.

Based upon this preliminary analysis, two other observations merit note. First, even with this apparent fourfold increase in reporting of male-male sex, the 4.7% estimate is still considerably below what would be reasonable based upon the retrospective reports of adolescent sexual behaviors made by recent generations of adult men. Secondly, although the sample size is to too small for the result to be statistically reliable (and the preliminary 1995 sample is not appropriate for making population estimates), the paper-and-pencil SAQ in the 1995 NSAM-2 presently yields an (unweighted) estimate of any male-male contact (1.1%) that is half the size of the weighted estimate derived from the 1988 NSAM-1 (2.1%).

Table 1. Estimates of prevalence of different types of male-male sexual contact in a national sample of males aged 15 to 19 in 1995 by mode of data collection: Preliminary results from the 1995 NSAM-2

	Paper SAQ		ACASI			
Measurement	Estimated %	Base N	Estimated %	Base N	OR	p
Ever masturbated another male	1.1	176	2.3	731	2.07	0.29
Ever been masturbated by another male	0.6	176	3.0	730	5.44	0.03
Ever had insertive oral sex with another male (your penis in his mouth)	0.6	176	2.5	730	4.42	0.07
Ever had receptive oral sex with another male (his penis in your mouth)	0.6	176	2.1	730	3.67	0.13
Ever had receptive anal sex with another male (his penis in your rectum or butt)	0.0	176	1.2	730	—[a]	0.05
Ever had insertive anal sex with another male (your penis in his rectum or butt)	0.6	176	1.6	729	2.93	0.23
Any male-male sex[b]	1.1	176	4.7	728	4.26	0.01

NOTE: Preliminary data from the first 928 cases of 1995 NSAM-2. p-values are those for likelihood ratio chi-square for fit of independence model to the two-way table of mode by reporting of behavior.

[a]Odds ratio cannot be calculated due to zero denominator.

[b]Composite measure of any male-male sex is derived from the six individual measurements. Cases with missing data for any of the six behaviors were excluded from the analysis of the composite measure.

Patterns of Response

While the results presented in Table 1 suggest that ACASI has effects that are larger than we anticipated, there remain several other concerns worthy of analysis. Most important, perhaps, is an assessment of the impact of ACASI on the internal consistency of responses. One fear in moving to computer-based self-interviewing is that any apparent increase in a low-prevalence behavior may reflect nothing more than an increase in the error rates (e.g., respondents accidentally pressing 1 ["yes"] when they meant to respond "no").

To provide some initial evidence, we examined the patterns of response to the six male-male sex questions. If one ignores item nonresponse, there are 64 (2^6) possible combinations of answers respondents could have given to the six male-male sexual behavior questions. Actual sexual behaviors, however, are much more structured; some patterns should be rare or nonexistent. For example, we would not expect to find large numbers of males reporting no sexual activity other than insertive anal intercourse. To the extent that such structure is lacking in the ACASI response patterns, one may legitimately wonder whether the ACASI responses are meaningful. Similarly, a proliferation of response patterns with very small frequency counts might encourage suspicion that random errors in keying were inflating the prevalence estimates.

The reports obtained under ACASI do evidence substantial structure. Only 20 of the 64 possible response patterns are observed in the ACASI data, and the patterns observed most frequently conform well to our expectations as to the patterning of male-male contacts among adolescents. The most frequent pattern observed was, of course, the reporting of no experience with each of the six male-male sexual behaviors (694 respondents). The next most frequent patterns were (a) only masturbation (13 respondents), (b) masturbation plus oral sex (5 respondents), and (c) masturbation plus oral and anal sex (9 respondents). An additional 3 respondents reported only oral sex, and 2 respondents reported oral and anal sex but not male-male masturbation. These response patterns account for all of the ACASI reports but 2.

Conclusion

While we are anxiously awaiting completion of the second half of the survey, our preliminary data strongly suggest that ACASI is diminishing (but not eliminating) the underreporting of male-male sexual contacts. The evidence we can adduce at this time also suggests that this result is unlikely to be due to random measurement errors.

References

Aquilino, W. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. Public Opinion Quarterly, 58, 210–240.

Cooley, P. C., Turner, C. F., O'Reilly, J., Allen, D. A., & Paddock, R. E. (in press). Audio-CASI: Adding sound to a

computer-assisted interviewing system—hardware and software considerations. Social Science Computer Review.

Fay, R. E., Turner, C. F., Klassen, A., & Gagnon, J. H. (1989). Prevalence and patterns of same-gender sexual contact among men. Science, 243, 338–348.

Fienberg, S., & Mason, W. (1985). Specification and implementation of age, period and cohort models. In W. Mason & S. Fienberg (Eds.), Cohort analysis in social research: Beyond the identification problem. New York: Springer-Verlag.

Glenn, N. (1977). Cohort analysis. Beverly Hills, CA: Sage.

Jones, E. F., & Forrest, J. D. (1992). Underreporting of abortion in surveys of U.S. women: 1976 to 1988. Demography, 29, 113–126.

Ku, L., & Kershaw, J. (1991). Notes on attrition bias and weighting in Wave 2 of the National Survey of Adolescent Males. Unpublished manuscript, The Urban Institute, Washington, DC.

Ku, L., Sonenstein, F., & Pleck, J. (1992a). Patterns of HIV risk and preventive behaviors among teenage men. Public Health Reports, 107, 131–138.

Ku, L., Sonenstein, F., & Pleck, J. (1992b). The association of AIDS education and sex education with sexual behavior and condom use among teenage men. Family Planning Perspectives, 24, 100–106.

Laumann, E. O., Gagnon, J. H., Michael, R. T., & Michaels, S. (1994). The social organization of sexuality: Sexual practices in the United States. Chicago: University of Chicago Press.

Mierzwa, F. J. (1995, May 12). Preliminary field report on experiences with audio-CASI in NSAM. Unpublished memorandum, Center for Survey Research, Research Triangle Institute.

Miller, H. G., Turner, C. F., & Moses, L. E. (Eds.). (1990). AIDS: The second decade. Washington, DC: National Academy Press.

O'Reilly, J., Hubbard, M., Lessler, J., Biemer, P., & Turner, C. F. (1994). Audio and video computer assisted self-interviewing: Preliminary tests of new technologies for data collection. Journal of Official Statistics, 10, 197–214.

O'Reilly, J., & Turner, C. F. (1992, March). Survey interviewing using audio-format, computer-assisted technologies. Presented at the Washington Statistical Society, Washington, DC.

Pleck, J., Sonenstein, F., & Ku, L. (1990). Contraceptive attitudes and intention to use condoms in sexually experienced and inexperienced adolescent males. Journal of Family Issues, 11, 294–312.

Pleck, J., Sonenstein, F., & Ku, L. (1991). Adolescent males' condom use: The influence of perceived costs-benefits on consistency. Journal of Marriage and the Family, 53, 733–745.

Pleck, J., Sonenstein, F., & Ku, L. (1992). Problem behaviors and masculinity ideology in adolescent males. In R. Ketterlinus & M. Lamb (Eds.), Adolescent problem behaviors. Hillsdale, NJ: Erlbaum.

Project Light. (1991, February 28). Minutes. Meeting of the Steering Committee for the Multisite Trial of Behavioral Interventions to Prevent the Spread of HIV, Washington, DC.

Project Light. (1991, May 23). Draft summary notes. Meeting of the Subcommittee on Measurement, Multisite Trial of Behavioral Interventions to Prevent the Spread of HIV, Baltimore, MD.

Rogers, S. M., & Turner, C. F. (1991). Male-male sexual contact in the U.S.A.: Findings from five sample surveys, 1970–1990. Journal of Sex Research, 28, 491–519.

Sonenstein, F. (1986). Risking paternity: Sex and contraception among adolescent males. In M. Lamb & A. Elster (Eds.), Adolescent fatherhood. Hillsdale, NJ: Erlbaum.

Sonenstein, F., Pleck, J., & Ku, L. (1989a). Sexual activity, condom use and AIDS awareness among adolescent males. Family Planning Perspectives, 21, 152–158.

Sonenstein, F., Pleck, J., & Ku, L. (1989b). Missing data on the missing male: Measuring fertility and paternity behavior among young men. Paper presented at the meeting of the American Statistical Association, Washington, DC.

Sonenstein, F., Pleck, J., & Ku, L. (1991). Levels of sexual activity among adolescent males. Family Planning Perspectives, 24, 100–106.

Sonenstein, F., & Wolf, D. (1991). Satisfaction with child care: Perspectives of welfare mothers. Journal of Social Issues, 47(2), 15–31.

Spencer, L., Faulkner, A., & Keegan, J. (1988). Talking about sex (Publication P.5997). London: Social and Community Planning Research.

Turner, C. (1989). Research on sexual behaviors that transmit HIV: Progress and problems. AIDS, 3, S63–S71.

Turner, C. F. (1991, May 23). Voice-administered CAPI: Memorandum to Subcommittee on Measurement for the Multisite Trial of Behavioral Interventions to Prevent the Spread of HIV. Unpublished memorandum.

Turner, C. F., Danella, R., & Rogers, S. (1995). Sexual behavior in the United States: 1930–1990: Trends and methodological problems. Sexually Transmitted Diseases, 22, 173–190.

Turner, C. F., Lessler, J. T., & Devore, J. (1992). Effects of mode of administration and wording on reporting of drug use. In C. F. Turner, J. T. Lessler, & J. Gfroerer (Eds.), Survey measurement of drug use: Methodological studies. Washington, DC: Government Printing Office.

Turner, C. F., Lessler, J. T., & Gfroerer, J. (Eds.). (1992). Survey measurement of drug use: Methodological studies. Washington, DC: Government Printing Office.

Turner, C. F., Miller, H. G., & Moses, L. E. (Eds.). (1989). AIDS, sexual behavior, and intravenous drug use. Washington, DC: National Academy Press.

Turner, C. F., Miller, H. G., Smith, T. K., Cooley, P. C., Catania, J. A., Rogers, S. M., & Von Colln, L. (1995). Preliminary evaluation of a new technology for telephone surveys of sexual and other sensitive behaviors. Unpublished manuscript, Research Triangle Institute, Rockville, MD.

Special Populations, Sensitive Issues, and the Use of Computer-Assisted Interviewing in Surveys

Joseph Gfroerer

Introduction

This session addresses three separate but related areas of interest to health survey researchers: surveys of special populations, collecting data on sensitive issues, and the use of computer-assisted interviewing (CAI). Each of these three topics has received increasing attention over the past few years.

Surveys of special populations have become more important as health planners and policy makers require data to address the health care needs of specific population subgroups. Although sometimes these data can be obtained from ongoing broad-based surveys, often it is necessary to conduct separate surveys targeting special populations. Available ongoing surveys may not have sufficient numbers of cases in the population of interest or may not even include the population in their universe (e.g., most household surveys exclude the homeless). Many of the same methodological issues apply whether the special population is surveyed in a limited study or as part of a larger survey with broader coverage.

Surveys of sensitive issues have become more prevalent, in part to provide critical data describing emerging health problems such as AIDS and drug abuse. These health problems require survey researchers to collect data on sensitive topics such as sexual behavior and illegal activities. This requires new and innovative methods for ensuring the validity of the data collected in surveys.

The third major area discussed in this session is the use of CAI, which is rapidly becoming the standard for all large-scale surveys. Many studies have documented the data collection, processing, and quality benefits of CAI. As costs continue to decline and improved technology and software become available, we continue to see the conversion of many surveys to CAI, including, for now at least, computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), computer-assisted self-interviewing (CASI), and audio computer-assisted self-interviewing (ACASI).

The emergence of each of these three major areas in health survey research has occurred somewhat independently and for unrelated reasons. However, it is difficult to discuss them separately. The methodological questions that they generate often overlap, and the answers given by research in these areas are sometimes meaningful only in the context of the others. For example, some topics that are sensitive for some special populations may not be sensitive for others. An example of this is alcohol use, which is thought to be sensitive for underage youths but not for adults. Similarly, some types of CAI (e.g., CASI) may work very well for most populations but may be problematic for some special populations, requiring specialized methods. And respondents' willingness to report sensitive data may vary with different types of CAI.

During this period of rapid conversion of surveys to CAI, it is critical that methodological research include studies of the benefits and effects on data that CAI has in surveys of special populations and of sensitive topics. The six papers presented in this session add important new knowledge to this growing body of research and will be useful to government agencies conducting surveys of special populations and on sensitive issues and that are considering the use of CAI.

The Importance of Privacy in a Survey of Drug Use Among Youths

The paper by Horm, Cynamon, and Thornberry addresses the collection of sensitive data from a special population, youth. The results support previous research showing that youths are more likely to report their drug use when the threat of disclosure to parents is minimized (Turner, Lessler, & Devore 1992; Schutz, Chilcoat, & Anthony, 1994). Horm et al. show that the presence of a parent during the interview inhibits reporting in two different surveys, the National Household Survey on Drug Abuse (NHSDA) and the Youth Risk Behavior Survey (YRBS). Furthermore, the YRBS, which used extraordinary efforts to enhance privacy (i.e., the use of cassette recorders and headphones), found higher rates of reported drug use than the NHSDA. The most dramatic difference between the two surveys is seen for the reporting of cigarette use among 12 and 13 year olds when their parents were present. In this case, the YRBS estimate is seven times greater than the NHSDA estimate. In the NHSDA,

Joseph Gfroerer is Chief of the Prevalence Branch, Office of Applied Studies, Substance Abuse and Mental Health Services Administration, U.S. Department of Health and Human Services, Rockville, Maryland.

this estimate was probably the most vulnerable to under-reporting because the cigarette questions were interviewer administered. It is not surprising that many 12- and 13-year-old smokers would be reluctant to tell an interviewer about their smoking when their parent is nearby.

I have some comments regarding the comparison of the YRBS to the NHSDA. Besides the mode of interview, there were other differences between the two surveys that could have affected youths' reporting of drug use. First of all, the question wordings were not exactly the same in the two surveys. Secondly, the context of the questions was different. In the YRBS, a small set of drug use questions was embedded in a larger set of questions about health behaviors. The NHSDA questionnaire focused almost entirely on drug use and asked detailed questions about patterns of use of various licit and illicit drugs. A third difference between the two surveys was the affiliation of the interviewers. The NHSDA was conducted by a private contractor, while the YRBS was conducted by Census Bureau interviewers. Youthful respondents may have been more willing to reveal sensitive information to the YRBS interviewers because of the reputation of the Census Bureau in collecting and protecting confidential data. The level of trust may also have been greater in the YRBS because the interviews were done as a follow-up to the National Health Interview Survey (NHIS), giving respondents increased familiarity with the survey and rapport with the interviewer. A fourth factor potentially affecting youths' reporting was the procedure used in the YRBS in which the questions in the booklet shown to parents were placed in a different order from those on the cassette tape. This additional step taken to assure youths that their parents would not be able to ascertain their answers to survey questions was not done in the NHSDA. On the other hand, the NHSDA procedures provided greater anonymity than the YRBS. In the NHSDA, no names were collected, and address information was kept separate from the interview responses. In the NHIS, identifying information such as names, addresses, and Social Security numbers were collected. Independent of the mode effect, all of these factors could account for part of the difference in reporting sensitive behaviors.

The paper concludes that the level of privacy had a much smaller effect in the YRBS than in the NHSDA. I would suggest that further analysis might help support this conclusion. In the collapsed data presented (see Horm et al.'s Table 4), the effect of privacy on reporting in the YRBS is understated because of the way the categories are grouped. The more detailed presentation (see Horm et al.'s Table 2) shows consistently higher reports of sensitive behaviors when parents were not home than when parents were close by. Finally, it should be pointed out that the privacy scales used in the YRBS and NHSDA were different.

These caveats do not diminish the validity of the main conclusion of the paper, that greater privacy enhances youth reporting of sensitive behavior. This conclusion leads to the question of which modes and settings provide the greatest privacy. Clearly, an interviewer-administered questionnaire gives inadequate privacy if observers are present. The procedures used in the YRBS seem to enhance privacy better than self-administered questionnaires (SAQs), but more research is needed. And since most youth surveys are school based, we should also be concerned about ensuring privacy in that setting when collecting sensitive data. Comparison of estimates of youth drug use from data collected in classrooms and from households has suggested that youths are more likely to report drug use in the classroom (Gfroerer, 1993), with differences greatest for the youngest students. However, it may also be true that some behaviors are sensitive in the household setting but not in the classroom and vice versa.

Preliminary results from the 1994 NHSDA again show the importance of privacy and the impact of mode on the reporting of sensitive behavior by youths. In 1994, the NHSDA converted the tobacco questions from interviewer administered to self-administered. This apparently has greatly improved measurement of smoking (and also smokeless tobacco) among teenagers. The rates obtained from the new SAQs are nearly twice that obtained from the interviewer-administered mode, resulting in estimates that are consistent with YRBS estimates for 12 and 13 year olds, 14 and 15 year olds, and 16 and 17 year olds.

The Effect of Incentives and Interview Mode in a Survey of Women 15–44

The paper by Duffer, Lessler, Weeks, and Mosher includes some interesting results from the National Survey of Family Growth pretest. The pretest was designed to study the effect of incentives on response rates and also the effect of ACASI on the reporting of sensitive behaviors in a special population, women aged 15 to 44. The results indicate that modest incentives appear to improve response rates in minority populations. The data also suggest that privacy enhances the reporting of sensitive behaviors, although the evidence is weak. Privacy from family members was achieved by conducting some interviews at a neutral site. Rates for a few of the sensitive behaviors were higher for interviews conducted at the neutral site than for interviews conducted at home. Within the in-home sample, respondents were given a second opportunity to report abortions in a short ACASI follow-up after completion of the main CAPI interview. With ACASI, the percentage of women reporting one or more abortions increased but only from 23.6% to 27.1%.

Another conclusion in the paper was that the incentive may have increased the reporting of sensitive behaviors in households. However, it is possible that coverage differences could explain this finding. First of all, the differences between the in-home, no incentive and the in-home, incentive rates are generally very small. Secondly, the higher rates of reported sensitive behaviors among women who received incentives may have been the result of increased participation rates in groups with high rates of sensitive behaviors. In other words, the incentive payment

brought into the survey groups of women who had higher abortion rates and had more sexual partners than others.

Surveying Non-English-Speaking Populations

The paper by Hendershot, Rogers, Thornberry, Miller, and Turner shows how ACASI can be adapted to surveys of non-English-speaking populations. This research demonstrates the potential that the ACASI technology has to improve the coverage of national surveys as well as our ability to survey special populations previously inaccessible. This study points out that it is sometimes preferable and even mandatory to design data collection methods specifically to address the needs of individual subpopulations. However, this raises the question of how such specialized procedures could be incorporated into a large, national survey that would only encounter these special situations occasionally and without prior knowledge. The potential impact on a large, national survey can be seen from interview completion statistics from the NHSDA. The NHSDA utilizes an English and a Spanish questionnaire, but persons who cannot complete the interview in either of these languages become nonrespondents. In 1992, about 0.6% of sampled persons could not be interviewed due to a language barrier other than Spanish. However, this component of nonresponse was 1.8% in Los Angeles and was only about 0.3% outside of the largest metropolitan areas.

Surveying Elderly Populations

The paper by B. Kahana, Kercher, E. Kahana, Namazi, and Stange provides evidence that reliable health data can be obtained from elderly community residents, despite mild cognitive deficits.

This paper is the only one of the six that does not involve the study of new interviewing technologies such as ACASI. This raises a question: What would the results of the study have been if the interviews had been done using ACASI? With the lack of respondent-interviewer interaction provided in ACASI, elderly persons with mild cognitive impairment might have given less reliable data. Whether or not this would have been the case, we should be concerned about the possibility that new data collection technologies may not be appropriate for all populations. Another good example of a special population for which ACASI would not work well is the hearing-impaired population, many of whom are elderly. CASI and ACASI may be less desirable in surveys of elderly populations, regardless of any impairment, simply because the elderly are less comfortable with computers. All of these potential pitfalls lead to the consideration of specialized procedures tailored to specific subpopulations.

Converting a Survey From Paper and Pencil to CAI

Shepherd, Hill, Bristor, and Montalvan's paper nicely outlines many of the issues that may need to be confronted when converting an ongoing survey from paper-and-pencil interviewing (PAPI) to CAPI. The paper demonstrates that careful attention to all aspects of the survey field and data-processing components is necessary. It also shows some of the improvements in data quality that can result from conversion to CAPI.

The situation described in this paper is somewhat unusual because the survey changed methodology during the middle of a wave of data collection. Phase 2 of the National Health and Nutrition Examination Survey (NHANES) is a nationally representative sample conducted over a 3-year period (Ezzati et al., 1992). CAPI was introduced during the second year of Phase 2. Thus, Phase 2 estimates will be based on two different interviewing methodologies assigned to the sample nonrandomly. Although we would generally not expect conversion to CAPI to have much effect on responses because both PAPI and CAPI are interviewer administered, there were some modifications made to answer categories, question presentations, and probes used in the CAPI version for some variables. This could possibly complicate later analyses involving these variables.

A more common situation would be the conversion of a continuous, annual survey from PAPI to CAI. The CPS recently converted to CAI, and the NHIS is planning to convert next year. These situations involve not only the operational issues described by Shepherd et al., but also analytical issues such as trend measurement. When trend measurement is important, the conversion to CAI should be made using a split sample using both old and new methods (Nicholls & Matchett, 1992).

In the conversion to CAPI, it was decided that the NHANES household screener needed to remain as hard copy because of concerns that a CAPI screener might have an adverse impact on the screening response rate. No evidence was presented or references given to support this decision. Research on this issue would be useful because CAPI has some important potential benefits for household screening. In particular, it allows more complex sample person selection algorithms to be implemented.

Comparison of ACASI Versus SAQ in a Survey of Sexual Behavior Among Young Men

The paper by Turner, Ku, Sonenstein, and Pleck involves the study of the impact of ACASI in collecting sensitive data from a population of young adult males. The authors describe another example of the conversion of a survey from PAPI to CAI. In this case, the conversion was done at the start of a new wave of data collection and included a well-designed experiment to test the effect of ACASI on the

reporting of sensitive behaviors. The initial results appear to strongly support the use of ACASI to improve reporting of sensitive behaviors. The paper also provides an excellent discussion of the advantages of ACASI over SAQs, including the ability to employ complex skip patterns and the ability to maintain privacy for illiterate respondents.

One of the advantages of ACASI cited by the authors is that it provides a completely standardized measurement system. This may not always be the case, however. For some special populations (e.g., the hearing impaired), ACASI may not be advisable or even possible. Some respondents will either require or just prefer different modes of providing data. Rather than lose these respondents, a better strategy might be to allow respondents to choose the mode they are most comfortable with. Designers of surveys may have to make a choice between standardized measurement and high response rates.

Summary and Recommendations

In summary, these six papers cover a number of issues concerning surveys of special populations, collecting data on sensitive issues, and the use of CAI. These papers will be useful to government and private agencies that are planning to collect these types of data and that are considering CAI. The results of these studies support a number of recommendations for the sponsors of surveys and for researchers conducting methodological studies:

1. Ensure privacy when collecting sensitive data, especially for youths in households. SAQs provide more privacy than interviewer-administered questionnaires, but ACASI and similar techniques seem to provide the most privacy.
2. Numerous studies have documented the data-processing, analytic, and other advantages of CAI over PAPI. With the decreasing costs and improving technology of CAI, it should be given first consideration as the mode for every large survey.
3. When analyzing results across surveys or waves of ongoing surveys that used different modes, it should not be assumed that the estimates are comparable. This is particularly important for data on sensitive topics.
4. For ongoing surveys where trends are important, conversion from PAPI to CAI should use split-sample designs that provide measurement of the effect of the new method.
5. When designing surveys that will use CAI, consider the different effects that the mode might have on special populations. Conduct field tests to assess these effects. Consider using multiple modes, giving respondents the option of which mode to use.
6. Much of the methodological research has been done for specific purposes and for application with specific surveys. Thus, it is often difficult to find results that are generalizable to other surveys. For example, the NHSDA is probably unique in having such a substantial portion of the questionnaire covering sensitive topics. To implement ACASI for all sensitive items in the NHSDA would require the development of a 45-minute ACASI interview. I am not aware of any research that addresses the question of how long respondents would be willing to cooperate in an ACASI interview. Of course, some methodological issues may be so unique to a particular survey that a special field test designed for that survey is required. Nevertheless, researchers should try to use designs that will generate results more widely applicable.
7. Much of the research on ACASI and other new technologies seems to be overly optimistic about the benefits and advantages of adopting them. More discussion and data on potential problems and limitations of the new technologies is needed in the literature.

References

Ezzati, T. M., et al. (1992). Sample design: Third National Health and Nutrition Examination Survey. Vital and Health Statistics (Series 2, No. 113). Hyattsville, MD: NCHS.

Gfroerer, J. (1992). Overview of the National Household Survey on Drug Abuse and related methodological research. American Statistical Association 1992 Proceedings of the Section on Survey Research Methods.

Nicholls, W. L., & Matchett, S. D. (1992). CASIC issues at the Census Bureau as seen by members of outside panels. Proceedings, 1992 Annual Research Conference.

Schutz, C. G., Chilcoat, H. D., & Anthony, J. C. (1994). Breach of privacy in surveys on adolescent drug use: A methodological inquiry. International Journal of Methods in Psychiatric Research, 4, 183–188.

Turner, C. F., Lessler, J. T., & Devore, J. W. (1992). Effects of mode of administration and wording on reporting of drug use. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), Survey measurement of drug use: Methodological studies (DHHS Publication No. [ADM] 92-1929). Washington, DC: NIDA.

Discussion of Session on Special Populations and Sensitive Issues

Robert M. Groves

Most of the papers in this session address various types of survey measurement errors, arising either because of attributes of persons (as with elderly respondents or non-English-speaking persons) or attributes of questions (as in reports of drug usage and sexual behavior). In many of them, there is also at least an undercurrent of interest in survey nonresponse error. I am sincere in my admiration of the authors' tackling of important problems in their work and believe they benefitted the conference noticeably.

Two of the papers in the session report on an implementation of a new idea without any experimental contrasts. These are papers of the class that present to the field a demonstration that some innovation can work under some circumstances. Over the years, I have rarely read such papers describing the failure of such ideas, and the lack of experimental contrasts heavily limits what inference one can draw from the work. Presence of contrasts opens the researchers to the failure of their favorite solution; absence of contrasts assures its perceived success.

Shepherd, Hill, Bristor, and Montalvan

One such paper is the Shepherd, Hill, Bristor, and Montalvan paper describing the conversion of the National Health and Nutrition Examination Survey (NHANES) to computer-assisted personal interviewing (CAPI). The application makes interesting demands on CAPI (multiple instruments, need for nonforward movement through the instrument). The only contrast available to the authors is the results of NHANES pre-CAPI. The paper clearly reports work in progress. The initial results show reduced need for editing and adjudication steps, a common finding of the literature. I look forward to more empirical evidence on the quality, cost, and timeliness differences between pre-CAPI and CAPI.

Hendershot, Rogers, Thornberry, Miller, and Turner

This paper describes two small pretests of audio computer-assisted self-interviewing (ACASI) on health topics for a population of Hispanic and Korean respondents with very limited English skills. There are two implicit contrasts to the use of ACASI here: (a) use of an in-home or paid translator of an English instrument and (b) use of a Korean- or Spanish-speaking interviewer. A reinterview component provides a contrast with (b). The results suggest more enjoyment and consistency of data (for ACASI relative to native-speaking interviewer) for younger respondents. One clever idea that merits further tests is the use of a cellular phone to link the respondent to someone who speaks Korean or Spanish. I hope the authors continue their work, with more explicit contrasts among alternative designs. I would also recommend that future such small tests use a crossover design that randomly varies the order of the two data collection methods and the use of a reconciliation step.

B. Kahana, Kercher, E. Kahana, Namazi, and Stange

This paper presents an assessment of measurement error from a very different tradition of analysis than most survey researchers follow. The authors discuss thresholds of measurement error indicators that will render a data record excluded from analysis. They investigate the problem of cognitive impairment among the elderly as a source of measurement error and seek some measure that would empirically guide the handling of cases suspected to have large measurement errors. They employ multiple indicators of cognitive deficits and choose three measurement error indicators—accuracy of height and weight estimation, interitem reliability of physical health and psychological well-being, and missing data on health and well-being items. The sample consisted of about 600 persons, with a mean age of 83 years, from three retirement communities.

The perspective taken by this research, therefore, uses measures of hypothesized causes of measurement error and estimates their empirical relationship with different types of estimates of measurement errors on a small number of variables. The conclusion is a null finding: There is little evidence of the relationship of mild cognitive deficits on measurement error.

I have no expertise in the measures of cognitive deficits, but I do worry about the generalizability of these findings and caution readers to consider three limitations of the study: (a) The respondents are in the fourth wave of a panel

Robert M. Groves is a Professor and Research Scientist at the University of Michigan, Ann Arbor, and serves as Associate Director of the National Science Foundation–sponsored Joint Program in Survey Methodology at the University of Maryland, College Park.

survey and have been trained in the respondent role; others without such training may behave differently; (b) only a small number of variables were chosen for error consideration; and (c) the more severely cognitively impaired may have dropped from the panel, and thus there are important selectivity issues in the inference.

Horm, Cynamon, and Thornberry

This paper takes a perspective on the survey interview quite compatible with early notions of it as a social interaction with a well-defined purpose. The social psychology of the interview situation has stimulated concern about influence from the interviewer and from other persons privy to the conversation about the survey questions. The paper focuses on effects of the presence of other persons on survey responses. In this, it joins past research observing tendencies of respondents to provide answers more acceptable to those persons when they are present.

This paper can be considered a case study in this regard, focusing on youth; the presence of parents; and the measurement of sex, drug, and alcohol experiences. The implicit hypothesis is that the young respondents believe their parents are opposed to their engaging in certain behaviors in these realms and will alter their responses when they suspect their parents may learn of their response. The reader should note that this hypothesis, like all social desirability hypotheses, is itself a function of the true values on the survey measures. That is, only those youths who, for example, have smoked marijuana need fear that their truthful answer may yield negative sanctions from their parents.

The paper is an evaluation of the use of audiotape recorders and paper-and-pencil answer sheets to shield all but the respondent from the survey questioning on these topics. Like most research in this area, the criterion is the percentage of persons reporting the socially undesirable behavior—more is better.

There are several design features in the work that complicate inference. First, there is no controlled comparison of the audiotape method (portable audiocassette tape player with headphones [PACTAPH]) to some other mode (except indirectly using National Household Survey on Drug Abuse [NHSDA] data). Second, this was an observational study, and the presence of parents was an uncontrolled aspect of the interviewing situation. That means that if youths who didn't tend to engage in the risky behaviors also tended to have their parents present, the measurement of parental presence was confounded with true differences in behavior. Third, the measure used for parental presence was a nested one: whether or not the parent was at home; if so, whether they were in sight of the child; if so, whether they could see the answers. This may have had some multidimensionalities that caused problems of interpretation. For example, on several characteristics, it appears that the youths whose parents were home but out of sight were less similar to those whose parents were in sight than to those whose parents weren't home at all. Why is this important? It may reveal quite different reasons for parental presence. The absence of a parent during the interview may have had little to do with the parenting style on issues of sex and drugs. However, it is more likely that among those parents at home, ones who exert close supervision on those issues may have chosen to be within sight of their child during the interview.

The best approach for these problems, in my belief, is to attempt an explicit repair of the lack of randomization of parental presence. That is, the analyst of these data needs to incorporate into the model those correlates of parental presence that could also be correlates of the real behavior being measured. This might be done with a logit model, given most of the dependent variables, using a continuous measure of age and all the predictors of the dependent variable available, and then measuring the marginal effect of parental presence. This attempts to create statistical controls that assure that youths with parents present actually do behave as those without parents present, conditional on the other predictor variables.

The paper concludes that the audiotape protocol provides a level of privacy necessary for maximum disclosure of sensitive behaviors, but it still measures an effect of parental presence, one that is stronger than noted in the paper when attention is focused on degree of closeness of the parent among those with a parent within sight. It does appear that the effects are smaller than those found in the NHSDA, which used oral questioning by an interviewer and written answers by the youth.

This finding that some of the effects of social desirability are removed by a device offering full privacy is unfortunately common in the survey methodological literature on social desirability. It resembles, for example, findings of tests of the randomized response and item count methods. In this case, one interpretation of the finding is that the visible presence of a parent makes more salient the norms they promulgate in the family. Regardless of the perceived likelihood of the parent learning the response, the youth might use the norm saliency to change the rigor of their memory search about the risky behavior and respond in the negative.

This research might profit from attempts to more directly measure the nature of parental influence on youth response as a way to interpret the effects of different survey techniques.

Duffer, Lessler, Weeks, and Mosher

Duffer, Lessler, Weeks, and Mosher also examine a survey measuring sensitive topics but add nonresponse as well as measurement error to their attention. The authors report on a pretest of the National Survey of Family Growth at six sites. With six sites, they were burdened with an investigation of three separate design factors: site of the

interview (in-home or neutral site), incentive to respond, and use of CAPI or ACASI. There are no explicit hypotheses in the paper, but I infer that the authors conceptualized incentives as acting to increase cooperation, not necessarily reduce measurement error, and the other two factors as acting principally on measurement error.

The design is not a factorial design on the three factors. It might be best to think of it as a balanced two-factor design of incentives ($0 or $20) by mode (CAPI or ACASI) among women interviewed at home. In addition, they have another treatment group—$40 incentive, neutral site, CAPI.

At this level of description, it seems that inference on the effects of the $20 incentive and ACASI could be made for in-home interviews. However, another complication arose— the belief that within–primary sampling unit (PSU) variation on incentives would contaminate the experiment. Hence, they assigned all in-home cases in three PSUs to the no incentive and all in the other three PSUs to the $20 incentive. That in itself merely increased the sampling variance of estimates of experimental effects if the PSUs were randomly assigned and they showed evidence of equivalence. My Figure 1 shows the assignment of different treatments to the six PSUs in the design, with Xs denoting that no sample cases were assigned to a particular cell.

What are the findings? In late 1993, they were able to locate about 80% of the sampled women who were interviewed in the 1991 National Health Interview Survey (this suggests that locating problems deserve scrutiny equal to that of compliance, given contact). Although they don't present estimates of standard errors, it appears that the $20 incentive group has lower refusal rates (the paper also presents response rates, but the other components of nonresponse should not be affected by the experimental stimulus).

Is there any complication with the interpretation of these differences? Unfortunately there is. Their Tables 1 and 2, if I read them correctly, illustrate that the PSUs receiving the $20 incentive had a different demographic mix of women than did the no incentive PSUs. The $20 incentive PSUs had more black and Hispanic women, many fewer nonblack, non-Hispanic. They only had about 40% of the women with incomes over $20,000 versus about 70% in the no incentive PSUs. Should we care about these demographic differences? Only if they portend different base cooperation rates. There is some evidence that the attributes are related to cooperation—but the evidence varies by incentive group, probably because the groups are also differentially sensitive to the incentive treatment. This is complicated material, impossible to interpret correctly given the data in the paper, but probably interpretable with a different analysis using a logit model controlling on the covariate variables that seem to affect cooperation and differ across the two sets of PSUs. Given the weight of prior evidence on incentives, I have little doubt that the incentive effect will be found to be positive; I do doubt that the 7.5% reduction in the refusal rate would apply to the full survey.

Figure 1. Assignment of treatments to PSUs in the Duffer et al. design

Three experimental variables
1. CAPI versus ACASI
2. $0 incentive (PSUs 1–3), $20 incentive (PSUs 4–6), $40 incentive (PSUs 1–6)
3. In-home versus neutral site

	CAPI		ACASI	
	Neutral			
Neutral	In-home	site	In-home	site
$0 incentive	PSUs 1–3	X	PSUs 1–3	X
$20 incentive	PSUs 4–6	X	PSUs 4–6	X
$40 incentive	X	PSUs 1–6	X	X

The best interpretation of the refusal rate on the neutral site, $40 incentive CAPI cases is that that package of design features seems similar on raw cooperation to that of the $20 incentive, in-home group. (It might have been interesting to see a comparison of the $40 group with the $20, in-home, CAPI group, just to eliminate one other source of difference in the design packages.)

The interpretation of the authors regarding the incentive effects on refusals is that the incentive group will bring in more lower-income minority women, and consequently, the number of reported abortions will be higher. I understand this to be a comment on nonresponse error properties; that is, the true values of the respondents added to the pool given the incentives reflect more abortions. This interpretation becomes important later.

With regard to measurement error effects of the design features, the paper presents a comparison of in-home versus neutral site interviewing, but the comparison cannot be easily interpreted. The only neutral site interviewing was the $40 incentive CAPI group. When a comparison of abortion rates is made incorporating incentives and site (i.e., breaking the in-home into no incentive and $20 incentive), the authors find an increase in reports for the $20 incentive group. Recall, however, the conclusion that nonresponse differences could account for this. They also note a similar set of higher reports for the neutral site group, but that group also contains the effects of a $40 incentive.

Thus, the comparisons of rates of reporting sensitive attributes are confounded with nonresponse differences that the authors found earlier. Again, one is tempted to propose multivariate models that attempt to control for differences across treatment groups on key attributes of the respondents.

Turner, Ku, Sonenstein, and Pleck

Another part of the research on ACASI, a companion to the Hendershot et al. work, examines reporting of male-male sexual contact in a panel of young men aged 15 to 19.

It is a split-sample randomized experiment, with half using a paper self-administered form and half using ACASI. The interpretation of differences between the two samples is that more reporting of male-male sexual activity reflects reduced measurement error.

We should note seriously the authors' observation that if social desirability and threat are the major influences on measurement errors in these data, then both the self-administered questionnaire (SAQ) and ACASI offer considerable privacy to the respondent. In neither case is immediate or overt revelation of the respondents' answers possible, either to the interviewer or to others present. Small differences between modes might be expected. However, preliminary estimates show large odds ratios reflecting higher reports of male-male sexual contact in ACASI than in the SAQ.

Logical extensions to this work that merit attention are (a) multivariate analysis testing whether younger men show larger effects of using ACASI, following the theoretical assertions in the paper; (b) analysis of the effects of the presence of others in the interviewing situation; and (c) the effect of using ACASI on other sensitive measures.

In short, this session was filled with diverse designs and insights into the processes that produce measurement errors in special populations and for sensitive items. Several of the papers propose solutions that may reduce some of those errors, and they deserve our careful scrutiny in planning future studies.

Discussion Themes From Session 4

Lu Ann Aday, Rapporteur, and Mary Grace Kovar, Chair

A common theme through all the papers presented in this session is the interaction between respondent and questionnaire characteristics and the mode of interviewing. The papers seem to be assessing the general assertion that the newer technological advances offered by computer-assisted interviewing (CAI) strategies would improve the data quality in surveys conducted on sensitive topics or with special populations. From this general focus, the discussion centered around three more specific topics: (a) the interaction between mode and respondent characteristics, (b) the interaction between mode and question characteristics, and (c) methodological issues in assessing the interaction between respondent characteristics, question characteristics, and mode of data collection.

Mode and Respondent Characteristics

The discussion focused on the ways in which audio computer-assisted self-interviewing (ACASI) might be used to address special response problems that occur with two specific populations that present particular issues: adolescents and non-English-speaking respondents.

An assertion was made that ACASI is a general strategy that can be used with almost any population. However, there were also those who suggested that this assertion might be too optimistic, especially concerning elderly populations, whose members may have mild cognitive deficits or may simply be uncomfortable with such technology. Respondents with hearing impairments may also find difficulty with ACASI, and these also include a disproportionate number of older respondents. The data presented seem to indicate that younger respondents—that is, adolescent males and non-English-speaking younger respondents—prefer the ACASI approaches. Older non-English-speaking respondents seem to be less comfortable with this technology.

A second general theme related to respondent characteristics was also observed across several of the papers: Adolescent respondents seem to have major concerns about privacy. Younger respondents seem to prefer ACASI forms to other modes for sensitive topics, even other forms that provide privacy, such as self-administered paper-and-pencil versions of the same instrument. The positive benefits of ACASI were also observed during in-home interviews, for which it was noted that younger respondents were bothered even by the presence of a parent or other supervising adult.

Interviewing of older respondents raises some additional concerns related to the cognitive capacity of the respondents. Assessing the cognitive capacity of very old respondents is an important problem in the increasing number of surveys that focus on what are described as "old old" respondents (generally over 75 years of age). The methods for determining cognitive capacity are not well formulated or generally agreed upon. It was noted that more objective criteria are needed to determine when to proceed with the interview and when to request a proxy.

Overall, however, there was concern that the mode effects by respondent characteristics were not well integrated into existing theoretical models of survey response. For example, in his discussion of the Horm, Cynamon, and Thornberry paper, Groves indicates that the influence of the presence of an adult during an interview on a sensitive topic is consistent with the general social psychology of the interview situation, but during the discussion, concerns were raised that both the mode effects and the effects of observers in the interview session were not clear-cut and that more theoretical work on the issues of privacy, mode, and respondent characteristics was definitely indicated. Another theoretical issue that was raised related to the themes in Session 2 regarding the interaction between ethnicity and educational attainment on the understanding of questions would definitely impact the utility of ACASI, since the content of the question would not be altered unless translation and back translation were conducted.

Mode Effects and Question Characteristics

The most consistent theme related to mode effects and question characteristics was consistency and accuracy of reporting. These issues were particularly of concern in surveys of adolescents that require reporting of sensitive behaviors. The principal question was whether higher reports of sensitive behaviors necessarily imply greater accuracy in the absence of externally verifiable criteria. This was seen as a particularly critical issue for adolescents

Lu Ann Aday is at the School of Public Health at the University of Texas in Houston. Mary Grace Kovar, Senior Health Scientist, is at the National Opinion Research Center's Washington, DC, office.

because, as the discussants for this session note, there may be demand characteristics in the interview setting that would encourage overreporting as well as underreporting, depending on the composition of the interview setting. For example, in the presence of peers, it might be more important to overreport certain behaviors, whereas with parents or other authority figures present, underreporting may be more desirable.

The work by Bradburn, Sudman, and Associates (1979) on asking sensitive questions was the standard used by some. During the discussion, Bradburn offered three criteria that might be applied, depending upon the direct availability of a relevant criterion source: (a) use of validation data at the individual level, such as records of drunk driving arrests; (b) use of aggregate external data, such as sales tax data or arrest reports of adolescents; and (c) seeding the sample with verified cases and use of the resulting data to adjust for over- or underreporting in the entire sample. Mathiowetz also reported on the quality of data on the smoking behavior of adolescents. She reported on the fairly well-known results of work by Evans, Hansen, and Mittelmark (1977) using the "bogus pipeline" strategy of indicating that self-reports will be validated and actually collecting saliva or hair samples. Evans et al.'s data indicate that the reports given by adolescents who believed their reports were being validated had the same level of accuracy as those that were actually validated. However, either way, these methods require some collection of specimens, whether or not they are actually tested. It is the standard in school studies to validate self-reported behavior. A difficulty arises with studies that are conducted in respondents' homes or in other settings. Whereas adolescents in school settings and adults in clinic settings feel constrained to cooperate with validation, less success has been obtained in other settings where the demand for cooperation is less evident. Moreover, when the interviews are conducted by telephone, such validation procedures are not possible, and when home visits are made after a telephone contact, the refusal rates are often quite high (Warnecke, Langenberg, Gruder, Flay, & Jason, 1989).

Another type of demand characteristic that might affect accuracy of reporting is the use of incentives. In the Duffer, Lessler, Weeks, and Mosher paper, incentives were used as an inducement to increase cooperation in the study. However, the question is whether the incentives also created an implicit demand for respondents to report more accurately or even to overreport. The question is whether there is likely to be an interaction between incentive amount and reporting. One suggestion was that sensitivity analyses might be done to evaluate the effects of varying levels of incentives versus other changes, such as shifts in overall cooperation.

Assessing Interactions Between Respondent Characteristics, Question Characteristics, and Mode of Data Collection

One of the difficulties encountered in this session was that the mode effects attributed to ACASI and the methodological approaches discussed in this session could not easily be separated from the characteristics of the respondents and the questions. Most of the data reported were descriptive or demonstrational and were not evaluated in experimental formats. Thus, a general observation made by many of the participants was the need for theoretically driven experimental evaluation of some of the questions that these papers raise. Several suggestions were made for other analytic strategies that might help sort out some effects. Included in this category were multivariate analyses in which some of the confounding effects due to respondent or question characteristics or interview environment could be controlled while the mode effects were assessed.

More methodological research is needed in ongoing research before substantive inquiry. This may require federal and private funders to allow more of this kind of preliminary work as part of the funding for major surveys. It was noted that this kind of preliminary methodological research is costly and is often eliminated from study budgets by reviewers.

The issues of data validity in surveys on threatening topics and the consistency of data reported over time are still unresolved, and validation will continue to be a problem, particularly as the number of face-to-face studies decreases due to cost. However, even where these new technologies for self-administered interviews are implemented in face-to-face settings, without some kind of standard, the confounding between mode and validity due to the nature of the topics or characteristics of the respondents may not be sorted out.

Themes to Be Pursued in Future Research

1. These new technologies need to be evaluated in the context of the same kind of rigorous theoretical framework that has been used for previous survey methodology, as exemplified in the work of Dillman (1978), Sudman and Bradburn (1982), and Groves (1989), among others.
2. A difficult aspect of this research is that much of it is being conducted as pilot studies or in conjunction with ongoing work in which the priority is not the methodological effects on the data. As a result, some of the papers presented in this session reflect some rudimentary hypotheses often tested on small samples and select population groups. Thus, it is difficult to generalize from them to broader applications. However, this is the next step in assessing some of these seemingly promising technologies, such as ACASI.
3. It is important that agencies sponsoring these kinds of studies include some support for more rigorous assessment of the effects of new methodologies that are emerging from this small but very important pilot research.
4. Research in this area needs to take into consideration both variable and systematic error in the design of

ongoing studies comparing innovative computerized strategies with more traditional approaches.

5. Also, the research must be able to separate the effects of these new technologies from the influences resulting from respondent and question characteristics and from the context in which the interview is taking place.

References

Bradburn, N. M., Sudman, S., & Associates. (1979). Improving interview method and questionnaire design: Response effects to threatening questions in survey research. San Francisco: Jossey-Bass.

Dillman, D. (1978). Mail and telephone surveys: The total design method. New York: Wiley.

Evans, R. I., Hansen, W. B., & Mittelmark, M. B. (1977). Increasing the validity of self-reports of smoking behavior in children. Journal of Applied Psychology, 62, 521–523.

Groves, R. M. (1989). Survey errors and survey costs. New York: Wiley.

Sudman, S., & Bradburn, N. A. (1982). Asking questions: A practical guide to questionnaire design. San Francisco: Jossey-Bass.

Warnecke, R. B., Langenberg, P., Gruder, C. L., Flay, B. R., & Jason, L. A. (1989). Factors in smoking cessation among participants in a televised intervention. Preventive Medicine, 18, 833–846.

Integrating Survey and Other Data

The five papers in this session together with the discussion address the topic of data integration and its utility in enhancing data quality, analytic capability, or both. The papers are fairly eclectic and illustrate both the advantages and problems that result from this strategy. The common themes of this session are well described by Andersen in his critique as securing a match between data sets on the time when the data are collected, the place where the data are obtained, and the unit of observation. Collectively, the papers indicate that these integration processes balance significant gains and significant costs in data quality, timeliness, and flexibility in analysis. Thus, their use is justified only in instances in which costs of other procedures are high or the data needed can be obtained in no other way.

Computer Matching of Medicare Current Beneficiary Survey Data With Medicare Claims

Franklin J. Eppig Jr. and Brad Edwards

The Medicare Current Beneficiary Survey (MCBS) is a continuous panel survey of Medicare beneficiaries in the United States.[1] Interviews are conducted three times a year with a sample of about 12,000 to collect information about the use and cost of health care services. All household interviews are conducted in person by computer-assisted personal interviewing (CAPI). In addition to the usual features of computer-assisted interviewing (CAI), the MCBS CAPI design includes extensive abstracting of documents, especially explanations of Medicare benefits and statements that reflect private insurance coverage for specific events. Because a critical MCBS goal is to estimate payments by various sources for services that Medicare covers (but does not pay in full), for each reported service, the survey attempts to identify the total charge (or the Medicare-approved charge, for participating providers) and the Medicare payment in order to determine the amount for which the Medicare beneficiary or other payment sources are responsible.

When a Medicare enrollee receives a Medicare-covered service, the medical provider submits a claim for payment directly to Medicare.[2] Even if the provider refuses to accept assignment and requires the patient to pay for the service and seek reimbursement from Medicare, the provider is still required to submit the claim for payment. After Medicare claims are processed for payment by Medicare's fiscal agents, they are forwarded to the National Claims History database (NCH). An estimated 97% of the Medicare claims are posted to NCH within a year, even with processing delays related to adjudication of disputed claims. Thus, NCH data provide a nearly complete picture of the Medicare utilization and reimbursements for all but the 6% of the Medicare population enrolled in capitate plans.

However, the NCH database contains no information about other payment sources for events covered by Medicare, nor does it include items/events that Medicare does not cover (such as most prescribed medicines or physician services for persons covered by Part A but not by Part B). The survey interviewer asks the beneficiary about all events and attempts to collect data on all payment sources and amounts for those events. The best estimate for total expenditures for all events is derived from a combination of the two data sources.

Objectives for Matching MCBS Survey Data and Medicare Claims

Matching survey data with claims data has two primary objectives: to adjust for underreporting of the use of health care services by survey respondents and to fill gaps and make corrections in the survey expenditure data.

Underreporting health care events has been a subject of considerable interest in the survey literature. Memory of specific events is prone to decay, and even the best efforts to probe respondents' memories and to assist their recall are unlikely to boost reporting to desirable levels, particularly for events that are not very salient and for recall periods that are very long. A person level comparison of survey-reported events with events in the Medicare claims can identify events that the respondent may have forgotten. Other events may be difficult or impossible for the respondent to report, not because of memory limitations, but because of the way the events are experienced. For instance, laboratory services may be classified as events in their own right, but the respondent may never be conscious of them—it's a mystery to the patient what happens to the blood once it's drawn. The Medicare records system, however, treats laboratory services like other events and services, so it is a better source for these "hidden" event categories.

Franklin J. Eppig Jr. is with the Health Care Financing Administration in Baltimore, Maryland. Brad Edwards is with Westat, Inc., in Rockville, Maryland.

[1]The Medicare program is a federal health insurance program for people 65 or older and certain disabled people. Approximately 34,000,000 Americans are enrolled in Medicare. Medicare Hospital Insurance, Part A, covers inpatient hospital care, inpatient care in a skilled nursing facility following a hospital stay, home health care, and hospice care. Medicare Medical Insurance, Part B, helps pay for doctors' services, outpatient hospital care, diagnostic test, durable medical equipment, ambulance services, and many other health services and supplies.

[2]This is not true for Medicare beneficiaries who are enrolled in capitate plans. Because their services are not provided on a fee-for-service basis no claim is submitted to Medicare for payment. As a result, Medicare administrative claims databases do not capture utilization and expenditures for medical services provided through capitate arrangements. In 1992 about 6% of those in the Medicare population were members of capitate plans.

Survey respondents experience even more difficulty in reporting expenditures for medical care than they do in reporting the occurrence of health care events. This is not surprising, especially given the complexity of the current health care financing systems in the United States. The survey respondent may be the best source for information on out-of-pocket payments, but the Medicare program is likely to be the best source for information on Medicare payments. For some events, such as inpatient hospital stays, Medicare and the provider may be the only sources for expenditure data because Medicare payments (under the Diagnostic Related Group [DRG] system) are not related to charges. Matching the survey events with the Medicare claims also allows us to check the respondent's reported expenditure data and to fill gaps when the respondent does not know the charges or the payment sources or amounts for covered services.

MCBS Matching Strategy

The first step in matching survey-reported events to Medicare claims is the association of all Medicare claims with a given sampled person. The MCBS design accommodates person level accumulation of Medicare claims data through its use of the Medicare health insurance claim number (HICN). The HICN appears on every Medicare claim submitted for payment and is the key to collecting all of a sampled person's Medicare claims. Since the MCBS sample is drawn from the Enrollment Data Base, the HICN for each sampled person is known prior to the start of field operations.

MCBS interviewers verify the sampled person's HICN during the initial interview using the HICN from the Enrollment Data Base. This circumvents the problems of misreporting and incorrect transcription associated with the collection of the HICN in the field. Having the correct HICN for each sampled person means that a sampled person's Medicare claims can be extracted from the NCH with complete accuracy.

A potential problem with using the HICN to capture an individual's Medicare claims is that a Medicare enrollee's HICN can change. For example, if an individual is entitled to Medicare benefits under both his or her own and a spouse's health insurance account, the HICN may change with the death of the spouse. The MCBS staff track claim number changes using internal Health Care Financing Administration (HCFA) files. This allows MCBS staff to capture all of an individual's Medicare claims regardless of claim number changes.

The next step is to determine the extent of overlap between the survey-reported events and claims data, which requires event level matching of survey data and claims data. Matching survey-reported data to Medicare claims at the event level is significantly more difficult than person level matching. Unlike the HICN at the person level, no data element or combination of data elements provides a consistent and reliably reported basis for conducting event level matches. Discrepancies in the reporting of the same event can occur because of differences in the perspective of the parties or the faulty recollection of specific details of events by respondents. The MCBS relies on Medicare explanation of benefits forms, insurance statements, and other receipts to assist the respondent's memory whenever possible (and as a source of other data elements, such as the claim control number, that were never stored in respondent memory). Often, however, the unaided memory of the respondent is the only source available for event details.

There are several other reasons for the lack of a consistent set of data for event matching. First, the MCBS does not capture a consistent set of variables for the different types of service. For example, the MCBS does not collect total charges or reimbursements for inpatient hospital events, since Medicare beneficiaries usually don't know this information. However, event total charges is a key match field for other survey event categories. Similarly, the MCBS does not capture date-of-service information for prescription drugs, home health events, and "other" medical expenses, but the date of service is a key match field for all other types of service. Second, there are different file layouts and different data elements on the Medicare claims for different service types. Third, for certain classes of beneficiaries (e.g., end stage renal disease [ESRD]) and certain repeat service situations, Medicare claims contain aggregate monthly billing information instead of event level data.

Differences in the categorization of medical services between the Medicare claims and the survey further complicate event level matching. The Medicare claims are essentially organized by type of provider, whereas the type of service categories used in the MCBS are more closely related to the way in which individuals think about the medical care they receive (see Figure 1). In matching the survey event to the Medicare claims data, MCBS staff

Figure 1. Comparison of Medicare claims categories with MCBS event categories

Medicare claims categories	MCBS event categories
Inpatient hospital	DU — Dental
Skilled nursing facility	ER — Emergency room
Hospice services	IP — Inpatient hospital
Home health agency	OP — Outpatient hospital services
Outpatient hospital	MP — Medical provider services
Part B physician/supplier	PM — Prescribed medicine
	HF — Home health services—friend
	HP — Home health services—prof.
	OM — Other medical
	IU — Institutional utilization
	SD — Separately billing doctors
	SL — Separately billing labs

frequently must match a Medicare claim category with multiple MCBS event categories and vice versa.

There are only 6 claims categories versus 12 MCBS event categories. Some of these discrepancies are readily explained. For example, dental services are not included in the claims list because Medicare does not cover most dental services. One of the most noteworthy categories missing from the claims list is emergency room services. In the Medicare claims system, emergency room services that are immediately followed by an inpatient stay are included in the DRG for the inpatient stay and thus are not associated with any separate charges or claims. Emergency room visits that stand alone are classified as outpatient services.

Event level matching is actually a series of matches between different categories of Medicare claims and MCBS service types. In conducting these matches, MCBS staff employ different match algorithms depending on the data elements available for the particular event categories being matched. The sequence of the matches is arranged so that the most similar MCBS event and Medicare claims categories are compared first (see Figure 2).

Each match algorithm employs a hierarchy of match criteria that are progressively less restrictive. For example, reported doctor visits are initially compared with claims data by doctor name, date of service, and total charge. If there is no exact match, the algorithm checks for a match on physician name and date of service or on total charge and date of service. If there is still no match, the program looks for an exact match on physician name and total charge with the date-of-service match relaxed to within a week. Thus, the match algorithms not only link a survey event and Medicare claim, but also indicate the strength of the link.

MCBS staff designed the match algorithms to allow survey-reported events to be linked to multiple Medicare claims and vice versa. There are several reasons for this. First, multiple links are often valid. For example, a survey-reported doctor visit may be linked to both a Medicare claim for physician services and a Medicare claim for lab services connected with the visit. Second, sometimes a stronger match occurs later in the series of matches than the initial, weak match. For example, a survey-reported doctor visit may have a weak match to a Medicare Part B physician/supplier claim and a strong link to a Medicare Part B outpatient claim. MCBS staff use the match strength indicator to resolve situations in which the multiple matches are logically inconsistent.

Figure 2. Overview of event category matches conducted during event level matching

Matches between similar service types
- IP to inpatient hospital
- MP, OM, SD, SL to Part B physician/supplier
- OP to outpatient hospital
- IU to SNF claims
- DU to Part B physician/supplier claims
- ER to outpatient hospital
- HF & HP to home health agency claims

Match between less similar service types
- ER to inpatient hospital claims
- OP to inpatient hospital claims
- IU to inpatient hospital claims
- IP to SNF claims
- IP to outpatient hospital claims
- OP to Part B physician/supplier claims
- MP, OM, SD, SL to outpatient hospital claims

Our strategy can be contrasted with a more probabilistic approach, such as that used by National Medical Expenditure Survey (NMES) for matching Medical Provider Survey data with household-reported data (Cohen & Carlson, 1994; Felligi & Sunter, 1969; Newcombe, 1988). Although many elements of the match process are comparable between the two surveys, for MCBS we did not assign a weight to the outcomes of the matching rules. Rather, the rules were arrayed in hierarchical fashion, reflecting the strength of the matches for each event category and across categories. Stronger matches were accepted before weaker matches for the same event.

A major concern in matching data from the two sources is potential double counting of medical events. MCBS staff have sought to minimize situations in which it is unclear whether an unmatched survey-reported event and an unmatched Medicare claim represent the same event or two different events. Such ambiguities were minimized by conducting the event level match within the data for each person. After organizing the data on a person basis, there are four possible outcomes: (a) a 100% match of the survey-reported events and Medicare claims; this does not present any reconciliation problems; (b) a 100% match of survey-reported events with unmatched Medicare claims; this does not present any reconciliation problems if we assume that the unmatched Medicare claims represent forgotten utilization additive to the sampled person's reported utilization; (c) a 100% match of Medicare claims with unmatched survey-reported events; this does not present any reconciliation problems if we assume that the unmatched survey-reported events are for non-Medicare services, unless the sampled person has reported that Medicare was a source of payment for the service; and (d) there are both unmatched Medicare claims and unmatched survey events; here there is a reconciliation problem.

MCBS staff attempt to address the fourth outcome by classifying unmatched survey events and unmatched claims into discrete service categories and determining whether the unmatched events and claims are in mutually exclusive categories. For example, an unmatched survey-reported dental visit and an unmatched Medicare inpatient hospital claim would be considered mutually exclusive and therefore classified as two separate events. The HCPCS[3] codes on the Medicare Part B physician/supplier claims are used to

[3]Codes that contain procedure specific information at several levels using the American Medical Association's Common Procedure Terminology (CPT) for physician services, HCFA codes for supplier services such as ambulance, and local codes that vary by carrier.

classify Medicare claims into a number of discreet subcategories. With this finer classification scheme, MCBS staff can be more precise in determining whether survey events and Medicare claims are mutually exclusive.

Event Level Match Results for 1992 Data

The first calendar year of MCBS utilization and expenditure data is 1992. Interviewers completed the collection of these data in August 1993. In June 1995, matching activities for most event types are essentially complete, and imputation activities for missing data are in progress. The post-matching file contains more than 300,000 events. Raw match results for the 1992 data by survey event type are presented for four major event classes in Table 1. Nearly one-half of the events are unmatched, and the proportion of false negatives is unknown. The difference between the minimum and maximum number of events is about 26% across these four event types, though it is only 11% for inpatient stays (which are among the most salient types of events for survey respondents) and it is 0% for hospital emergency room visits, since the Medicare system does not have that category as an event type in its own right.

Table 2 presents the results of our review of the unmatched claims and survey events at the person level to identify unmatched events, which must be nonduplicative (i.e., additive) because the individual did not have both unmatched survey events and unmatched claims. We were able to reduce the difference between the minimum and maximum number of events from 26.3% to 16.7% across these four event types.

It is informative to review the effect of the matching process on the expenditure data. For three event types, Table 3 presents the expenditure information as it looks after the match (but before imputation for missing data and editing for inconsistent data) by data source: administrative (i.e., Medicare claims) data or survey data. An event is classified as reported in both sources if it matches and has total charge (or Medicare-allowed charge) and at least some payment data from both sources. In the second group, an event is found in the administrative data that either does not match any survey event or that matches a survey event that has no reported dollars. In the third group, we see the opposite: a survey-reported event with dollars but either no matched event in the administrative data or a matched event with no dollars. The fourth group represents events for which dollars are missing from both sources.

For about 60% of the inpatient stays, expenditure data exist only in the administrative data. Most Medicare beneficiaries are unable to report any dollars associated with hospital stays that are covered by Medicare. For the other two event types shown in Table 3, medical provider visits and hospital outpatient department visits, about three-fourths

Table 1. MCBS raw match results

	A Matched survey-reported events	B Unmatched survey-reported events	C Unmatched claims	Maximum A + B + C	Minimum A + (B or C, whichever is greater)	Difference	
Hospital inpatient	2,853	1,474	493	4,820	4,327	493	(11.4%)
Medical provider	87,862	35,416	44,628	167,906	132,490	35,416	(26.7%)
Hospital outpatient	16,507	7,456	9,499	33,462	26,006	7,456	(28.7%)
Emergency room	1,160	1,030	—	2,190	2,190	0	(0.0%)
Total	108,382	45,376	54,620	208,378	165,013	43,365	(26.3%)

Table 2. MCBS match results after determining which nonmatches cannot be duplicates

	A Matched survey-reported events	B Non-duplicate survey-reported events	C Non-duplicate claims	D Unknown survey-reported events	E Unknown claims	Maximum A + B + C + D + E	Minimum A + B + C + (D or E, whichever is greater)	Difference	
Hospital inpatient	2,853	278	41	1,196	452	4,820	4,368	452	(10.3%)
Medical provider	87,862	11,254	3,009	24,162	41,619	167,906	143,744	24,162	(16.8%)
Hospital outpatient	16,507	2,311	537	5,145	8,962	33,462	28,317	5,145	(18.2%)
Emergency room	1,160	360	—	670	—	2,190	2,190	0	(0.0%)
Total	108,382	14,203	3,587	31,173	51,033	208,378	178,619	29,759	(16.7%)

Table 3. Preliminary distribution of source-of-expenditure data for three event categories

Group	Administrative data	Survey data	No. events	%
Hospital inpatient stays				
1	Reported	Reported	467	9.7
2	Reported	Missing	2,879	59.7
3	Missing	Reported	234	4.9
4	Missing	Missing	1,240	25.7
Total			4,820	100.0
Medical provider events				
1	Reported	Reported	74,505	44.4
2	Reported	Missing	57,985	34.5
3	Missing	Reported	13,302	7.9
4	Missing	Missing	22,114	13.2
Total			167,906	100.0
Hospital outpatient events				
1	Reported	Reported	9,843	29.4
2	Reported	Missing	16,163	48.3
3	Missing	Reported	2,771	8.3
4	Missing	Missing	4,685	14.0
Total			33,462	100.0

of the expenditure data is in the first two groups; that is, most of the events have dollars reported in both sources or in the administrative data alone. This reflects the dominance of the claims data in the MCBS design, even for those covered services for which many survey respondents are able to report expenditure data. The survey design focus is on amounts that are not covered by Medicare and on noncovered events.

It should be noted that Table 3 is based on preliminary data. Through additional editing and imputation, we expect some events will move from the top three groups into the fourth group and some events may move into different categories. However, even at this interim stage, the table shows how relatively dependent the MCBS is on administrative data (the Medicare claims) as opposed to survey data, at least for these three services that are covered by Medicare. In contrast, a similar analysis of the final data from the 1987 National Medical Expenditure Survey (NMES; a household-based survey that collected records from a sample of the medical providers reported by the household respondents and then matched these data to survey-reported events) showed a much higher proportion of total expenditure data reported by household respondents (Cohen & Carlson, 1994). This difference is expected, given the basic design differences between MCBS and NMES.

NMES reported the effects of the matching on estimates of total medical expenditures. We are unable to compare MCBS directly with NMES on this score, because the MCBS was not designed to produce independent estimates from administrative and survey data. However, we can compare (unweighted) data for the dollars on the average claim with dollars on the average survey report for the three event types. Table 4 shows that for hospital events (both inpatient and outpatient) in the first group (expenditures reported in both sources), the average survey report is much higher than the average claim. This reflects the effect of the Medicare program rules governing allowed charges for covered services. On the other hand, dollars for hospital stays reported by the survey respondent but not matched to a claim (the third group) are lower than the average claim amounts in the other groups. These inpatient stays may include a number of events that are more properly classified as outpatient services, including many surgical procedures.

Conclusions

Although matching survey data with Medicare data can introduce a number of ambiguities, the process improves estimates by increasing the amount of utilization and enhancing the accuracy of expenditure information. It reduces the need for imputation of missing data; through matching, we are able to supply total charges and at least some payment amounts by source for 86.4% of events in several major categories. Further research on MCBS match rates could be extraordinarily useful for informing decisions about optimal reference period lengths and for designing improved instruments, editing processes, and imputation strategies. We encourage future investigations of match rates by interviewer and respondent characteristics, proxy versus self-report, type of insurance coverage, length of panel experience, use of respondent records, Medicare claims service category, and Medicare fiscal agent.

Table 4. MCBS matching: Comparing dollars on Medicare claims and survey reports (unweighted data)

	Medicare dollars	MCBS dollars	No. events	Average $ claim	Average $ survey
Hospital inpatient stays					
	Reported	Reported	467	$6,508	$8,110
	Reported	Missing	2,386	$6,435	—
	No claim	Reported	234	—	$3,332
	No claim	Missing	1,240	—	—
	Reported	No survey-reported event	493	$5,833	—
Medical provider events: Reimbursement					
	Reported	Reported	74,505	$85	$89
	Reported	Missing	13,357	$71	—
	No Claim	Reported	13,302	—	$75
	No Claim	Missing	22,114	—	—
	Reported	No survey-reported event	44,628	$89	—
Hospital outpatient events: Reimbursement					
	Reported	Reported	9,843	$202	$353
	Reported	Missing	6,664	$201	—
	No claim	Reported	2,771	—	$139
	No claim	Missing	4,685	—	—
	Reported	No survey-reported event	9,499	$181	—

References

Cohen, S., & Carlson, B. (1994). A comparison of household and medical provider reported expenditures in the 1987 NMES. Journal of Official Statistics, 10, 3–29.

Felligi, I., & Sunter, A. (1969). A theory of record linkage. Journal of the American Statistical Association, 64, 1183–1210.

Newcombe, H. (1988). Handbook of record linkage: Methods for health and statistical studies, administration, and business. New York: Oxford Medical Publications.

Aging in Manitoba: Integrating Survey and Administrative Data

Betty Havens

Introduction

The Aging in Manitoba Longitudinal Studies (AIM) began in 1971 with a random sample (N = 4,803) stratified by age and gender using a small area probability sampling frame of both community- and institutional-dwelling Manitobans aged 65 or over. A second cross section of seniors was surveyed in 1976 (N = 1,302), and a third cross-sectional sample (N = 2,877) was interviewed in 1983. Also in 1983–84, the panel of survivors from 1971 and 1976 (n = 2,401) was reinterviewed. They were reinterviewed again in 1990 along with the survivors from 1983 (n = 3,223). The interviews collect information on sociodemographic and social psychological characteristics, physical and mental health status and functioning, economics, leisure activities, care and support networks, and consumption of services. The interview data have been linked to the full spectrum of health services utilization from the administrative databases of the Manitoba insured services and of the services delivered directly by Manitoba Health for the period 1968 through 1988. Death certificates have also been obtained for 5,275 of the 5,548 known decedents, and these data have been merged with the AIM database.

AIM will continue to study the panel of survivors (n ≅ 2,800) in 1996 and continue linking service utilization data (medical, hospital, nursing home care, home care, and pharmacare data) of survivors and death certificates of decedents. Because 1996 is a census year and also the year of several relevant national surveys (GSS 1996, National Population Health Survey 1996, and the Canadian Study of Health and Aging 1995/96), comparisons of selected characteristics of older Manitobans with their prairie region and national counterparts will be uniquely possible. This will provide a broader background and increase their usefulness to program developers and managers, policy makers, and researchers across Canada. By 1996, the shortest panel will comprise three waves over 13 years, while the longest panel will include four waves over 25 years. The utilization data for these individuals cover a total of 30 years.

The AIM database is unique in two ways: First, it is one of the largest and most extensive population-based longitudinal studies of aging in existence; and second, to date, it is still the only longitudinal study of aging that combines complete utilization data with the interview data. The most similar study with interview and utilization data is the U.S. Longitudinal Study of Aging, with four interviews between 1984 and 1990 (N ≅ 7,000). Data from the Medicare files (for all those respondents who could be matched) are linked in the database (Kovar, 1993). AIM is the only Canadian longitudinal study of aging that is population based and by 1996 will be the longest Canadian study. An annotated bibliography of the over 200 papers, articles, chapters, and reports based on the AIM database has recently been completed (Hall, 1994).

Background

Much of the early analyses of the AIM cross-sectional data were envisioned as testing and adapting the Andersen model of utilization (Andersen & Newman, 1973) based on predisposing, enabling, and need variables. None of these analyses provided complete support for the model. On the other hand, many of these analyses supported a consistency or continuity model; that is, previous use based on need as measured by objective health status scores was the most or among the most important determinants of subsequent service use (e.g., Mossey, Mutran, Shapiro, & Andrews, 1984). Another early line of inquiry found support for the link between self-rated health and mortality (Mossey & Shapiro, 1982). Another important variable throughout the past two decades has been functional capacity, most frequently measured by activities of daily living (ADLs; Katz, Ford, et al., 1963) and instrumental activities of daily living (IADLs; Lawton & Brody, 1969). Some of the most recent AIM publications continue this work (e.g., N. P. Roos, Havens, & Black, 1993). A renewed interest in ADL and IADL has emerged with our own and other longitudinal studies.

None of these analyses were strikingly able to explain variations in mortality, morbidity, disability, health status or health service utilization. Even in the most complete models using all of these variables, less than 50% of the

Betty Havens is a Professor in the Department of Community Health Sciences, University of Manitoba, Winnipeg, and a Visiting Research Fellow with Statistics Canada.

The author would like to thank her AIM co–principal investigators and associates for their intellectual contributions, both the National Health Research Development Program, Health Canada, and Manitoba Health for funding support and the Department of Community Health Sciences for administrative support over the past 25 years.

variance was explained. In another avenue of investigation, researchers began looking at the compression of mortality (Fries, 1980) versus the expansion of morbidity (Manton, 1982). While several reports include some of this reasoning, one of the most specifically relevant is Black, N. P. Roos, Havens, and MacWilliam (1995). This strategy appears to have greater explanatory capacity but needs to be pursued further to include cognitive impairments and the "quiet" disabilities, like arthritis. The rectangularization of the survival curve and whether it yields decreased disability and morbidity in the final years of life or accompanies an increase in the same and the related concept of active life expectancy (Branch, Guralnik, Foley, et al., 1991) is probably the most energetically debated issue in health and aging today. Ensuring sound longitudinal research on this issue is critical to appropriately informing the policies that determine resource allocations in the health care system and especially those resources consumed at the end of life.

Closely related research measures health status transitions and the trajectories of diseases in old age as they are usually related to active life expectancy. While active life expectancy is a prevalent concept in North American studies, the transitions and trajectories are more common in European research (e.g., Euridiss, 1990). There is sufficient overlap in content between the AIM database and many of these studies to enable testing alternative models.

A relatively recent concept to be empirically explored is that of successful aging. Many early gerontological studies have focused on adjustment to retirement or to old age, but the measures have invariably been retrospective recall items. The newer approach seeks to identify the determinants of successful aging or of good health in old age and explicitly recognizes aging as a process (Rakowski, Mor, & Hiris, 1991). N. P. Roos and Havens (1991) have demonstrated that the AIM database includes over 100 indicators with the potential to predict successful aging.

The utilization of health care services is another research avenue. As early as 1981, N. P. Roos and Shapiro published preliminary findings on utilization from the AIM data. This was followed by several other studies that compared utilization for those in the community and in facilities and for survivors and decedents. Of related interest are those studies based on AIM that have looked at specific diseases, such as diabetes, dementia, and arthritis.

One of the exciting potentials of this rich database is contribution to the evolving knowledge about the oldest old (e.g., Suzman, Willis, & Manton, 1992). The AIM database contains in excess of 400 oldest old respondents in each panel wave, with centenarians being the most extreme cases. They may be viewed as the most robust survivors. Preliminary analyses of this relatively rare subpopulation are planned using the 1996 data. It is interesting that those aged 100 or older in 1996 were 75 when these studies began and the youngest 1996 panel members will be aged 73.

Living arrangements is another AIM topic. In the original 1971 cross section (Manitoba, 1973), household composition was a major variable. Household composition figured prominently in analyses of proximity or accessibility of resources, as contrasted with the availability of resources. Given the emerging interest in the relationship between proximity in living arrangements and caregiving, these analyses demonstrate newly emergent relevance. The AIM database also provides opportunities to examine both the existence of and changes in proximity of supportive living arrangements among elderly Manitobans over the course of 25 years.

Of related interest is the existence of and changes to caregiving and care receiving over time. Caregiving and -receiving is conceptualized within the broader context of social support and interdependence. Stone (1991) has called attention to the conceptual fuzziness in defining "family" and the resulting problems in policy discussions about caregiving. The AIM data provide specificity to the definition of informal support within and external to the family. This enables considerations of family and nonfamily support and self-care in the matrix of self-care, informal care, and formal care, that is, service utilization (Penning, 1995).

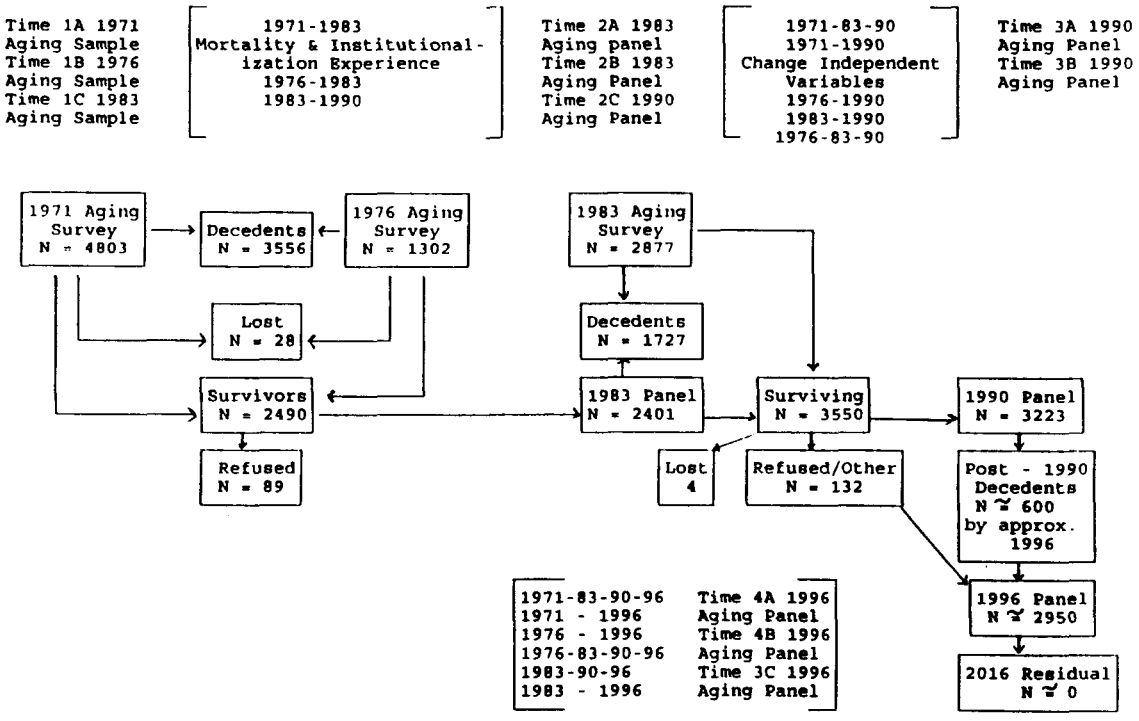
The database has been used to analyze service utilization changes in the years preceding death (N. P. Roos, Montgomery, & L. L. Roos, 1987) and preceding placement in nursing homes (Shapiro & Tate, 1988). AIM is uniquely flexible and capable of analysis using any pivot point over almost three decades of service data. Typically, the analyses have used the interview dates as pivots; however, the articles cited in this paragraph demonstrate the use of alternative pivots.

"Sample mortality" has been defined to refer to those who were in the sample at Time 1 (T_1) and were neither a part of the panel nor among the decedents to $T_2 \ldots T_n$ (Ekland, 1968). Sample mortality therefore arises from refusals, being too ill to participate and having no available proxy, relocation without a new address (including some moves into institutions), and being lost to follow-up through administrative errors but not from death or case mortality. When case mortality is treated as a dependent variable or outcome measure (as is typical in longitudinal research on aging), a second form of sample mortality occurs; that is, no death certificates are secured, although deaths are confirmed from other sources. The 1983–84 panel survey sustained a 1.9% sample mortality rate after three years of very intensive follow-up. The 1990 panel suffered a 3.8% sample mortality rate with only passive follow-up. Similarly, the sample mortality rate among decedents was 1.0% in 1983–84 and 3.1% in 1990 (Havens, 1994).

Data Linking Design

The overall integrating AIM longitudinal study design is displayed in Figure 1. The successful retention of panel respondents, matching and merging utilization data and securing death certificates over the past 25 years is documented elsewhere (Havens, 1994).

Figure 1. AIM longitudinal study design



Interviewers are hired, trained, and then deployed throughout the province in teams of five members. Over almost 25 years of interviewing older Manitobans on a wide range of topics—including several that have been deemed by researchers, interviewers, and older people themselves to be sensitive—refusal rates have never exceeded 3.0%. The most important means to secure adequate responses is to ensure that the interviewers are thoroughly trained, have overcome any of their own sensitivities, are secure with all the items, are meticulous in coding all responses immediately following each interview, are able to make multiple attempts to reach the interviewee, and are thoroughly familiar with the geographic areas.

The management and maintenance of the AIM database, including both the interview data and the utilization data, are not trivial concerns. Each respondent may have up to 1,450 interview item responses, over 300,000 additional health service encounters, and a further multitude of derived variables. Merging the health services utilization data is complex. Despite continuing literature that disputes the reliable and unbiased use of administrative data, the AIM databases have been well documented as containing high-quality data (Mossey, Havens, N. P. Roos, & Shapiro, 1981; L. L. Roos, N. P. Roos, Cageorge, & Nicol, 1982; L. L. Roos, Nicol, & Cageorge, 1987; N. P. Roos, L. L. Roos, Mossey, & Havens, 1988).

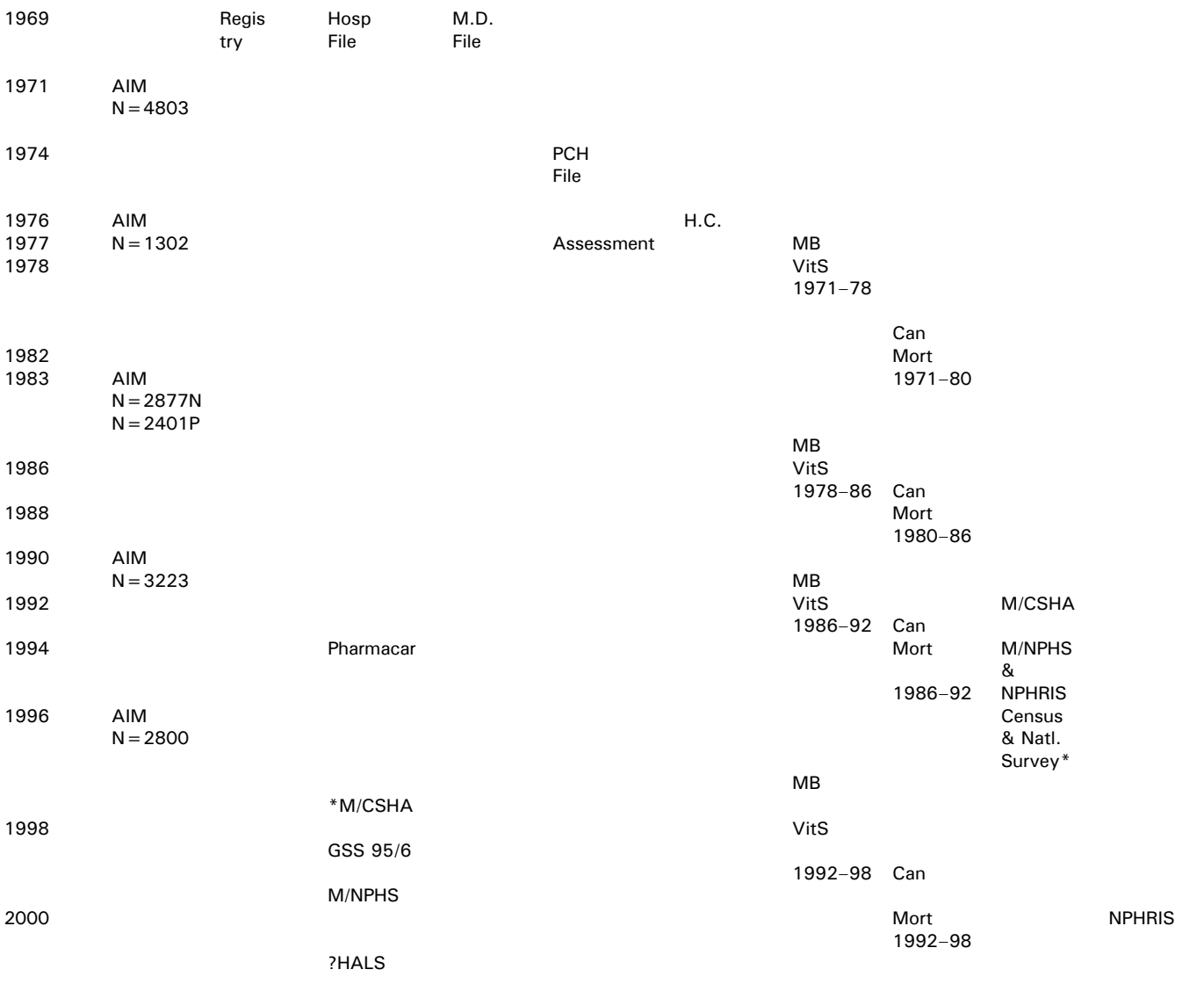
The utilization portions of the AIM database (see Figure 2) are subsets (based on the 9,000 AIM respondents) of the Manitoba Health health services databases for insured services and of the Manitoba Health administrative data for services delivered directly by or through the provincial home care program. The health services databases contain electronic records of all insured services used by all Manitobans, including any services consumed outside the province. The home care data are manual records of those Manitobans registered with the home care program for in-home services or for assessments leading to placement in nursing homes and chronic care units.

Because the home care data are manual records housed in every local office of Manitoba Health, only data from the central registry have been merged into the AIM database. The central registry data include the date of first assessment for care, dates of entry (and re-entry) into the program, dates of discharge(s) from the program, disposition at closure or discharge, and the coordinating office location.

The insured health services data sets include all physician visits, all hospitalizations except emergency and outpatient contacts, nursing home assessments and reassessments, and nursing home admissions and discharges. From the standpoint of encounters with the health system, it is possible for any one person to have up to 31 encounters on any given day over the almost 30 years of utilization data in the AIM database. Therefore, each respondent could have up to 342,300 encounters with the system by 1998 (including pharmacare since mid-1994) or as few as 2 encounters (registration and death). This variability in record length has lead to analytic complexities and to the development of utilization summary variables.

AIM has previously tried unsuccessfully to link pharmacare data to respondents; however, until mid-1994, these

Figure 2. AIM timeline 1969–2000



data were family files, not individual-specific files. Since mid-1994, data have been captured per individual that will enable linking them to AIM respondents and consequently will enable their inclusion in the AIM database. This addition will support new avenues of research and analysis.

The accumulation and merging of death data from AIM decedents have been ongoing since 1980. Deaths of respondents may be identified from several sources: the health services registration file, the hospital database, the nursing home database, the home care administrative records, the medical database, and interviewers who identify decedents not previously recorded by the health system records. The provincial databases document deaths by date of death and location of death. They usually document place of residence at death and generally document primary cause of death.

The actual death certificates contain considerably more information than is captured by Vital Statistics, which is more than is recorded in the Manitoba Health databases. The missing information of most interest to AIM is cause of death, especially multiple or underlying causes of death. Many chronic diseases such as arthritis, diabetes, and even dementia, may contribute to death even if they are seldom fatal. As a result, the primary cause of death is of limited value in gerontological research, which, like AIM, includes longitudinal health services utilization data. Therefore, we have acquired copies of the complete death certificates.

Linking utilization and death records with the interview data is primarily based on the respondent's health services registration number (provided to AIM in an algorithmic manner by Manitoba Health). However, an off-site, confidential alphabetic file is also maintained to link data from the home care central registry and the death data that are only available by name of respondent. The additional confirming variables available within the health services registration files for purposes of data linkage are the first three letters of the surname, the given name (or its first three letters or first or first and middle initials, depending on the data set of origin), date of birth (year, month, and day), gender, location (six-digit postal code or three-digit municipal code), residence (applies only to those in hospitals, nursing homes, or other residential facilities), and date of last contact. The last contact enables one to trace previous addresses and may identify a potential contact person.

Of related interest is the Manitoba/Canada data linkage pilot project (in part based on experience with the AIM database), which is testing whether Canadian census records can be linked with Manitoba Health administrative data. This linkage must be pursued without benefit of a unique identifier held in common by the two databases, one federal and the other provincial (David, Berthelot, & Mustard, 1993). An even smaller feasibility assessment has been initiated by the AIM principal investigator and colleagues at Statistics Canada to determine whether other "synthetic" linkages, specifically between the AIM database and several national survey databases, may be possible (Stone, Hagey, Norris, & Havens, 1994; Havens, 1989a).

Conclusions

AIM is particularly well suited to analyses related to health and social policy issues as they relate to seniors in general and specifically to those aspects of policy that can best be served by analyses that draw on longitudinal interview data and health services utilization data. Examples of analyses of this type include hospital use versus nursing home use versus home care use, each with or without informal network supports. Utilization data are most complete from health services databases (as opposed to respondent recall items), but data on support networks require interview responses. The AIM studies have provided many examples of this unique blend of data used for policy relevant research over the past 15 years (Havens, 1989b; Shapiro, 1991, pp. 38–66; Shapiro & Tate, 1988; N. P. Roos, Shapiro, & L. L. Roos, 1984; Roos & Havens, 1991).

For over two decades, these data have been used to shape policies and programs affecting senior Manitobans and to foster related research and evaluation of public policies and a broad range of programs that target seniors. As such, AIM is both research in the current interests of the health of Canadians, especially senior Canadians, and a research interest in its own right. The 1996 wave will continue the tradition of policy relevant research and will be particularly important from the standpoint of those policy issues that relate to the oldest old, the subpopulation that proportionately uses the most health services. Further, the longitudinal survey data coupled with the utilization data will enable us to investigate the policy implications of changes in available services and in family structures and support systems among our respondents.

Finally, knowing that "chronic diseases are the leading causes of premature death and disability" (Health Canada, 1994, p. 1) and that both the resultant service costs and the potential savings from postponing disability are major concerns in this era of shrinking resources, the AIM longitudinal study will contribute to informing policies relative to ameliorating chronic diseases. This research will continue to enable program developers and policy makers to tailor programs in the most appropriate manner to support the continuing independence of older Canadians as we approach the 21st century.

References

Andersen, R. M., & Newman, J. (1973). Societal and individual determinants of medical care utilization in the United States. Milbank Memorial Fund Quarterly, 51, 95–124.

Black, C., Roos, N. P., Havens, B., & MacWilliam, L. (1995). Rising use of physician services by the elderly: The contribution of morbidity. Canadian Journal on Aging, 14, 225–244.

Branch, L. G., Guralnik, J. M., Foley, D. J., et al. (1991). Active life expectancy for 10,000 Caucasian men and women in three communities. Journal of Gerontology, 46,(4) M145–M150.

David, P., Berthelot, J.-M., & Mustard, C. (1993). Linking survey and administrative data to study determinants of health. American Statistical Association 1993 Proceedings of the Social Statistics Section, 155–160.

Ekland, B. K. (1968). Retrieving mobile cases in longitudinal surveys. Public Opinion Quarterly, 32, 51–64.

Euridiss. (1990). European research on incapacitating diseases and social support. International Journal of Health Sciences, 1, 217–228.

Fries, J. F. (1980). Aging, natural death and the compression of morbidity. The New England Journal of Medicine, 3, 130–135.

Hall, M. K. (1994). Annotated bibliography of the Aging in Manitoba studies. Winnipeg, Canada: Aging in Manitoba Project.

Havens, B. (1989a, July). Can the Manitoba data base be a test case or a case study? Presented at the Public Health Conference on Records and Statistics, Washington, DC.

Havens, B. (1989b). Linking theory, research, practice and policy. In S. Lewis (Ed.), Aging and health: Linking research and public policy. Chelsea, MI: Lewis.

Havens, B. (1994, July). Sample mortality: Aging in Manitoba 1971–1994. Presented to the Department of Community Health Sciences and Centre for Health Policy and Evaluation, University of Manitoba, Winnipeg, Canada.

Health Canada. (1994). Bureau of Chronic Diseases, program description: Fiscal year 1994/95. Ottawa: Laboratory Centre for Disease Control.

Katz, S., Ford, A. B., et al. (1963). Studies of illness in the aged. The index of ASL: A standardized measure of biological and psychosocial function. Journal of the American Medical Association, 185, 914–919.

Kovar, M. G. (1993). A longitudinal study based on a continuing national survey: The Longitudinal Study of Aging. In J. Armstrong, N. Darcovich, & P. Lavallee (Eds.), Symposium 92: Design and analysis of longitudinal survey proceedings (pp. 179–185). Ottawa, Canada: Minister of Industry Science and Technology.

Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. The Gerontologist, 9, 179–186.

Manitoba, Division of Research, Planning and Program Development. (1973). Aging in Manitoba: Needs and resources, 1971 (Vols. 1–10). Winnipeg, Canada: Manitoba Department of Health and Social Development.

Manton, K. (1982). Changing concepts of morbidity and mortality in the elderly population. Milbank Memorial Fund Quarterly, 2, 183–245.

Mossey, J. M., Havens, B., Roos, N. P., & Shapiro, E. (1981). The Manitoba Longitudinal Study on Aging: Description and methods. The Gerontologist, 21, 551–558.

Mossey, J. M., Mutran, E., Shapiro, E., & Andrews, K. (1984). Determinants of consistency and level of physician use by the elderly over an eight year period. Presented at the Gerontological Society of America Meetings, San Antonio, TX.

Mossey, J. M., & Shapiro, E. (1982). Self-rated health: A predictor of mortality among the elderly. American Journal of Public Health, 72, 800–808.

Penning, M. J. (1995). Health, social support, and the utilization of health services among older adults. Journal of Gerontology Series B, 50B(5).

Rakowski, W., Mor, V., & Hiris, J. (1991). The association of self-rated health with two-year mortality in a sample of well elderly. Journal of Aging and Health, 3, 527–545.

Roos, L. L. Jr., Nicol, J. P., & Cageorge, S. M. (1987). Using administrative data for longitudinal research: Comparisons with primary data collection. Journal of Chronic Diseases, 40, 41–49.

Roos, L. L. Jr., Roos, N. P., Cageorge, S. M., & Nicol, J. P. (1982). How good are the data? Reliability of one health care data bank. Medical Care, 20, 266–276.

Roos, N. P., & Havens, B. (1991). Predictors of successful aging: A twelve year study of Manitoba elderly. American Journal of Public Health, 81, 63–68.

Roos, N. P., Havens, B., & Black, C. (1993). Living longer but doing worse: Assessing health status in elderly persons at two points in time in Manitoba, Canada, 1971 and 1983. Social Science and Medicine, 36, 273–282.

Roos, N. P., Montgomery, P. R., & Roos, L. L. Jr. (1987). Health care utilization in the years prior to death. Milbank Memorial Fund Quarterly, 65, 231–254.

Roos, N. P., Roos, L. L. Jr., Mossey, J. M., & Havens, B. (1988). Using administrative data to predict important health outcomes: Entry to hospital, nursing home and death. Medical Care, 26, 221–239.

Roos, N. P., & Shapiro, E. (1981). The Manitoba Longitudinal Study on Aging: Preliminary findings on health care utilization by the elderly. Medical Care, 19, 644–657.

Roos, N. P., Shapiro, E., & Roos, L. L. Jr. (1984). Aging and the demand for health care services: Which aged and whose demand? Gerontologist, 24, 31–36.

Shapiro, E. (1991). Manitoba health care studies and their policy implications. Winnipeg, Manitoba: Manitoba Centre for Health Policy and Evaluation.

Shapiro, E., & Tate, R. B. (1988). Who is really at risk of institutionalization? Gerontologist, 28, 237–245.

Stone, L., Hagey, J., Norris, D., & Havens, B. (1994, September 8). Proposal for visiting research fellow. Personal Communication.

Stone, R. (1991). Defining family caregivers of the elderly: Implications for research and public policy. The Gerontologist, 31, 724–725.

Suzman, R. M., Willis, D. P., & Manton, K. G. (Eds.). (1992). The oldest old. Oxford: Oxford Press.

Linking Primary and Secondary Data for Outcomes Research: Methodology of the Total Knee Replacement Patient Outcomes Research Team

John E. Paul, Catherine A. Melfi, Timothy K. Smith, Deborah A. Freund, Barry P. Katz, Peter C. Coyte, and Gillian A. Hawker

Introduction

This paper describes the data collection strategy used by the total knee replacement (TKR) Patient Outcomes Research Team (PORT), which is one of about 15 PORTs funded by the Agency for Health Care Policy and Research (AHCPR) to study patient-based outcomes following a particular medical condition and/or procedure (PORTs, 1990; Salive, Mayfield, & Weissman, 1990). In what follows, we discuss the use of Medicare claims data for sampling purposes, issues of confidentiality of patient level information, and the patient survey and medical records abstraction components of the study. We also describe the building of analysis files through linking the patient survey and medical records data with the administrative data.

Linked multiple data sources are necessary in order to test the many research questions related to patient-based outcomes—clinical, functional, patient satisfaction, and charges for and utilization of the procedure. Such data sets also offer unique opportunities for assessment of data reliability and validity across multiple sources of information. This paper describes the data-linking methodology used by the TKR PORT, along with comments regarding the usefulness of the methodology for other outcome studies.

Sample Design

Administrative or claims databases provide an important source from which to construct a sampling frame for subsequent data collection, provided the population represented in the large database is fully understood (Paul, Weis, & Epstein, 1993). Since there is no national registry for either severe knee arthritis or TKR, we used Medicare Provider Analysis and Review (MEDPAR) records from the Health Care Financing Administration (HCFA) as our sampling frame. The universe of eligible Medicare beneficiaries aged 65 and older is generally representative of the U.S. population over 65, when compared with Census data (Fisher, Baron, Malenka, Barrett, & Bubolz, 1990). However, using Medicare files as a sampling frame causes one to miss those whose care was paid for outside the traditional fee-for-service Medicare system. This problem is most pronounced for Medicare enrollees in HMOs, for whom utilization of services is either underreported or largely missing in the Medicare files. The problem also applies to persons served by Veteran's Affairs (VA) hospitals. Over the period 1985 through 1989, enrollment of Medicare beneficiaries in HMOs nationwide was approximately 3% to 5%, with substantial regional variation (U.S. Department of Health and Human Services [DHHS], 1990; Group Health Association of America, 1989).

Since TKR is exclusively an inpatient procedure, the MEDPAR files, which contain data on 100% of the Medicare-reimbursed hospitalizations, were appropriate for constructing annual cohorts of patients receiving at least one TKR during the study period. The cohorts were constructed based on procedure and diagnosis codes recorded in the claims data. It is crucial to select and specify an appropriate algorithm for identifying the correct records. An incomplete or incorrectly specified algorithm could lead to a data set that is not appropriate for the intended analysis. For patients who received more than one TKR during the study period, their first TKR in the period was used to designate the cohort or index year.

John E. Paul is Director of Clinical Economics in the Care Management Division, Glaxo Wellcome, Inc., Research Triangle Park, North Carolina. Catherine A. Melfi is a Research Scientist at Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, and an Adjunct Scientist at the Indiana University School of Medicine, Indianapolis. Timothy K. Smith is a Survey Methodologist in the Survey Methods Research Program of the Survey Research Division at Research Triangle Institute, Research Triangle Park. Deborah A. Freund is a Professor of Public Affairs in the School of Public and Environmental Affairs and Director of the Bowen Research Center at Indiana University. Barry P. Katz is an Associate Professor in the Department of Medicine at Indiana University. Peter C. Coyte is an Associate Professor in the Department of Health Administration, Institute for Policy Analysis and the Centre for the Study of State and Market, University of Toronto, and Adjunct Senior Scientist in the Institute for Clinical Evaluative Sciences, North York, Ontario. Gillian A. Hawker is an Assistant Professor in the Departments of Medicine and Health Administration at the University of Toronto and Staff Rheumatologist at Women's College Hospital in Toronto.

This research was funded by the U.S. Agency for Health Care Policy and Research Grant No. 06432 to Indiana University and subgrantees to establish a Patient Outcomes Research Team to study total knee replacement surgery. Opinions expressed in this paper do not necessarily reflect the opinions of any funding agency.

Address correspondence to Catherine Melfi at the address listed on page 237.

Five types of exclusions were applied sequentially to the TKR-selected MEDPAR files in order to better define the population under study: (a) non–U.S. residents and all those under 65 who were Medicare eligible due to disability or end stage renal disease; (b) beneficiaries who received their TKRs while enrolled in an HMO, since complete claims data would not be available for these beneficiaries; (c) hospitalizations with records containing empirical criteria indicating that a TKR was likely not performed, for example, hospitalizations with length of stay (LOS) less than three days and zero total charges and discharge to home with self-care; (d) hospitalizations in which a single TKR procedure code was contradicted by a diagnosis code (V64.x) indicating "procedure not performed due to contra-indication"; and (e) transfer hospitalizations indicating a TKR in sites where TKR surgery would not be expected, such as psychiatric or rehabilitation facilities. Total exclusions amounted to 7.4% of the original number of TKR-selected hospitalizations.

The resulting cleaned TKR hospitalization file was converted to a person level sampling frame by further excluding individuals currently indicated as deceased in the Medicare vital status files and taking only the first or index TKR hospitalization during the study period if an individual had more than one TKR hospitalization. The final sampling frame contained 261,823 persons. (By year, the range was from 42,477 in 1985 to 63,086 in 1989.) In addition to serving as a sampling frame, the administrative claims files provided key information for use in outcomes analysis (e.g., charges, LOS, discharge destination, and mortality).

Next, we specified a stratified random sample with six strata based on patient demographic characteristics so that we could conduct subanalyses on selected groups of patients. Three independent samples were drawn: a U.S. (national) sample, an Indiana sample, and a sample from 29 counties in western Pennsylvania served by Blue Cross and Blue Shield of Western Pennsylvania. We drew separate samples for Indiana and western Pennsylvania to allow special focus on these areas using additional data available to the TKR PORT collaborators from these areas. Although the national sample included people in Indiana and western Pennsylvania, there was no overlap of individuals in the three samples.

The sample design accounted for differences in TKR utilization by race, urban/rural residence, and age, based upon preliminary analysis of the Medicare claims data. The six sampling strata were (a) blacks, (b) "unknown" and "other" race, (c) rural whites under 80 years of age, (d) rural whites over 80, (e) urban whites under 80, and (f) urban whites over 80. Equal numbers were sampled across each of the 5 years of TKR discharges. By stratum, equal numbers were sampled for the national sample; however, the strata were sampled with differing proportions for the Indiana and western Pennsylvania samples to reflect different characteristics of the underlying population in these areas and to allow for adequate power in our subanalyses. The final sample sizes were as follows: 750 TKR patients in the national sample; 500 in the Indiana sample; and 500 in the western Pennsylvania sample, for a total of 1,750.

The unique beneficiary health insurance claim numbers and corresponding inpatient medical care provider numbers of the 1,750 patients selected from the MEDPAR files were submitted to HCFA for matching with Social Security eligibility and Medicare provider files. We thus obtained (a) the name and current address of the TKR recipient to whom the lead letter and patient survey were sent and (b) the name and address of the inpatient provider who performed the TKR, to whom we sent the request for copies of medical records for the TKR hospitalization. We were not able to obtain telephone numbers from HCFA, however.

Data Collection

Patient Survey Field Procedures and Outcomes

The National Survey of TKR (the "TKR patient survey") was a mail survey with telephone follow-up fielded in the first half of 1992. PORT researchers developed the patient survey instrument during the first 18 months of PORT activities and pretested it in multiple sites prior to finalization. The instrument made use of existing validated scales, including a generic health status instrument, the Medical Outcomes Study 36-item short form (MOS SF-36) (Ware & Sherbourne, 1992), a disease-specific health measure; the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain and activity limitation scales (Bellamy, Buchanan, Goldsmith, Campbell, & Stitt, 1988; Bellamy, 1989); and selected scales from the Knee Society assessment instrument (Insall, Dorr, Scott, & Scott, 1989). Table 1 provides an outline of the patient survey instrument and key variables.

Prior to the first survey mailing, a letter was sent to each sample member from the HCFA administrator on DHHS HCFA letterhead explaining the purpose of the survey, advising them of the intended request for copies of medical records relating to their TKR hospitalization, and stressing that their participation was entirely voluntary and would in no way affect their Medicare coverage or participation in any other government program. A toll free telephone number for questions was provided.

Protection of confidentiality for sample members and respondents in data collection efforts is of utmost concern to government (AHCPR) and PORT investigators. Moreover, HCFA, as the source of patient and provider names and addresses, has its own stringent set of review and clearance requirements. Initially, we obtained institutional review board (IRB) approval of the data collection protocol from IRBs both at the TKR PORT lead organization and at the PORT data collection subcontractor in accordance with 45 CFR 46 (Protection of Human Subjects). Concerns about the rights of sample members to decline participation and to decline use of their TKR medical records in the research were addressed in the cover letters and questionnaires.

Table 1. Key elements from the TKR patient survey

General/background to TKR
· Laterality of replaced knee
· Previous TKR, osteotomy
· Reason for TKR
· Subsequent knee surgery on same knee
· Perceived health status/social functioning, time of TKR
Knee-related pain (WOMAC and Knee Society measures)
· Prior to TKR; currently
Physical functioning/activity limitations/stiffness (WOMAC and Knee Society measures)
· Prior to TKR; currently
Access to TKR
· Locating surgeon
· Timing of initial surgical consultation
· Timing of surgery
Post-TKR physical therapy
Satisfaction with TKR
· Current status of replaced knee
· Overall satisfaction with TKR
· Right decision made regarding TKR
Medication for joint pain/swelling
· Prior to TKR; currently
Current health
· SF-36 general health status and functioning scales
· Comorbid conditions
· General arthritis-related questions
· Status of other knee and lower body joints
Patient demographics and other information
· Unique patient identifier and linking variable
· Living arrangements prior to TKR and currently
· Employment status prior to TKR and currently
· Current height and weight; weight around time of TKR
· Educational level
· Race and ethnicity
· Insurance coverage at time of TKR and currently
· Income level at time of TKR and currently
· Use of VA health care
· Identification of proxy respondents
· Assistance in completing questionnaire
· Permission to use medical records information in study

AHCPR and HCFA reviewed all protocols and draft data collection materials. HCFA also required that its own lead letter go out to all sample members prior to mailing the survey. The HCFA lead letter was essentially informed consent information signed by the HCFA administrator stressing the voluntary nature of participation in the survey. The informed consent statement was repeated in the cover letter that accompanied each mailed survey instrument.

One of the unusual characteristics of the sample population was its age. Since we began with aged Medicare beneficiaries who had TKRs between 1985 and 1989, the youngest a person in our sample at the time of the survey was 67 years old. The oldest person in the sample was 99 years old. The mean age at time of TKR was 79.6 (SD = 6.53). Researchers have expressed concerns regarding whether this population would participate in a mail survey (Herzog & Rodgers, 1988) and whether they could recall their functional status and the events surrounding their TKR, which may have occurred up to 7 years prior to receiving the survey questionnaire. Other studies (Kovar, 1989) and commentaries (Kalton, 1989), however, have established the basic feasibility of surveys of older persons, specifically by telephone. A 54.7% response rate to the first mailing (adjusted for ineligible, incapacitated, and deceased sample members) was achieved after 5 weeks, giving us confidence in the willingness and ability of an elderly population to respond to a mail survey.

Following standard mail survey techniques, 5 weeks after the initial mailing, we sent a second mailing of the questionnaire to the 713 sample members from whom we had not yet received a response or for whom we did not have a final disposition code. Five weeks following the second mailing, the patient survey had achieved a cumulative response rate of 68.7%. The remaining 437 nonresponding cases were referred for telephone tracing and follow-up. The 4-week telephone follow-up resulted in another 160 completed surveys, for an overall patient survey response rate of 80.3%.

Data quality was addressed by establishing a priori unacceptable data ("fail-edit") criteria, defined as missing or bad data on one or more of the key questions relating to pain or physical functioning following the TKR. Approximately one-third of the respondents who completed the questionnaire by mail failed at least one of the edit criteria. For each case, we attempted to resolve the problem by telephone. In the survey instrument, we requested that respondents provide telephone numbers and a convenient time to call regarding any questions about the returned questionnaire. Since most of the respondents (about 90%) complied with this request, contacting those cases that failed the edit was straightforward. We carried out telephone tracing for the 10% who didn't provide a contact number. All of the respondents whom we contacted for fail-edit resolution were willing and able to provide the key data we needed. We also attempted to correct any problems with other questions while the respondent was on the telephone. Fail-edit callbacks were carried out throughout the data collection period, successfully addressing the potential problem of item nonresponse.

Examination of the response rate by stratum and site indicated the range was from a low of 67.3% (blacks—national sample) to a high of 92.2% (urban whites under 80—Indiana sample). Older patients were less likely to respond than were younger patients, as were patients with more distant TKRs compared with patients with more recent TKRs.

The only major problem that arose during the patient survey was related to losing nearly 13% of the sample due

to the sample members being deceased. Despite efforts to identify and remove deceased individuals from the administrative data sampling frame, a large number remained among the names and addresses sent by HCFA. Unfortunately, these persons were identified only after the first mailing was sent. A total of 15.1% of the sample overall was found to be either deceased, ineligible, or incapacitated.

Medical Records Field Procedures and Outcomes

Concurrent with the mailing of the TKR patient survey, we mailed a request for copies of medical records to each hospital where TKRs were performed on patient sample members. The request consisted of a cover letter on AHCPR stationery to the hospital's medical records administrator explaining the research project and a series of one or more patient specific identification forms, containing patient name; date of birth; dates of admission and discharge; and, when available, the hospital's medical record number for the patient. Key data elements abstracted from the hospital medical records are listed in Table 2.

The number of sampled patients per hospital ranged from 1 to 38. The concentration of sample members per hospital was much greater in the Indiana and western Pennsylvania samples than in the national sample because there were fewer hospitals from which sample members could receive their TKR. Overall, we identified 714 unique hospital providers for the 1,750 total sample members.

Table 2. Key elements from hospital medical records abstraction

Unique patient identifier and linking variable
Preoperative clinical and other characteristics
· Unique patient identifier and linking variable
· Diagnosis and procedure codes
· Insurance status (in addition to Medicare)
· Identification of knee(s) operated on
· Previous joint prostheses (hip and knee)
· Preoperative ambulation status
· Patient weight on admission
· Preoperative range of motion of replaced knee
· Preoperative malalignment of replaced knee
Peri-operative surgical characteristics and postoperative outcomes
· Date of surgery
· Type of anesthesia
· Prosthesis type and manufacturer
· Use of cement and cement antibiotics
· Use of bone graft
· Type of patellar resurfacing
· Surgery time (anesthesia time; tourniquet time)
· Blood transfusion information
· Postoperative complications
· Other technical features of surgery
· Postoperative physical therapy

Copies of medical records were received in both hard copy and microfiche formats. The first mailed request for records resulted in receipt of nearly 42% of the needed patient records and represented cooperation from over 60% of the hospitals. A second mailing was sent out to 278 nonresponding hospitals (representing 737 patients) 6 weeks after the first request. Telephone follow-up 4 weeks after the second mailing was done to answer any further questions and encourage participation in the study by the hospitals. Overall, 84.9% of the hospitals complied fully to our request for copies of medical records by the end of the 12-week data collection period. Another 2.9% complied partially by providing copies of some, but not all, of the requested patient records. In all, we obtained useable copies of inpatient medical records relating to the index TKR hospitalization for 79.5% of our sampled patients.

Approximately 5% of the hospitals inquired about confidentiality of patient records and patient permission beyond the information contained in the AHCPR cover letter that accompanied the request. In the cases in which patient permission was explicitly requested by the hospital, copies of the signed page from the patient survey instrument giving permission to include the records, if available, were photocopied and sent to the institution. Among all respondents to the patient survey, 85.8% answered affirmatively to the request for permission to use copies of their medical records in the study. Eight percent denied permission to include their medical records in the study, with the remainder leaving the question blank. In cases in which patient permission was refused, we made no further attempts to acquire records and destroyed any records that were already received.

Construction and Use of Linked Files

The separate field efforts resulted in differing numbers of patient surveys and medical records. Specifically, there were 1,193 patient surveys and 1,391 medical records. Constructing a data set that consisted of sample members for whom we had MEDPAR claims data, patient survey data, and medical records data resulted in a data set with 962 records.

In addition to the data specifically collected as described above, two other sources of administrative data containing useful covariates or other explanatory variables are being accessed by the PORT. First, the American Hospital Association (AHA) Annual Hospital Survey data were linked with the MEDPAR data using the unique inpatient provider number. The AHA survey data provide important organizational level variables such as ownership characteristics, teaching status, and number of beds. Second, the Area Resource File (ARF) data from the DHHS Bureau of Health Professions, HRSA, were linked to administrative and patient survey data based upon county of residence of the TKR recipient. The ARF provides up-to-date health resource and medical marketplace covariates such as the

number of physicians by specialty and the number of hospital, nursing home, and rehabilitation beds in a county area. The ARF also has county level demographic and socioeconomic data for potential inclusion in analysis models.

Verification of Data Across Sources

One way that linked files such as those described here may be used is for purposes of validating data from one source and/or filling in missing information about some subjects. For example, with the increased use of administrative claims data for analysis, there has been some concern regarding the accuracy of the procedure and diagnosis coding on claims databases (Fisher et al., 1992). This is especially important if this information is to be used to control for comorbidities, preexisting conditions, complications, and/or overall severity of illness. By comparing the medical records with the claims data, we were able to examine how well the claims data reflected what was contained in the more complete medical records.

Both medical records and the claims data have a place where a primary diagnosis is to be included. Any other diagnoses are considered secondary diagnoses. The MEDPAR file allows for up to five diagnosis codes per record, whereas we could abstract up to 14 diagnoses from the medical records. In our linked data set, 1,339 of the 1,391 records (96.3%) agreed completely on primary diagnosis. Since diagnosis is recorded using ICD-9-CM codes, we also examined whether the first three or four digits matched if there was not complete agreement between the medical records and claims data. There was agreement in the first four digits for 1,346 records (96.8%) and agreement in the first three digits for 1,355 of the records (97.4%). Most of the secondary diagnoses matched completely (82.7%), with disagreement mainly due to the room for extra codes in the medical records.

There may be up to three procedure codes in the MEDPAR data, while we could extract up to 14 procedure codes from the medical records. As with diagnoses, we compared agreement on primary procedure as well as secondary procedures. For 1,357 of the 1,391 records (97.6%), the primary procedure was the same in the claims data and on the medical records. Again, discrepancies were mainly due to extra procedures listed on the medical records.

Another area of concern in conducting analysis using only claims data is the reliability of data on patients' racial classification. The MEDPAR data have categories for "black," "white," "other," and "unknown." If it is possible to determine the approximate distribution of races in the "unknown" category, then results can be interpreted more accurately than if that race category is just referred to as "unknown." Based on our patient survey data, 94% (132/141) of patients who were classified as "unknown" race on the claims data were "white"; 4% were "black," and 2% were "other." This is very close to the overall distribution of races in the claims data for "white," "black," and "other." This implies that there is no systematic bias resulting from overrepresentation of a particular race in the "unknown" category.

There were very few variables that could come from both the patient survey and medical records, so we did not make comparisons between those two sources of data. There may be some applications, however, for which this type of comparison would be useful. For example, self-reported information about drug use, number and timing of medical visits, or comorbidities often may be checked against information on medical records if there is a way to link patient survey data with data from medical records.

Summary and Conclusions

The purpose of this paper has been to illustrate the feasibility and usefulness of linking primary and secondary data for outcomes research. Administrative claims files of Medicare-reimbursed inpatient hospitalizations linked to Social Security enrollment files proved to be a tractable sampling frame broadly representative of the U.S. population over 65. The administrative claims files also provided charges and utilization outcomes related to the TKR hospitalizations. This elderly population was successfully surveyed by mail, using a lead letter from HCFA and a two-mailing-with-telephone-follow-up approach. Recall of a significant inpatient event (TKR) 2 to 7 years distant from the hospitalization did not appear to be a problem, based upon assessment of item nonresponse in the survey. It is worth noting that a parallel TKR patient survey and medical records abstraction was also carried out in the province of Ontario, Canada, by TKR PORT collaborators at the University of Toronto. They have also linked their claims, patient survey, and medical records data for analysis, and cross-national analyses are being conducted with the two data sets.

Patient permission for the inclusion of medical records in the study proved straightforward, with more than 85% of patients answering the permission question affirmatively in the patient survey. Hospital willingness to provide photocopies of records was similarly positive, with nearly 88% of institutions fully or partially complying with our request for records. Reflective of the fact that confidentiality and release of personal records are becoming increasingly sensitive issues, most hospitals were appropriately cautious, but they were cooperative with our request. Only a few hospitals were stridently opposed and flatly refused to participate. It is unlikely, however, that gaining cooperation from institutional providers will be any easier in subsequent years. A linked data collection methodology such as that described here should be considered by other projects with similar research goals.

References

Bellamy, N. (1989). Pain assessment in osteoarthritis: Experience with the WOMAC osteoarthritis index. Seminars in Arthritis and Rheumatism, 18, 14.

Bellamy, N., Buchanan, W. W., Goldsmith, C. H., Campbell, J., & Stitt, L. W. (1988). Validation study of WOMAC: A health status instrument measuring clinically important patient-relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. Journal of Rheumatology, 15, 12.

Fisher, E. S., Baron, J. A., Malenka, D. J., Barrett, J., & Bubolz, T. A. (1990). Overcoming potential pitfalls in the use of Medicare data for epidemiological research. American Journal of Public Health, 80, 1487–1490.

Fisher, E. S., Whaley, F. S., Krushat, W. M., Malenka, D. J., Fleming, C., Baron, J. A., & Hsia, D. C. (1992). The accuracy of Medicare's hospital claims data: Progress, but problems remain. American Journal of Public Health, 82, 243–248.

Group Health Association of America. (1989). HMO industry profile: Vol. 1. Benefits, premiums, and market structure in 1988. Washington, DC: GHAA.

Herzog, A. R., & Rodgers, W. L. (1988). Age and response rates to interview sample surveys. Journal of Gerontology, 43, S200–S205.

Insall, J. N., Dorr, L. D., Scott, R. D., & Scott, W. N. (1989). Rationale of the Knee Society clinical rating system. Clinical Orthopaedics, 248, 13–14.

Kalton, G. (1989). Surveying older adults. In F. J. Fowler Jr. (Ed.), Conference proceedings: Health Survey Research Methods (DHHS Publication No. [PHS] 89-3447, pp. 147–151). Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment.

Kovar, M. G. (1989). Collecting health data from and about older people: The Longitudinal Study of Aging. In F. J. Fowler Jr. (Ed.), Conference proceedings: Health Survey Research Methods (DHHS Publication No. [PHS] 89-3447, pp. 115–119). Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment.

Patient Outcomes Research Teams: A new strategy for health services research on medical care quality and effectiveness [Special supplement]. (1990). Health Services Research, 25.

Paul, J. E., Weis, K. A., & Epstein, R. A. (1993). Data bases for variations research. Medical Care, 31(Suppl. 5), YS96–YS102.

Salive, M. E., Mayfield, J. A., & Weissman, N. W. (1990). Patient Outcomes Research Teams and the Agency for Health Care Policy and Research. Health Services Research, 25(suppl.) 697–708.

U.S. Department of Health and Human Services. (1990). Medicare and Medicaid data book, 1990. Baltimore, MD: Health Care Financing Administration, Office of Research and Demonstrations.

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). Medical Care, 30, 473–483.

Collecting Survey and Medical Records Data to Measure Intervention Outcomes in Medical Practices Serving Urban Minorities

Ronald Czaja, Clara Manfredi, and Richard B. Warnecke

Introduction

Organizations are frequently the units of analysis in interventions aimed at altering medical practice. Even in relatively large intervention studies that involve organizations, the number of units rarely exceeds 50, and frequently, the number is less. In such studies, it is critical to obtain complete information on all the participating organizations. Consequently, response rates that would be considered quite acceptable for regular sample surveys can seriously undermine the ability to evaluate outcomes in these studies.

In this paper, we discuss the successes and pitfalls of implementing and evaluating an intervention aimed at improving cancer screening practices in 51 clinics located in Chicago's inner city that serve low-income African American and Hispanic patients. These practices have diverse administrative characteristics, but most experience the usual problems associated with high employee turnover, patient enrollment overload, missed and unscheduled appointments, patient-provider language barriers, low computer skills and access, and manual record keeping and maintenance. These characteristics make access to patient records and obtaining information from providers and clinic staff about clinic operations and policies problematic. As a result, we were unable to obtain complete data from at least one-third of the clinics on key variables related to the evaluation of the intervention. Similarly, and not surprisingly, we were unable to obtain complete data from the providers. Thus, to enable us to effectively evaluate the outcomes of the intervention, we had to devise ways to impute the necessary information by pooling data from a practice survey and a provider survey and from information we obtained from the Census, medical directories, and HMO records. To illustrate the problems and results, we have chosen one key independent variable, whether the practice has ". . . an established, written policy for preventive care consisting of specific services at appropriate intervals."

Ronald Czaja is an Associate Professor in the Department of Sociology and Anthropology at North Carolina State University at Raleigh. Clara Manfredi is an Associate Director of the Prevention Research Center, School of Public Health, University of Illinois at Chicago. Richard B. Warnecke is Director of the Survey Research Laboratory, Professor of Public Administration in the College of Urban Planning and Public Affairs, and Professor of Epidemiology and Biometry in the School of Public Health, University of Illinois at Chicago.

The Intervention and Evaluation Design

This was a randomized trial of a clinic-based intervention designed to increase early cancer detection screening in 51 medical practices believed to serve predominantly minority or low-income patients that accept Medicaid HMO patients. The primary objective was to increase screening for breast, cervical, oral, and colorectal cancer in HMO patients receiving their care in the intervention practices. A secondary objective was to increase the rates of screening for non-HMO patients. A key aspect of the intervention was to assess the effectiveness of the HMO in acting as an intermediary requiring that practices that treat HMO patients provide these screening services. The rationale for examining non-HMO patients in the same practices was to assess whether required changes in practice for HMO patients would generalize to the care of non-HMO patients. The basic hypothesis was that following the intervention, screening rates for both HMO and non-HMO patients in the intervention sites would increase more than the rates in the control sites.

Prior to the intervention, a sample of 60 records was selected from each site and abstracted. Based on these data, preintervention screening rates were computed for each site, and the practices were matched and then randomized to experimental or control conditions. The intervention took place over 17 months. After the intervention, a second sample of 60 records from each intervention and control site was selected and abstracted. Rates were calculated for each site, and then the change in the rates was computed. Six elements comprised the intervention: medical records reminder flagging, physician training carrying one continuing education credit, periodic records audits with feedback to physicians, training of office support staff, distribution of patient health promotion cards to remind patients when their next screenings were due, and flow sheets placed in the records recording which tests were performed and when the next screening was due.

The Quality of Data Sources

Organization and Environmental Characteristics

Census data were used to characterize the neighborhood or environmental characteristics of the practices selected for the study (U.S. Bureau of the Census, 1993). Practices selected for study were located in Census tracts that were on average 56% black. Census tracts that were less than 50% black had an average Hispanic population of 45%. Across all tracts in which the practices were located, the mean of female-headed households was 37%. The mean of the median annual household income was $26,239, the mean percentage unemployed was 25%, and the mean percentage who had changed residence in the previous 5 years was 45%.

At the beginning of the intervention, 166 practices accepted members of the HMO. In consultation with the medical director of the HMO, 53 practices thought to be all the practices serving primarily minority and low-income patients were selected. There was wide variability in staffing arrangements among the selected practices. Only 19 practices had a full-time registered nurse (RN) or licensed practical nurse (LPN), and 1 other practice had a part-time RN. The majority of sites had at least one full-time medical assistant (n = 37), and nine sites had neither an RN, LPN, nor medical assistant. Thirty-four practices had an office manager. The number of primary care physicians at a site varied from 1 to 27, with the median being 3.

There was wide variation in the amount of space at the practices. A few practices were located in hospital settings or professional office buildings and had more than adequate space for patients and staff. Most, however, were in small and crowded quarters in storefront locations or in shared space with other establishments. Space inadequacies created problems for records abstraction. At locations where there was little or no place to work, the abstractors worked in waiting rooms or sat at the reception desk examining patient records.

The intervention targeted primary care physicians in family practice, general practice, internal medicine, and obstetrics and gynecology. Nationally, the percentage of primary care physicians who are in family or general practice is 38%; 45% are in internal medicine, and 17% are gynecologists or obstetricians (American Medical Association [AMA], 1994). At the start of the intervention, the HMO directory listed 176 primary care physicians, of whom 37% were in family or general practice, 41% were internists, and 22% were gynecologists or obstetricians. At the end of the intervention, the respective percentages were family and general practice 41%, internal medicine 45%, and obstetrics and gynecology 14%. Thus, the overall distribution of primary care physicians in the study practices reflected the national distribution.

Physician and Practice Surveys

Two physician surveys were conducted: one at baseline and the other at the end of the intervention. Both were mail surveys using essentially the same procedures: An initial mailing containing a questionnaire and cover letter was sent to all physicians; there followed a second mailing with a new cover letter and a copy of the questionnaire to all who did not respond to the initial mailing; then a reminder postcard was sent, followed by phone call reminders to each practice asking our contact person to remind the physician to return the questionnaire; and finally, a third mailing with a questionnaire and a further revised cover letter from the medical director of the HMO went to all who had not responded to the previous contacts. Also in the first survey, the principal investigator called every physician nonrespondent after the second mailing. Data collection for the first survey took 4 months; the second survey took 4½ months.

In 1992, 176 physicians were listed in the HMO staff directory. Eighteen were ineligible (15 had left the HMO, and 3 were duplicate listings). Of the remaining 158 eligible physicians, 97 responded, resulting in a 61.4% response rate. Seven of the nonrespondents were in offices that were undergoing reorganization, and the remaining 54 refused to respond for other reasons.

The HMO roster listed 180 physicians in 1994, 33 of whom were ineligible. Most of the ineligible physicians had left their practices in the 6 months preceding the survey. Of the 147 eligible physicians, 101 completed the questionnaire, yielding a response rate of 68.2%. The 46 nonresponding physicians either refused or claimed that they had returned the questionnaire even though we never received it.

In addition to nonresponse to the entire survey, item nonresponse was a moderate problem in both survey waves. The 1992 questionnaire was 16 pages in length and contained 33 questions and subparts, constituting approximately 158 variables. Mean item nonresponse was 2.9 $\pm$ 2.4. Nine percent of the items (n = 15) received complete response, 40% (n = 63) were not answered by 1 or 2 respondents, 46% (n = 73) were not answered by 3 to 8 respondents, and 4.5% (n = 7) were left unanswered by between 10 and 15 respondents. There seemed to be no discernible pattern to the item nonresponse. One of the questions was on page 2 and asked the respondent to indicate the age at which he/she generally recommends performance of a skin exam with the patient fully undressed. Another question was on page 11 and asked, "To what extent is your day to day delivery of preventive services for cancer influenced by the National Medical Association?" The eight demographic questions had patterns of nonresponse similar to other questions. One demographic item was answered by all respondents, four items were not answered by one or two respondents, the gender item was not answered by three respondents, three respondents did not provide the year they graduated from medical school, and six respondents did not provide their year of birth.

The 1994 questionnaire was shorter, but it had a higher number of items for which there were no responses. The questionnaire was 10 pages in length with 22 questions and subparts, constituting approximately 95 variables. Mean

item nonresponse was 6.0 ± 3.5. Nine percent (n = 9) of the items elicited complete information from all respondents, 14% (n = 13) were not answered by 1 or 2 respondents, 59% (n = 56) had 5 to 9 missing respondents, and 18% (n = 17) were not answered by as many as 10 to 14 respondents. Questions with no missing data included those that asked the physician (a) to rate his/her ability to counsel and teach patients about breast and cervical cancer and (b) to describe the influence on day-to-day delivery of preventive services by medical journals, the American Cancer Society, the National Cancer Institute, and third-party payers.

Items with the highest nonresponse were asked on page 4. They requested the physicians to estimate the percentage of patients in their practices who were current and up-to-date with various cancer prevention or early detection procedures, such as clinical breast exam and mammography. The high proportion of nonresponse so early in the questionnaire indicates that the question tasks were most likely the reasons for item nonresponse. The completion rates for the demographic questions were better than for most other items. Three demographic questions were answered completely, three were not answered by one respondent, and one was not answered by two respondents.

Stability of the medical staff in the practices included in the study had implications both for continuity of care and for the research design, in addition to the implications for survey nonresponse. In practices with high turnover, a panel design of individuals, for example, is not recommended. Collecting data at the organizational level was expected to provide better information about practice continuity.

Table 1 shows that three physicians was both the mode and median number of primary care physicians per practice. In 1992, 25% of the practices had 1 or 2 physicians, about half had 3 or 4 physicians, and 27% had between 5 and 25 physicians. Some downsizing in the number of primary care physicians at the individual practices occurred between 1992 and 1994, even though both the median and mode remained the same. In 1994, 44% of the practices had 1 or 2 primary care physicians, 25% had 3 physicians, 8% had 4, and 23% had between 5 and 27 primary care physicians.

Table 1. Number of physicians per clinic in 1992 and 1994

	1992		1994	
No. physicians	No. clinics	%	No. clinics	%
1	5	10	10	21
2	7	15	11	23
3	14	29	12	25
4	9	19	4	8
5	3	6	3	6
6	4	8	2	4
7–27	6	13	6	13
Total	48	100	48	100

Table 2 provides an indication of the amount of turnover in the 2-year period. "Turnover" is defined as the number of primary care physicians that were added to or left the practice (Price & Mueller, 1986). Six practices (12%) had no physician turnover in the 2-year period. Nine practices either increased or decreased by 1 physician, 13 added or lost 2 physicians, 19 practices gained or lost between 3 and 8, and in 1 practice there was a turnover of 23 physicians. Tables 1 and 2 provide some indication of the amount of change in the practices, but they do not describe overall stability or retention.

Table 3 shows the percentage of physicians in 1992 that were still at the practice in 1994, which indicates the retention rate by practice. Eleven practices (23%) had the same physicians in 1994 as they did in 1992, 12 practices (25%) maintained between 57% and 91% of the physicians they had in 1992, 10 practices (21%) retained at least half of the physicians from 1992, 11 (23%) retained between 17% and 43%, and 4 practices (8%) lost all of the physicians they had in 1992. In total, by 1994, 40.7% of the physicians who were listed in the 1992 HMO directory had left the practices.

Many of these figures are similar to data reported by Willke (1991) concerning the practice mobility among physicians under age 40. He reported that 34.7% of those physicians who were in practice for 2 through 5 years had changed from their initial practice, that 31% of those in primary care had changed, and that 46.8% of those first employed by HMOs had changed. Foreign medical graduates were least likely to change practices. The race of

Table 2. Physician turnover at clinics, 1992–1994

No. physicians left or added	No. clinics	%
0	6	12
1	9	19
2	13	27
3	5	10
4	5	10
5–6	6	13
7–8	3	6
23	1	2
Total	48	99

Table 3. Physician retention by clinics, 1992–1994

% physicians that stayed	No. clinics	%
100	11	23
57–91	12	25
50	10	21
17–43	11	23
0	4	8
Total	48	100

Willke's sample was 82% white, 4% black, 4% Hispanic, and 10% other. Although our findings are similar, the characteristics of our sample are considerably different from those reported by Willke. Our sample is not of recent medical school graduates, is much older, and is more ethnically diverse. The median age of our physician respondents in 1992 was 48, and in 1994 it was 46. With regard to ethnicity, in 1992, 35% of our respondents were Asian; 34% were white, non-Hispanic; 20% were black, non-Hispanic; 8% were Hispanic; and 3% were other.

The medical practice survey was conducted at about the same time as the first physician survey. This survey asked questions about the practice as an organization: Who made decisions and set policy? What were the responsibilities and tasks of relevant personnel? What types of patients were seen? What were the most common health problems? These data were collected by a mailed questionnaire, 15 pages in length and containing 32 questions and subparts or about 100 variables. Data were collected over 3 months, and every nonrespondent was contacted at least five times by telephone. At the time of the survey, 50 of the 53 practices identified by the HMO were participating in the study: Two had been dropped because they had no adult patients, and 1 had left the HMO. Questionnaires were received from 37 practices, resulting in a response rate of 74%. Sixty-two percent of the respondents (n = 23) were office or business managers, medical assistants, or other nonmedical staff; 38% were physicians (n = 5) or nurses (n = 9).

Item nonresponse was lower than in the physician surveys. The mean item nonresponse was 1.5 ± 2.3. Fifty-three percent of the items obtained a complete response, 27% had one or two missing responses, 15% had between three and seven, and 5% had eight or nine missing responses.

Assessing Nonresponse Bias

Currently, a number of sources are being consulted to assess the potential bias due to nonresponse in the 1992 physician survey. Using the HMO physician directory, we compared respondents and nonrespondents by their gender, medical specialty, and the number of physicians in their practices. There was a statistically significant difference in response rates by specialty, with family and general practitioners more likely to respond than internists, obstetricians, and gynecologists. To assess the effects of nonresponse on other key variables, such as ethnicity/race, year of birth, year graduated from medical school, country of medical school, and board certification, we consulted the directories of the AMA (1992) and American Board of Medical Specialties (ABMS; 1994). In addition, the HMO medical director has been asked to supply other missing information from physician resumes that the HMO may have on file.

Dealing With Missing Data

Because the unit of analysis in this intervention is the organization, data quality and response rates assume considerable importance in the resulting analyses. When the eligibility and access issues were settled, there were only 47 practices available for analysis. It was critical, therefore, that data collection efforts achieve very high cooperation, since low or even moderate unit and item nonresponse would seriously limit the capacity to assess the outcome and the factors that predicted the outcome. Without such data, the results would be of limited value in making recommendations based on the intervention.

As we noted in the introduction, we hypothesized that a key issue in predicting whether screening was offered to patients was whether the practice had a written policy regarding screening procedures. Thus, one question asked in both the medical practice survey and in the physician survey was, "Does this [your] practice have an established, written policy for preventive care consisting of specific services at appropriate intervals?" The response categories were "yes," "no," and "don't know." Recall that 37 out of 50 practices responded to the survey. However, since 47 practices completed the intervention, we actually needed information about this policy for 47 practices.

Of the 47 practices that completed the intervention, 11 did not respond to the survey. In addition, three other practice respondents replied "don't know" to the question. The combined unit and item nonresponse resulted in a 30% (14/47) loss rate for this information at the practice level. Excluding 30% of the patient records (n = 805) from one component of the analysis would seriously hamper the study analysis.

To resolve this problem, we decided initially to impute a response for the missing practices from the answers to the physician survey. After further consideration, we decided to create a totally new variable by comparing the practice survey responses to the physician survey responses from the same practice. The cross-classification process produced five response pattern groupings, as shown in Table 4. Responses from both the physicians and the practices were consistent in 12 sites; the responses from practices and physicians in 10 sites showed inconsistent patterns, which varied from slight to considerable disagreement among physicians within sites; there was total disagreement between the practice response and the physicians' responses at 7 sites; at 13 sites, incomplete or uncertain practice data made it impossible to compare practice and physician responses; and at 5 sites, there were practice responses but no physician respondents, once again making comparisons impossible. Based on these patterns, a new categorical variable with four categories for classifying practices was created: "Yes, the practice has a policy," "No, the practice does not have a policy," "The responses were inconsistent," and "don't know." When the practice and the physician respondents totally agreed, the practice was classified using the agreed upon response category. When the practice and the majority of physicians were in agreement, the practice was classified according to the majority response. When the practice and physicians disagreed or the practice did not respond and the physicians who responded were not in

Table 4. Comparison of responses from practice and physician respondents (Q: Have preventive care policy?)

Practice data	Physician data	Coded
Group 1: 12 Sites with consistent answers		
Yes	2 yes/2[a]	Yes
Yes	1 yes/1	Yes
Yes	1 yes/1	Yes
Yes	2 yes/2	Yes
Yes	1 yes/3	Yes
No	1 no/2	No
Yes	1 yes/2	Yes
No	4 no/4	No
No	3 no/3	No
No	2 no/2	No
Yes	1 yes/2	Yes
Yes	1 yes/1	Yes
Group 2: 10 sites with some inconsistency across respondents, especially among physician respondents		
No	1 yes 1 no/3	Mixed
Yes	3 yes 1 no/4	Yes
No	2 yes 1 no/6	Mixed
Yes	1 yes 1 no/2	Mixed
Yes	1 yes 1 no/4	Mixed
Yes	1 yes 1 no/3	Mixed
Yes	1 yes 1 no/2	Mixed
Yes	2 yes 1 no/3	Yes
No	2 yes 7 no/9	No
No	4 yes 3 no/11	Mixed
Group 3: 7 sites with no consistency across respondents		
Yes	1 no/3	Mixed
Yes	1 no/2	Mixed
Yes	1 no/1	No
Yes	1 no/2	Mixed
Yes	1 no/4	Mixed
Yes	1 no/2	Mixed
No	3 yes/4	Yes
Group 4: 13 sites with incomplete or uncertain practice data and fairly consistent physician data		
No data	2 no/3	No
No data	1 yes/2	Yes
No data	1 yes/1	Yes
No data	1 no/1	No
No data	1 no/1	No
Don't know	1 no/1	No
No data	1 yes 1 no/4	Mixed
No data	2 no/3	No
No data	3 no/8	No
No data	8 yes 6 no/21	Mixed
Don't know	1 no/1	No
No data	1 yes/1	Yes
No data	1 yes/8	Yes
Group 5: 5 sites with no physician data to compare with practice data		
Yes	No data/2	Yes
No	No data/2	No
Yes	No data/1	Yes
Don't know	No data/1	Don't know
No	No data/2	No

[a]Number of eligible physicians at the site who were sent a questionnaire.

agreement and there was no clear majority among the physicians, the practice was coded as "mixed." Finally, when practice data were available but there were no physician respondents, the practice response was coded. The new variable classifies 17 practices as having a policy, 15 as not having a policy, 14 as "mixed" or indeterminate, and 1 as "don't know." In only one practice was it impossible to make any determination. By coding the new variable into two dummy categories with "mixed" as the reference category, we will be able to compare practices that do or do not have a policy with practices where the situation is ambiguous.

Conclusions

A number of things can be learned from this research. First, we discuss design implications. Interventions are typically conducted over a period of months or, at times, years. Therefore, attrition and turnover can be serious problems. We lost six practices (11%) over a 30-month period. The major losses, however, came at the individual level. Within a 2-year period, approximately 41% of the physicians left the practices through which they were originally contacted. We do not have data on staff members, but our impressions are that at least a similar percentage of staff also left during the study period. This level of turnover, in addition to unit and item nonresponse, does not allow for panel designs in which individuals are the units of analysis. For example, we had originally planned to use change in physicians' attitudes over the intervention period as a predictor of screening rates. This will not be possible given the levels of physician turnover and nonresponse. On the other hand, the organization may be the most appropriate unit of analysis for this evaluation.

Moderate amounts of unit and item nonresponse cause serious problems. Eleven of the practices that completed the intervention did not respond to the practice survey; three others responded "don't know" to a key question. This level of nonresponse had the potential for making 30% of the patient medical records data unusable. Fortunately, we were able to construct a surrogate variable by combining responses from two different surveys. The new variable should have higher validity than the responses to either of the questions that were asked in the medical practice and physician surveys from which it is created.

To assess potential sources of unit nonresponse, independent sources of information are being consulted. For the physician surveys, we are using background and demographic information from the AMA and ABMS directories in addition to the physician resumes on file at the HMO central office. These data will be used to compare nonrespondents with respondents. For comparing medical practice survey nonrespondents with respondents, we are using data from the HMO directory and information from informants in the HMO central office who frequently visit the practices. In addition, data from the 1990 Census are being used

to profile the community/neighborhood characteristics of each practice.

We have two suggestions for future studies of similar populations. The first recommendation is to plan for frequent communication between the research staff and the practices and to allow more time and staff to collect data than in an average study. Obtaining data from practices that have diverse administrative staff, high employee turnover, staff with low computer skills, and inadequate space and that use manual record keeping requires extensive interaction between the research team and the practice staff. Very seldom, if ever, do these practices participate in research projects. Adequate staff time is needed to explain fully on more than one occasion the purposes and benefits of the research. When visits are made or questionnaires are sent to the practices, a natural reaction on the part of the staff is to assume they are being evaluated. This issue needs to be clarified early on, and these types of potential tension-raising events must be addressed. This is especially critical when numerous contacts will be made during the research project. It is important to communicate with both the doctors and staff and to work within the limitations and needs of the practices. Frequently, the research time schedule cannot be imposed on the practices. For example, at many of our participating medical practices, space was at a premium, and there were severe difficulties in scheduling adequate time for patient records abstraction. Many practices allowed us to abstract for only one 2-hour period per day; others allowed us only one half-day per week. As a result, we were seldom able to develop an optimum schedule in which abstractors could make sequential visits to practices located in the same sections of the city.

We used various incentives to encourage participation and cooperation. Gifts such as pens and coffee mugs were given to all physicians and staff members after each survey. Letters of appreciation were sent to each practice with a copy to the HMO office following each major phase of data collection. We did not try monetary incentives such as offering $50 to nonrespondents because by having the cooperation of the HMO medical director, we did not believe it was necessary, and we were afraid it would set a precedent for the other data collection efforts. A recent paper on participation in a longitudinal study indicates that respondents do not expect remuneration for subsequent waves after being paid an incentive for their participation in the first wave (Lengacher, Sullivan, Couper, & Groves, 1995). It would be particularly worthwhile for future studies to explore whether an incentive can shorten the data collection period and/or reduce the amount of staff time required to collect the information.

Other studies have shown that monetary incentives to physicians do increase response rates (Gunn & Rhodes, 1981; Berry & Kanouse, 1987; Aday, 1989). Our survey response rates without monetary incentives were similar to surveys that do pay physician respondents. However, in intervention studies, in which there are usually a small number of units, typical levels of survey response are inadequate. In retrospect, we believe that monetary incentives would have increased our survey response rates. What is unclear, however, is how much higher the rates would have been and whether it would have been worth the cost.

Finally, we believe that collecting data from multiple sources is an important strategy in intervention studies. Having data from multiple sources allowed us to turn a study deficiency into a method whereby we could create practice level variables with higher potential validity than those from any single source.

References

Aday, L. A. (1989). Designing and conducting health surveys: A comprehensive guide. San Francisco: Jossey-Bass.

American Board of Medical Specialties. (1994). The official ABMS directory of board certified medical specialists. Evanston, IL: American Board of Medical Specialties.

American Medical Association. (1992). Directory of physicians in the U.S. Chicago: American Medical Association.

American Medical Association. (1994). Physician characteristics and distribution in the U.S. Chicago: American Medical Association.

Berry, S. H., & Kanouse, D. E. (1987). Physician response to a mailed survey: An experiment in timing of payment. Public Opinion Quarterly, 51, 102–114.

Gunn, W. J., & Rhodes, I. N. (1981). Physician response rates to a telephone survey: Effects of monetary incentive level. Public Opinion Quarterly, 45, 109–115.

Lengacher, J. E., Sullivan, C. M., Couper, M. P., & Groves, R. M. (1995, May). Once reluctant, always reluctant? Effects of differential incentives on later survey participation in a longitudinal study. Paper presented at the American Association for Public Opinion Research, Fort Lauderdale, FL.

Price, J. L., & Mueller, C. W. (1986). Absenteeism and turnover of hospital employees. Greenwich, CT: JAI Press.

U.S. Bureau of the Census. (1993, August). Population and housing characteristics for Census tracts and block numbering groups, Chicago, IL PMSA. Washington, DC: U.S. Government Printing Office.

Willke, R. J. (1991). Practice mobility among young physicians. Medical Care, 29, 977–988.

Evaluation of the American Stop Smoking Intervention Study

Larry G. Kessler, Marcia Carlyn, Richard Windsor, and Laura Biesiadecki
for the members of the ASSIST Evaluation Work Group

Introduction

Background to ASSIST

The American Stop Smoking Intervention Study (ASSIST) is the largest, most comprehensive public health smoking control project ever undertaken in the United States. In October 1991, 17 state health departments were awarded ASSIST contracts: Colorado, Indiana, Maine, Massachusetts, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Rhode Island, South Carolina, Virginia, Washington, West Virginia, and Wisconsin. These 17 ASSIST states have a combined population of 91 million people, or slightly more than a third of the total U.S. population; 23 million are children and adolescents, and nearly 20 million are regular tobacco users. These states also contain significant minority representation; more than 10 million in their populations are African American, and 7 million are Hispanic or another racial/ethnic minority group.

As a demonstration project and not a controlled trial, the vast majority of ASSIST fiscal resources provide direct support of smoking control interventions at the state and local levels. The resources devoted to evaluation are quite modest relative to the large budget of the total program; thus, in designing the evaluation, existing databases are used where possible. Designing an evaluation of such a large, complex program with limited resources has proven a considerable challenge. The multifaceted evaluation, described briefly, presents a combination of data collection strategies, including the use of traditional health surveys, that provides an opportunity to learn about ASSIST and other antitobacco campaigns in the United States.

Goals and Objectives

ASSIST's overall purpose is to demonstrate that a widespread, coordinated application of the best available strategies to prevent and control tobacco use will significantly reduce the prevalence of smoking and tobacco use. The primary goal of ASSIST is to reduce cigarette smoking prevalence in ASSIST sites to no more than 17% of adults by 1998 and thereby play a significant role in achieving the year 2000 goal of a 15% smoking prevalence rate for adults.

A network of state and local coalitions provides a mechanism for delivery of the application of antitobacco strategies. ASSIST coalitions are charged with developing and implementing strategies through a variety of channels in order to reach priority populations. Five channels are targeted by ASSIST: the overall community environment, work sites, schools, health care settings, and community groups. ASSIST's primary and secondary goals and the specific program objectives for each channel are presented in Figure 1.

ASSIST Evaluation Design

The basic hypothesis underlying ASSIST is that the 17 ASSIST sites, with their network of state and local coalitions, will prove to be more effective than non-ASSIST sites in reducing smoking prevalence among adults and youths. Because ASSIST is not operating within an experimental or quasi-experimental context, such differences will be difficult to demonstrate.

The framework for the ASSIST evaluation is presented in Figure 2. The ASSIST model assumes that the types of interventions conducted by ASSIST sites (media, policy, and promotion of program services) will influence attitudes toward tobacco use and will result in the adoption of stricter tobacco control policies. Changes in these areas will be measured in order to assess whether progress is being made toward achieving the ASSIST program objectives relevant to each channel. The model also assumes that changes in social norms and policies will, in turn, stimulate individual behavior change in tobacco use and reduce smoking prevalence. For example, significant reductions in smoking prevalence are expected to create an environment that is conducive to additional changes in social norms and an increased demand for smoking cessation services.

Challenges of the ASSIST Evaluation

The evaluation of the ASSIST project is particularly challenging for a variety of reasons, including potential site

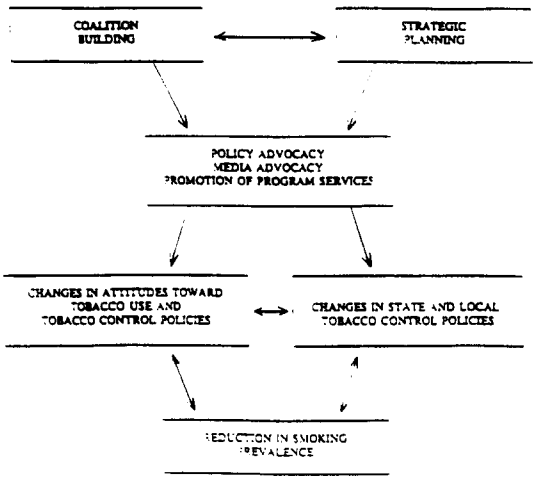
Larry G. Kessler is with the Applied Research Branch, National Cancer Institute, Bethesda, Maryland; Marcia Carlyn is with VC&A, Poolesville, Maryland; Richard Windsor is with the University of Alabama, Birmingham; and Laura Biesiadecki is with Prospect Associates, Rockville, Maryland.

selection bias, the high probability of diffusion from ASSIST to non-ASSIST sites, the inability to control a variety of other events occurring independently of ASSIST that could have a significant effect on attitudes toward smoking and tobacco control policies as well as smoking

Figure 1. ASSIST outcome goals and program objectives

	ASSIST outcome goals
Primary goal (adults)	Reduce smoking prevalence in adults aged 20 and older in ASSIST sites from the baseline rate of 25% (as measured by the September 1992 Current Population Survey) to achieve a smoking prevalence rate of 17% by 1998.
Secondary goals (youth)	Reduce smoking initiation in ASSIST sites, as measured by prevalence among individuals aged 18–24, to one-half the 1987 rate of 27% (as measured by the National Health Interview Survey) by 1998.
	Reduce smoking prevalence among individuals aged 12–18 in ASSIST sites to one-half the 1989 rate of 16% (as measured by the Teenage Attitudes and Practices Survey) by 1998.
	Reduce the prevalence of smokeless tobacco use in ASSIST sites to no more than 4% among males aged 18–24 by 1998.
	ASSIST program objectives
Community environment	By 1998, sites will substantially increase and strengthen public support for policies that (a) mandate clean indoor air, (b) restrict access to tobacco by minors, (c) increase economic incentives to discourage the use of tobacco products, and (d) restrict the advertising and promotion of tobacco.
	By 1998, cues and messages supporting nonsmoking will have increased, and prosmoking cues and messages will have decreased.
Work sites	By 1998, the proportion of work sites with a formal smoking policy that prohibits or severely restricts smoking at the workplace should increase to at least 75%.
	By 1998, work sites reaching major target populations will adopt and maintain a tobacco use cessation focus.
Schools	By 1998, 100% of schools serving grades K–12 and public vocational/technical/trade schools will be tobacco free.
	By 1998, 100% of all schools serving grades K–12 will be using efficacious tobacco use prevention curricula.
Health care settings	By 1998, all public health facilities, both outpatient and inpatient, will have enforced smoke free policies.
	By 1998, at least 75% of primary medical and dental care providers will routinely advise cessation and provide assistance and follow-up for all of their tobacco-using patients.
Community groups	By 1998, major community groups and organizations that represent the priority populations and have broad-based, statewide reach should be involved in ASSIST activities.

Figure 2. ASSIST framework for evaluation implemented through a variety of channels in order to reach priority populations



prevalence, the substantial size of the project (over $120 million from the National Cancer Institute [NCI] over 10 years), and the lack of national data sources to measure individual states' progress toward the achievement of many of the intermediate end points. The first three issues are addressed in more detail.

Potential Site Selection Bias

States were awarded ASSIST contracts primarily on how well their proposals demonstrated the capability to form effective coalitions that could deliver the three classes of interventions to reach populations having the highest smoking rates. Therefore, it is quite probable that ASSIST sites differ from non-ASSIST sites with respect to their ability to construct comprehensive smoking control plans. In addition, the site selection method also favored other characteristics related to smoking prevalence and population structure (Hall, Hershey, Kessler, & Stotts, 1992; Stotts, Kessler, Hershey, Hall, & Gruman, 1994), resulting in differences between ASSIST sites and non-ASSIST sites with respect to population demographics, tobacco use practices and trends, and preintervention levels of tobacco control activities and policies. In order to increase the power and possibly reduce bias in the detection of intervention effects, an analysis was conducted to assess the utility of selecting a sample of non-ASSIST sites to compare with the 17 ASSIST sites with respect to these factors.

Diffusion

A large amount of natural contamination is expected throughout the duration of the ASSIST project. A long-term increase in tobacco control activities and an accompanying decrease in tobacco use has been occurring for almost 30 years, and these trends are likely to continue in non-

ASSIST sites as well as those funded by ASSIST. In addition to the NCI and the American Cancer Society (ACS), a number of other federal, state, and voluntary organizations have targeted tobacco control as a major focus of their public health efforts, and it is quite likely that ASSIST methodologies and materials will be used by non-ASSIST sites to help them achieve their tobacco-related goals. For example, at the national level, the ACS is encouraging its divisions and units in non-ASSIST states to focus their resources on tobacco control, making no attempt to restrict the distribution of ASSIST methodologies and materials to only those ACS divisions and units located within ASSIST states. In fact, dissemination of tobacco control materials has been encouraged by the NCI to help achieve the national year 2000 goals.

In 1993, the Centers for Disease Control and Prevention's (CDC's) Office on Smoking and Health announced a $3 million program to help non-ASSIST states increase their capacity for smoking control, with the possibility of additional funding during the second year of the program. Also in 1993, the Robert Wood Johnson Foundation announced a new $10 million, 4-year tobacco prevention and control program to support efforts in an estimated 18 states to reduce tobacco use, particularly among youths, although these grants may go to ASSIST states.

Other Competing Activities and Events

In addition to diffusion, a variety of other events and competing influences are expected to occur independently of ASSIST and make it more difficult to measure the effectiveness of the ASSIST interventions. For example, the citizens of Massachusetts, one of the 17 ASSIST sites, voted for an increase in the state cigarette tax of 25¢ per pack, which took effect on January 1, 1993, prior to the ASSIST implementation phase. Although the ASSIST evaluation will include special studies to identify and analyze major competing influences, such as excise tax increases and media campaigns implemented independently of ASSIST, it will be extremely difficult to control for the variety of alternative hypotheses that could account for changes in tobacco use.

These are common challenges in the evaluation of demonstration programs with no easy answer. The ASSIST evaluation approaches this problem by collecting data from varied sources about the structure of ASSIST activities, the process of antitobacco activity at the state level, and tobacco prevalence as the main programmatic outcome. The limited scope of this paper precludes a detailed exposition of the evaluation components. Here, the basic components of the evaluation are presented, and the analytic strategy to handle the basic nonrandomized demonstration problem is outlined.

Outcome Evaluation

Measuring Changes in Smoking Prevalence

The primary goal of ASSIST is to reduce smoking prevalence in adults. The Current Population Survey (CPS) will be used as the principal means for measuring achievement of this goal. The CPS is a household sample telephone survey of the civilian noninstitutionalized population, conducted at regular intervals since 1950 by the U.S. Bureau of the Census to provide estimates of employment, unemployment, and other characteristics of the population as a whole, the general labor force, and various other subgroups of the population. It was chosen for ASSIST because it is the only ongoing survey funded by the federal government that provides a sufficient sample size to detect relatively small differences in changes in smoking prevalence between ASSIST and non-ASSIST states as well as to yield state specific estimates. It also includes state specific tobacco use trend data from previous years.

The Tobacco Use Supplement (TUS), a special supplement to the CPS sponsored by the NCI, was developed in 1990 for the ASSIST project and includes questions about attitudes toward tobacco use as well as individual patterns of smoking and smokeless tobacco use. The supplement consists of 41 self-report items that are asked of persons residing in sampled households. The NCI contracted for the CPS baseline survey to be conducted in three waves—in September 1992, January 1993, and May 1993—with approximately 115,000 individuals being interviewed for each wave. Although each of the three waves has only a modest sample size for use at the individual state level, the three waves taken together provide meaningful analyses of baseline data for each ASSIST site. CPS analyses will have ample power to detect a 2.5% or greater reduction in smoking prevalence in ASSIST sites as a group compared with the group of non-ASSIST sites, but there will be insufficient power to assess whether an individual ASSIST site has achieved a prevalence reduction of this magnitude when compared with all non-ASSIST sites.

Baseline data from the 1989 and 1992–93 CPS are shown in Tables 1 and 2 along with data on demographics. These data show the relatively consistent decline in tobacco prevalence, which has generally persisted. Trends by age, gender, and race show differing patterns but are not presented here for parsimony.

A secondary source of data for tracking changes in smoking prevalence is tobacco consumption estimates that are based on tobacco sales tax data and compiled on an annual basis by the Tobacco Institute. The ASSIST evaluation will include analyses of tobacco consumption data because there is evidence that measures of cigarette consumption may be more sensitive than prevalence measures to intervention effects.

Process Evaluation

Tracking Progress in Achieving Program Objectives

The framework for the ASSIST evaluation, presented in Figure 2, assumes that coalition building and strategic planning will be instrumental in the design and implementation of interventions focusing on policy, media, and the

Table 1. ASSIST and matched and unmatched non-ASSIST cohorts: 1992

Matching variables	Year	ASSIST cohort E N = 17	Non-ASSIST cohort C N = 34	Non-ASSIST matched cohort C_1 N = 17	Non-ASSIST unmatched cohort C_2 N = 17
Smokers	1992	25.2%	25.1%	25.3%	25.0%
Female smokers	1992	23.3%	22.9%	23.3%	22.5%
Male smokers	1992	27.3%	27.8%	27.6%	28.0%
Consumption[a]	1992	100.1	98.4	100.5	96.2
Cigarette tax rate[b]	1992	$0.24	$0.26	$0.27	$0.25
Black	1990	9.8%	11.6%	10.4%	12.7%
Hispanic	1990	5.1%	4.4%	4.9%	3.8%
Living below poverty line	1992	13.8%	14.3%	13.3%	15.1%
High school dropout	1990	10.1%	10.5%	10.4%	10.5%
Illiterate, aged > 20	1987[c]	12.1%	11.7%	12.0%	11.4%

NOTE: There are 17 ASSIST and 34 non-ASSIST sites—50 states plus Washington, DC.

[a]Number of cigarettes per capita per year.

[b]Averaged across states.

[c]Not available from the Department of Education for 1992.

Table 2. ASSIST and matched and unmatched non-ASSIST cohorts: 1989

Matching variables	Year	ASSIST cohort E N = 17	Non-ASSIST cohort C N = 34	Non-ASSIST matched cohort C_1 N = 17	Non-ASSIST unmatched cohort C_2 N = 17
Smokers	1989	26.1%	25.7%	25.9%	25.4%
Female smokers	1989	24.6%	22.8%	23.6%	22.7%
Male smokers	1989	27.9%	28.5%	28.4%	28.5%
Consumption[a]	1989	109.0	106.0	113.0	101.0
Cigarette tax rate[b]	1989	$0.21	$0.22	$0.22	$0.22
Black	1990	9.8%	11.5%	10.4%	12.5%
Hispanic	1990	5.1%	4.4%	4.8%	4.0%
Living below poverty line	1992	13.8%	14.3%	13.5%	15.1%
High school dropout	1990	10.1%	10.5%	10.5%	10.5%
Illiterate, aged > 20	1987[c]	12.1%	12.5%	12.0%	13.0%

NOTE: There are 17 ASSIST and 34 non-ASSIST sites—50 states plus Washington, DC.

[a]Number of cigarettes per capita per year.

[b]Averaged across states.

[c]Not available from Department of Education for 1989.

promotion of program services. Not all program objectives can be monitored, due to the lack of national data sources to measure individual states' progress toward the achievement of many intermediate end points.

The challenge for evaluating ASSIST was to find direct and indirect measures of program process objectives that were both nonintrusive (so as to minimize site burden) and that could be routinely collected on a statewide basis for all states. For certain objectives, this proved possible within the limited ASSIST evaluation budget. However, finding measures for other objectives became problematic, and program records and assessment by the NCI and other groups will have to suffice as a qualitative evaluation of the accomplishment of some objectives. Two of the major ASSIST evaluation components, legislative analysis and media tracking, are described in some detail below, while others are summarized more briefly.

Community Environment Objectives

There are two ASSIST program objectives for the community environment channel that relate to changes in the public's attitude toward tobacco control policies and changes in media coverage of tobacco-related issues, both of which are expected to be closely related to changes in social norms regarding the use of tobacco.

The State Cancer Legislative Database (SCLD), developed and maintained since 1989 by the Data-Based Intervention Research Program of the NCI, is the primary data

source for measuring changes in state tobacco control policies. The SCLD includes information about all enacted state legislation related to cancer control, including tobacco control. Information about each law, including an abstract describing the provisions of each law, is maintained in a single computerized record. With regard to the first community environment objective for ASSIST, the database currently tracks enacted state legislation related to tobacco vending machines, one means of restricting access to tobacco by minors. It also tracks legislation related to smokers' rights. To meet the needs of the ASSIST evaluation, the database is being expanded to include pending as well as enacted legislation. The ASSIST legislative analysis will include a content analysis of each piece of legislation tracked, using rating scales to quantify key aspects of the legislation, such as its breadth, restrictiveness, and enforcement provisions; actual enforcement will not be tracked, however.

One of the major interventions of ASSIST is media advocacy. The NCI hopes that during the implementation period of ASSIST, news coverage in the print media of legislative and policy issues that discourage smoking will be significantly greater in the 17 ASSIST sites than the 34 non-ASSIST sites, with increases through time in the number of articles in ASSIST sites supporting nonsmoking and decreases in the number of prosmoking articles.

Burrelle's Press Clipping Service collects data for the media analysis using a keyword search strategy to select articles appearing in U.S. daily newspapers that feature tobacco-related policies specified by ASSIST. A content analysis is then conducted by the ASSIST Coordinating Center to review the articles for relevance and sort them according to type of smoking policy (clean indoor air, restriction of access to minors, economic incentives, advertising and promotion of tobacco, or miscellaneous), point of view (prosmoking, antismoking, or neutral), type of article, and whether or not they appear on the front page of the newspaper. Although limited to print media, this methodology will permit comparisons through time of ASSIST and non-ASSIST sites in an unbiased and unintrusive manner. In addition to measuring changes in nonsmoking and prosmoking cues and messages, the extent and type of media coverage is expected to reflect community social norms regarding the use of tobacco.

Work Sites and Health Care Setting Objectives

The CPS TUS includes several questions to ascertain information about smoking policies and/or cessation services offered at the respondent's place of work. Some information on the achievement of the program objective on doctors and dentists discouraging tobacco use will be obtained from the CPS TUS. The ASSIST legislative analysis also offers a potential way to measure changes in publicly mandated smoking policies for health care settings, with state policies being tracked through the SCLD.

School Objectives

Although the ASSIST program has objectives for the school channel, no good source of information on a national basis is available to assess progress during this decade. A promising data source for tracking progress with respect to the school objectives is the School Health Policies and Programs Study (SHPPS). This new national survey of school policies and programs related to school health is sponsored by the CDC and supported by the ACS. Although it will not provide a definitive means for evaluating the achievement of the program objectives, it should prove to be very useful in analyzing progress during the 5-year implementation period and comparing ASSIST sites as a group with the group of non-ASSIST sites.

Community Group Objectives

Similar to school objectives, no useful data sources exist to measure progress with respect to the community group objectives. It is possible that program records maintained by the NCI and the ASSIST Coordinating Center plus the results of the ASSIST Coalition Assessment will provide qualitative information useful in understanding whether progress has been made in this area. Such qualitative data would help in the interpretation of relevant quantitative data obtained from the CPS and other sources.

Because the statewide coalition approach—using a network of state and local coalitions—is a relatively new concept in health promotion, a study is being undertaken to examine how this approach is being implemented in different contexts. The study is based on a conceptual framework of factors hypothesized to influence coalition effectiveness. The underlying theoretical proposition of the ASSIST Coalition Assessment is that certain environmental, structural, and functional characteristics of coalitions are indicative of their intermediate success as well as their long-term effectiveness in achieving ASSIST goals and objectives. The project focuses on the concept and experience of using state and local coalitions to implement tobacco control activities, rather than on the relative performance of individual sites.

Analysis Plans

The analysis of these diverse data presents a challenge for the ASSIST Evaluation Group. Each of the evaluation elements has a target objective related to the program to provide an assessment in that area. The principal challenge consists of developing an integrated framework for analysis to get a picture overall of the diverse ASSIST activities. The following discussion summarizes our preliminary plans in three areas: analysis of the CPS prevalence data, diffusion and lagged effects, and measurement of intervention exposure.

The CPS analysis involves multiple baseline assessments across a time frame approximating the intervals that will be

used for ASSIST. Because of potential site selection bias and suggestions made in the analysis of a community-oriented antitobacco project (Freedman, Green, & Byar, 1990), a matching approach to analysis was suggested to attempt to remove some of the selection effects of the process resulting in the ASSIST states. Tables 1 and 2 show an example of baseline CPS data from this matching procedure. However, preliminary power analyses (suggested by work of Martin, Diehr, Perrin, & Koepsell [1992] and using methods developed by Muller, LaVange, Ramey, & Ramey [1992]) compared this approach to a regression approach with CPS data. The findings suggest that the loss of states for comparison with ASSIST states (from 34 to 17) noticeably reduces power to detect differences over time. Therefore, the first analysis of the CPS data will use a regression approach rather than a matching analysis, with ASSIST participation serving as the treatment variable, and will examine differential changes over time.

One way of analyzing ASSIST is to investigate whether ASSIST states are leading the effort in building state and local antitobacco coalitions, with the effects of ASSIST showing up earlier than the effects of other competing efforts. The CPS data on knowledge and attitudes will allow the testing of hypotheses about these differences prior to looking for prevalence effects. In addition, data from both the SCLD and the media analysis will provide an opportunity to test the hypothesis that ASSIST states will lead the way in adoption of changes in social norms regarding tobacco use. If process measures such as these show ASSIST changes consistent with the underlying model, then tobacco consumption data, available on a monthly and yearly basis, can be analyzed using time series models to look for lagged effects. Unfortunately, little guidance exists about the nature and size of such a proposed lag, a fact that complicates hypothesis testing, although hypothesis development is possible.

The regression approach described earlier will allow the use of covariates in addition to the dichotomization of the country into ASSIST and non-ASSIST areas. This is necessary for several reasons already noted. ASSIST-like activities have been pursued by states outside and independent of the program. In addition, there are data available showing baseline differences in ASSIST states, which can be controlled for through the use of covariates.

Although the ASSIST evaluation remains at the core of the analysis, we plan to include in the evaluation additional major campaigns against tobacco and create an overall measure of exposure to antitobacco efforts at the state level that might be superior to measuring individual effects of each program. Variables will be constructed from several data sources to attempt to measure antitobacco activity at the state level. The principal sources for measuring state activity include the SCLD and the ASSIST media analysis. In addition, data from surveys of states by the Association of State and Territorial Health Officers that ask questions about their antitobacco activities might be combined with these other data. The object would be to develop one (or a few) measures of antitobacco activity that represent the overall level of exposure at the state level. After identifying and examining such process measures, the next step would be to attempt to detect whether ASSIST has had an effect in excess of changes occurring in non-ASSIST states.

If the evidence looks promising, then we can use this measure of antitobacco activity at the state level against which to regress the CPS prevalence and tobacco consumption data. Analyses of CPS data at the individual level can also be attempted, using state-based covariates describing antitobacco activity. However, a careful analysis of ASSIST versus non-ASSIST sites with regard to antitobacco activity will be necessary to appropriately interpret the CPS regression analyses because CPS data at the individual level do not permit trend analysis.

Conclusion

In conclusion, the challenges of the ASSIST evaluation are significant but well worth addressing in order to learn as much as possible about the delivery and impact of this major federal initiative. The development of new and efficient databases and the use of readily available data (state legislation and media information) along with health survey data from the CPS provide an opportunity to understand the complex relationship between social context, public health activity at the state level, and tobacco use. The availability of these diverse data on antitobacco activity promises to provide clues that have the potential to enhance program development in the future.

References

Freedman, L. S., Green, S. B., & Byar, D. P. (1990). Assessing the gain in efficiency due to matching in a community intervention study. Statistics in Medicine, 9, 943–952.

Hall, N., Hershey, J., Kessler, L. G., & Stotts, C. (1992). A model for making project funding decisions at the National Cancer Institute. Operations Research, 40, 1040–1052.

Martin, D. C., Diehr, P., Perrin, E. B., & Koepsell, T. D. (1992). The effect of matching on the power of randomized community intervention studies. Statistics in Medicine, 11, 1–10.

Muller, K. E., LaVange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association, 87, 1209–1226.

Stotts, R. C., Kessler, L. G., Hershey, J. C., Hall, J. G., & Gruman, J. G. (1994). Awarding contracts at the National Institutes of Health: A sensitivity analysis of the critical parameters. International Journal of Transactions in Operational Research, 1(2), 117–124.

Discussion of Session on Integrating Survey and Other Data

Steven B. Cohen

Introduction

A carefully designed data integration effort that combines survey data with other existing data sources will generally achieve gains in either data quality, analytical capacity, or both. These alternative data integration efforts can be characterized by levels of gradation in terms of the data elements that are linked between the core survey and alternative data sources. On one end of the continuum, a survey integration effort would result in data enhancement achieved by combining mutually exclusive data elements for the same sample units. At the other end of the spectrum, the resultant data quality for a given survey would be enhanced by supplementing survey data with administrative data on the same measures for the sample units. This data supplementation from administrative sources would help reduce levels of missing data in the host survey, expand the donor records available for imputation, and facilitate methodological investigations that examine the quality of the survey data. A variant of this data integration strategy is the linkage of comparable data elements for the same survey participants at two distinct points in time. This data integration plan is characteristically achieved when a larger survey is considered as the sampling frame for an additional survey effort with the data collected in both surveys used for analysis.

A well-designed survey integration effort also provides an opportunity to enhance the analytical capacity of the core survey with potential cost savings achieved through reductions in interview length and efficiencies in sample selection schemes. More specifically, consideration of an ongoing health care survey as a sampling frame for a subsequent health care survey that requires selective oversampling of policy relevant population groups will be more cost-efficient than the design and conduct of an independent sample that requires a separate screening interview to facilitate sample selection. In this setting, unnecessary data redundancies could be minimized by eliminating a significant number of questions that were asked in the host survey from the questionnaires to be administered in the subsequent survey effort.

It should be recognized that these data integration efforts are not without some risk. Minor differences in questionnaire wording across alternative data sources can introduce additional bias in resultant survey estimates. Unanticipated difficulties may arise in attempting to link survey data with alternative sources as a consequence of minor spelling differences in respondent names or errors in survey specific identifiers. Use of an existing survey as a sampling frame for a subsequent survey may result in a lower overall survey response rate than an independent survey effort. Often, the resultant analytical gains and related design efficiencies of a data integration effort outweigh the potential risks. In this regard, the papers presented in this session have emphasized the resultant gains achieved in each of the individual efforts to integrate surveys with other data sources.

Steven B. Cohen is Director of the Division of Statistics and Research Methodology at the Agency for Health Care Policy and Research in Rockville, Maryland.

Integration of Survey and Medical Records Data

The paper by Czaja, Manfredi, and Warnecke is directed to the problems encountered in the design and analysis of an intervention study aimed at improving cancer screening practices. Particular attention is given to the imputation strategies that were implemented to correct for high levels of missing data that characterized the practice survey. In this site specific study, the authors consider data supplementation strategies to correct for missing survey responses by pooling data from auxiliary sources, which include a related provider survey, Census data, medical directories, and HMO records. While the capacity to evaluate all the planned study outcomes is seriously constrained by the attrition encountered in the provider survey, the paper provides a number of insights with respect to the use of related auxiliary data to help reduce the bias in survey estimates attributable to both item and complete survey nonresponse. Rather than focusing on the key outcome measures that defined the study, however, the authors limit their focus to only one of the key independent variables of the study. The details of the study intervention are also not provided.

The practices that were selected for the study were not randomly chosen, but rather a set of 51 self-representing clinics located in Chicago's inner city that serve low-income minority populations. Consequently, the inferences that can be made with respect to the study intervention are constrained to this site specific population. The study design included a physician survey conducted at two time points—at baseline and at the end of the intervention. An

implied objective of the planned analysis of the physician survey data was to assess whether the study intervention altered medical practice with respect to cancer screening procedures. This would require a recontact with the same providers that were sampled in the sites at the baseline interview. In addition to an overall physician survey response rate of only 61.4% for the baseline survey, further compounded by item response rates, physician turnover between 1992 and 1994 was high, with less than half of the clinics retaining more than 50% of the physicians on staff by 1994. Perhaps an a priori knowledge of the interaction of the anticipated survey response rates with the high level of turnover for physicians in these practices would have resulted in a significant reduction in the planned scope of the physician survey.

The study also includes a practice survey, which was conducted at the same time as the first physician survey. Since the primary unit of analysis in this investigation was the organization, the combined impact of unit and item nonresponse would have resulted in a significant rate of loss of information. To correct for the missing data at the practice level for the key measure of analytic interest, an imputation strategy was considered that based the response for the practice on the response profile obtained from the physician survey for the same measure. Here, the use of auxiliary data to correct for missing response profiles is a standard approach that generally results in a reduction of bias associated with survey nonresponse. In the application that was considered, however, the researchers did not preserve the level of variation observed between physician and practice respondents on the measure under consideration. Under the imputation rule that was imposed, a site with missing practice data with consistent responses from physicians would be given the consistent physician data value. Relative to 19 sites with consistent physician responses, 37% of the practice responses were inconsistent with the physician data. The researchers should have considered data imputation strategies that are more robust than the approach taken to preserve the underlying observed variation in the response profiles observed for the respondents. Standard techniques such as hot deck or cold deck imputation strategies would have accomplished this and resulted in a potentially greater reduction in bias associated with nonresponse.

Matching Survey Data With Administrative Records

The paper by Eppig and Edwards provides a good example of the gains that can be achieved in improving the quality of survey data by linking Medicare Current Beneficiary Survey (MCBS) data with Medicare claims. The MCBS serves to provide national estimates of the annual health care utilization and expenditures by Medicare beneficiaries, including health care events not covered by Medicare. The primary rationale for matching to the Medicare claims data is to identify health care events not reported by survey respondents and help improve the quality of the expenditure data, particularly Medicare payments, that characterize the MCBS.

If both the MCBS and the Medicare claims data shared identical structures in the reporting of medical events, the identification of the matched and unmatched events would be achievable without the introduction of additional errors associated with alternative classification schemes. In actuality, the alternative data sources are characterized by a lack of consistent data elements, unreliable reporting patterns by survey respondents, and different file layouts and data elements on the Medicare claims for different services types. Having noted these limitations in attempting to match these two data sources, an analysis was conducted on a "one-ninth" sample of the 1992 MCBS survey data by event type. No details of the sample selection scheme were provided, but this review assumes that at a minimum, the sample that was selected contained the total MCBS utilization profile for a selected individual. This needs to be clarified in addition to the representativeness of the sample selected for this analysis.

The results of the matching study indicate a nonnegligible proportion of unmatched events between the two sources, even when controlling on a person basis for unduplicated survey-reported events. A first reaction to the number of events that were unmatched in each of the data sources suggests that if all of these cases were false negatives, the level of discordance between the two data sources would be minimal. Furthermore, a surprisingly large number of nonduplicative survey-reported events were identified in Eppig and Edwards's Table 4 for the medical provider events. Since the matching algorithm was based on a hierarchy of criteria that had been programmed, I was curious as to whether a truth set of matches had been determined and the matching algorithms tested against this truth set to determine the expected level of false matches and false nonmatches expected as a consequence of the matching rules that were implemented. If the expected number of false nonmatches were on the order of 10%, then the levels of nonmatches observed in the analysis would not be as disturbing as an initial observation would suggest.

The paper also includes a comparison of the MCBS matches to administrative data relative to data obtained from the 1987 National Medical Expenditure Survey (NMES). This particular analysis is misspecified, given the disparate aims of the respective methodological studies that have been contrasted. The data presented in Eppig and Edwards's Table 5[1] obtained from the MCBS-Medicare claims database represent the union of events for selected individuals from the two sources. Alternatively, the data from the 1987 NMES, also presented in their Table 5,[1] are conditioned on household-reported events. While the 1987 NMES included a Medical Provider Survey (MPS), which could be viewed as a source of data comparable to the Medicare claims data,

[1]This table appeared in Eppig and Edwards's original paper but was deleted in the revised version that appears in this proceedings.

the object of the survey was to serve as a data replacement source for medical expenditure survey data. As a consequence of the levels of nonresponse in obtaining permission forms to contact medical providers associated with household respondents and of the subsequent layers of provider nonresponse anticipated and realized in the survey, the attempt to obtain a complete utilization profile for sampled individuals in the NMES was not a design objective. A large portion of the missing data presented in Eppig and Edwards's Table 6[2] associated with the MPS data in the NMES for medical provider contacts is largely attributable to the fact that less than one-third of the household-reported events were eligible for an MPS follow-up. The NMES table from which the data were drawn was developed to indicate the availability of expenditure data from household respondents and medical providers relative to household-reported events. Consequently, statements such as, "The MCBS finds a much higher proportion of medical provider visits with dollars reported in both the administrative data and the survey,"[3] do not provide a correct comparison, given the selective nature of the NMES MPS sample. Otherwise, returning to the strengths of this investigation, further analysis of the MCBS match rates by reference period length and interviewer and respondent characteristics should inform future redesign decisions regarding desired reference periods and imputation strategies.

Survey and Administrative Utilization Data Integration

The presentation of the Aging in Manitoba Longitudinal Studies reflects a successful data integration effort that has served to enhance the analytical capacity of the panel survey component by linking the interview data to administrative health care utilization databases. Havens's paper provides a description of the design of the longitudinal survey in addition to the planned data collection effort for 1996 and a discussion of the content of the administrative data sets that have been used to augment the analytical capacity of the survey data. Another remarkable attribute of the longitudinal survey effort is the consistently low refusal rate that characterizes the respective survey contacts and the low rates of loss of sample as a consequence of movement out of Manitoba.

The utilization data that were merged into the survey data reflect health services utilization for insured services and administrative data on services directly delivered by the Manitoba Health Department or through the provincial home care program. The database also covers insured services consumed outside of Manitoba but does not include uninsured services obtained outside of the province. The further supplementation of the integrated database with electronic data from Manitoba Vital Statistics on information obtained from death statistics provides a rich source of information to facilitate analyses related to the health and health care utilization patterns of the elderly in Manitoba.

The plans for the 1996 survey effort call for a recontact with the longitudinal panel last surveyed in 1990. Under this plan, the most recent cross-sectional panel was selected in 1983, suggesting that the youngest members under consideration for analysis would be 78 years of age. Given the level of attrition that has characterized the longitudinal panel over 25 years, it is curious that an additional new cross-sectional sample has not been introduced in 1996. This new cross-sectional sample would serve to enhance future analyses and allow for cross-sectional comparison regarding the health status and the demographic composition of individuals that define the longitudinal panel with a representative current cross section of the elderly population in Manitoba. Furthermore, it has also been indicated that several relevant national surveys will be conducted in 1996 and will allow for comparisons with the elderly Manitoba population. However, the elderly Manitoba population represented by the longitudinal panel will not be representative of the current elderly population, particularly with undercoverage of elderly individuals who have moved into the province since 1983. Given the level of data enhancement achieved through this marriage of survey and administrative data, additional research should be conducted to determine whether existing province specific health care utilization databases are consistent enough in format and content to allow for data supplementation to the national Canadian health care surveys that are to be conducted in 1996.

Synthesis of Existing Survey Efforts to Evaluate an Intervention Program

The paper by Kessler, Carlyn, Windsor, and Biesiadecki provides an overview of the underlying framework and data collection efforts used to evaluate the American Stop Smoking Intervention Study (ASSIST). The research effort serves as an excellent example of economies that can be achieved by using existing surveys to assess the progress of the study intervention. More specifically, data from the Current Population Survey Tobacco Use Supplement will be utilized to assess whether the smoking prevalence for adults in ASSIST sites has been reduced to 17% in 1998. In a comparable fashion, data from the National Health Interview Survey Supplement on Cancer Epidemiology and Control will be used to determine whether smoking prevalence in teenagers is reduced by one-half the 1987 rate by 1998. Furthermore, the study will access the State Cancer Legislative Database to examine changes in state tobacco control policies and will examine press clippings for the media analysis component of the study.

The authors are careful to identify a number of inherent limitations of the design of the study and the planned

[2]This table appeared in Eppig and Edwards's original paper but was deleted in the revised version that appears in this proceedings.

[3]Quoted from the original paper; these words do not appear in the revised version included in this proceedings.

evaluation, which include potential site selection bias, the contamination of cases in non-ASSIST sites through diffusion, and the study's nonrandom design. In the planned analyses, a matching approach was considered to help remove some of the effects of site selection bias. The current plan appears to move away from this approach and recommends including all states in future regression analyses in order to improve the underlying power of the planned statistical tests. I would view the issue as a classic mean square error problem, in which the impact of bias on resultant estimates should be included in the final assessment rather than isolating the problem to one of only variance reduction. Out of curiosity, how would one evaluate the success of the intervention if the targeted reduction in the smoking prevalence rate was achieved in the ASSIST sites but an equivalent reduction in the prevalence rates was also achieved in the matched sites not subject to the intervention?

Linking Primary and Secondary Data for Outcomes Research

The research effort by Paul et al. provides an example of a three-way linkage of administrative, clinical, and patient-reported data to enhance analyses for the assessment of patient outcomes related to total knee replacement (TKR). The study used the existing Medicare Provider Analysis and Review (MEDPAR) files as a sampling frame to identify the universe of patients with Medicare-reimbursed TKR hospitalizations. The study database was further enhanced through linkage to the American Hospital Association Hospital Survey data to obtain additional organizational measures that characterized sampled hospitals and linkage to the Area Resource File for other supply-side analytical measures. Three independent samples were selected for the study—a national sample, an Indiana sample, and a sample from 29 counties in western Pennsylvania—with initial sample sizes of 750 patients for the national sample and 500 patients for each of the site specific samples. A first review of the sample sizes suggests limited power for planned analyses when the impact of nonresponse is factored in, particularly for the analytical subgroups identified. Furthermore, the study response rates were summarized across all of the three independent samples in addition to the sample that defined the linked analysis files. It is unclear whether the response rate results were consistent across samples, and given the disproportionate sample scheme considered across the sampling strata, the presentation of weighted study response rates would have been more informative about the resultant sample representativeness of the target population. In addition, a total of 15.1% of the overall target sample was found to be either deceased, ineligible, or incapacitated and not considered as eligible for the study. Given that a major component of the study is the assessment of patient outcomes, I would have thought that information on these patients may have been particularly insightful for the evaluation, and attempts at proxy responses for the decedents on their characteristics immediately prior to death would have been desired.

The study also obtained medical records for study patients from 87.8% of the eligible hospitals. This data allowed for a validation of the accuracy of the procedure and diagnosis coding on the Medicare claims databases. This patient specific information was obtained from the hospitals prior to obtaining permission from sampled participants. In cases in which patient permission was refused after the fact, medical records obtained for the sample participants were destroyed. The study, however, does not seem to have accommodated potential refusals of use of medical records information by survey nonrespondents. Perhaps the medical records component of the study should not have been initiated until permission to use medical records information was provided by study participants.

Summary

The papers presented in this session cover a broad array of ongoing integration efforts in the health field that link survey data with other related existing data sources to enhance analytical capacity. All of the authors in this session have helped expand our knowledge of both the inherent gains associated with data integration as well as the underlying potential limitations to the process.

Discussion of Session on Integrating Survey and Other Data: A Match Made in Heaven or a Shotgun Wedding?

Ronald Andersen

My remarks will concentrate on the papers by Kessler, Carlyn, Windsor, and Biesiadecki on tobacco use (American Stop Smoking Intervention Study [ASSIST]) and by Paul et al. on total knee replacement (TKR). I have subtitled these remarks on integration of surveys and other data "A Match Made in Heaven or a Shotgun Wedding?" to emphasize that successful integration depends on a good match between the data sources and the study design. Also, Diane and I are celebrating our 30th wedding anniversary. Ultimately, however, you will have to draw your own conclusions about the nature of all these matches—for marriages as well as integration papers.

My Figure 1 suggests there are four major reasons for integrating survey and other data: (a) First, and perhaps most importantly, multiple sources are required to carry out the study design. All key variables—be they independent, dependent, intervening, or control—can be measured only by incorporating more than one data source. (b) Comparisons of surveys and other data sources allow for the assessment and sometimes the improvement of the reliability and/or validity of study variables. (c) Integration of multiple sources into the design provides opportunities to examine threats to the internal validity of the design through the introduction of additional controls or intervening variables. And (d) comparisons of data from multiple sources increase the opportunities to apply the study findings to other population groups.

The thesis of my remarks (see my Figure 2) is that successful integration of alternative data sets depends on the degree of match with study design requirements. Elements of the match include (a) time—Do the alternative data sets cover the same time period? (b) place—Do the alternative data sets describe populations within the same geographical boundaries? and (c) unit of observation—Do the alternative data sets refer to the same people, families, organizations, or other units of observation?

Figure 1. Reasons for integrating survey and other data

1. Required to implement study design: Alternative sources measure key independent, dependent, intervening, or control variables
2. Establishes measurement reliability/validity
3. Establishes internal validity of design
4. Establishes external validity of design

Ronald Andersen is Chair of the Department of Health Services, University of California, Los Angeles, School of Public Health.

Figure 2. Successful integration of alternative data sets

Successful integration of alternative data sets depends on match with study design requirements according to

1. Time
2. Place
3. Unit of observation

We will discuss the success of ASSIST and TKR in matching and integrating data sources according to the major reasons for integration listed in my Figure 3, including (a) study design requirements, (b) measurement reliability and validity, (c) internal validity, and (d) external validity.

Study Design Requirements

Both the ASSIST and TKR projects have study designs requiring multiple data sets, including those from surveys and other sources.

The ASSIST design is especially complex, attempting to evaluate in 17 states a national demonstration program to reduce tobacco use, with special emphasis on youth, blue-collar populations, women, and minority populations. It includes both process and outcome components and various units of observation, such as individuals, counties, states, media releases, regulatory policies, and tobacco consumption reports. It would be virtually impossible to carry out

Figure 3. Assessing successful integration of data sets in the ASSIST and TKR studies

Reasons for integration	ASSIST	TKR
1. Study design requirements		
2. Measurement reliability/validity		
3. Internal validity		
4. External validity		

this evaluation without attempting to integrate multiple data sets.

How successful has ASSIST been in its creative effort to meld together these various data sources and match them to the design requirements? In part, it is too early to tell because the ASSIST evaluation is only in the second phase of a three-phase project. However, it appears that stronger matches have been attained for outcome measures of attitude change, state tobacco policies, and smoking prevalence, while some process measures regarding coalition building, strategic planning, and advocacy and promotion measures are, perhaps, proving more of a challenge.

The TKR study design is also a large-scale effort to match multiple data sources but is considerably less complex than ASSIST. The match includes Medicare claims, hospital records, and patient surveys at the individual level, with supplementary linking data provided through the American Hospital Association and Area Resource Files. While some rather serious threats to validity are present in the TKR design, the various sources appear to be quite well integrated and to meet the design's basic requirements.

Measurement Reliability/Validity

ASSIST has been able to put together measures from multiple sources that seemingly have a fair degree of measurement reliability and validity—particularly the outcome measures. The validity of some measures might be limited due to the uncertainty regarding lagged effects and the less-than-optimal fit between design requirements and availability of some secondary data sources.

The major comparisons are between the 17 ASSIST states and those states without ASSIST. Generally, there appear to be sufficient power to detect important aggregate differences between ASSIST and non-ASSIST states but not always to discover important differences at the state or local level.

The TKR study is providing some good opportunities to estimate and, in some cases, improve the validity and reliability of key diagnostic, procedural, and demographic measures. For example, some claims data have shown relatively high validity using medical records as criteria, and when one-third of the patient surveys failed critical edit checks, most of the respondents were recontacted and corrections could be made.

It is rather difficult to assess the power of the TKR samples since this paper does not deal specifically with key dependent variable analyses. However, the sample sizes appear adequate to detect important differences, given that analyses do not focus on highly stratified subgroups.

Internal Validity

Both projects face significant challenges in attempting to attribute outcomes to specific interventions. Neither has a randomized design, and neither is always able to control for serious threats to internal validity. However, one should probably not always expect especially strong internally valid designs in complex projects integrating multiple and often secondary data sources. Rather, their strengths lie in their multiplicity and diversity of information sources.

Kessler et al. document threats to internal validity including (a) diffusion of the innovation, as non-ASSIST states might adopt practices of the ASSIST states; (b) history and secular trends, as many other campaigns, regulatory efforts, and lifestyle changes are influencing tobacco use in the U.S. at the same time ASSIST is taking place; and (c) selection, as the states chosen for ASSIST may differ from non-ASSIST states in ways that will influence smoking trends independent of ASSIST. It is also of interest to note that ASSIST has adopted specific objectives, such as to reduce smoking prevalence of adults in ASSIST states to no more than 17% by 1998. It will be no easy feat to attribute such global changes specifically to ASSIST.

TKR also faces internal validity threats, but possibly of less magnitude than ASSIST since there is a more specific intervention (total knee replacement) being performed on specific patients. Also, as Paul et al. point out, the integration of multiple data sets allows for the introduction of many potentially important control variables in multivariate model building that could reduce threats to internal validity.

External Validity

Some internal validity is obviously necessary to be able to apply the findings of ASSIST and TKR about what works and does not work to other settings. However, these complex studies integrating surveys and other data might be considered relatively strong in terms of external validity. They can provide an array of relevant information and the documentation of intervention processes to others wishing to replicate their efforts.

One might ask, in the case of ASSIST, how relevant the findings in 17 selected states are to the others, and for TKR, how serious the exclusion of HMO and Veteran's Affairs patients might be in generalizing the results. However, they are both so comprehensive in the populations they cover and the types of information they collect that they potentially enjoy wide applicability.

Conclusions

ASSIST and TKR represent many of the strengths of studies integrating surveys and other data. TKR is a careful effort to match multiple data sources to study the outcomes of an important clinical procedure. ASSIST is an especially complex study requiring multiple data sets to richly document an ambitious effort to curtail tobacco use. Given the challenges to internal validity, I would caution some

moderation in the claims made and expectations for ASSIST in the conclusion to Kessler et al.'s paper. It will be somewhat difficult to attribute specifically to ASSIST the prevention or cessation of smoking among 6.5 million people or the avoidance of 1.2 million premature deaths. At an estimated cost of $150 million, I calculate the cost per averted death at $125—quite an astounding cost-effectiveness ratio!

Discussion of Themes From Session 5

Katherine Marconi, Rapporteur, and Richard Kulka, Chair

A major theme in this discussion concerned the selection of data sources to integrate with survey data. In his critique of the papers by Kessler, Carlyn, Windsor, and Biesiadecki and Paul et al., Andersen suggests that successful integration of alternative data sets depends upon the match with study design characteristics according to time, place, and unit of observation. Data sets to be integrated should be compiled during the same time period, be selected from observations from units in the same geographic area, and have the same unit of observation. Consequently, the use of data integrated from several sources is likely to constrain how the data can be used and the kinds of questions that can be addressed. Andersen's observations are applicable to all the papers in this session. For example, in the paper by Czaja, Manfredi, and Warnecke, it is noted that nonmatches at several stages of data collection constrain the amount of sample on which complete data are available and that there may be important differences between those that are matched and those that are not matched. Physicians who "disappeared" over the 2-year study period may, for example, differ from those who did not "disappear" in other ways related to the outcome variable. Thus, even if they are categorized using the imputation methods described by Czaja et al., these unobserved differences may affect the outcome results.

Similarly, in the discussion of Kessler et al.'s paper on the use of multiple data sources for the American Stop Smoking Intervention Study (ASSIST) evaluation, there was concern about the level of analysis and what could be said about the effectiveness of the program. As Kessler pointed out, the data are satisfactory for examination of trends in smoking at the national level comparing ASSIST states with states that do not have an ASSIST program. One cannot attribute causation to the program, but it is possible to note different patterns.

In their paper, Paul et al. illustrate another kind of trade-off in using certain secondary data sources, like the Medicare Provider Analysis and Review (MEDPAR) file, as a sampling frame. In this case, by merging the files, the authors got less statistical power than they had hoped for in some instances, but in other instances, they gained power. In particular, the Medicare file contained data on a larger-than-expected number of nonwhite recipients, which enhanced their power for doing analyses on nonwhite racial groups.

Mary Grace Kovar pointed to some other issues that need to be considered in studies such as those described in these papers. Raising a theme that had been considered in Session 1, on health care, she noted that gatekeepers would be an important concern, especially if the research concerns household samples and medical records. However, as noted in the discussion of the Harris, Tierney, and Weinberger paper in Session 1 and the discussion of the Park and Burt paper in Session 3, administrative policies may seriously affect access to administrative data. For example, in an effort to link Medicaid and Medicare data with data from the National Death Index and the Social Security Administration (SSA; Kovar, Chyba, & Fitti, 1992), she found that some agencies had standard, well-established procedures governing data access, whereas others did not. Hill reported similar experiences in dealing with the SSA, which requires individual written consent with the signatures having to be acquired within 2 months of when the data are needed.

Data quality in the various files was another concern raised by Kovar. She cited missing data on hospital admissions, data missing from entire states, and other problems in quality with the Area Resource Files (Stearns, Hayes, Koch, & Kovar, 1993). With regard to confidentiality, there is also the potential problem that the cell resulting from compiling data from several sources may contain so few cases that there is a risk of deductive disclosure.

The main question that arose from this session is "Why do it?" Andersen points to the fact that multiple sources are often required to carry out a study design. Key features of the study sometimes cannot be conducted without multiple data sources. Another reason is cost. Sometimes the savings derived from combining data sources are significant and the only way the study questions can be answered. This was certainly the case with the ASSIST evaluation and the project conducted by Eppig and Edwards. Sometimes it is the only way that the data can be obtained, or it enables better use of the data that are available, as is the case in all the studies described here. As noted by Steven Cohen, "A carefully designed data integration effort that combines survey data with other existing data sources will generally

Katherine Marconi is with the Bureau of Health Resources Development at the Health Resources Services Administrationin Rockville, Maryland. Richard Kulka is Research Vice President of the Statistics, Health, and Social Policy Unit at the Research Triangle Institute, Research Triangle Park, North Carolina.

achieve gains in either data quality, analytical capacity, or both."

However, as these studies also indicate, the potential gains are generally accompanied by significant costs in data quality, timeliness, and flexibility in analysis. Moreover, they require considerable effort, which is justified only if the costs of other approaches are also considerable.

Themes to Be Pursued in Future Research

1. Although we are becoming more sophisticated in the use of multiple sampling frames and the use of administrative data to supplement, replace, or correct survey data, we are less knowledgeable about techniques for data matching, dealing with duplicate data from multiple frames (see Session 3 summary), managing data quality in other data sets, and dealing with gatekeepers (see Session 1 summary). Thus, research is needed on these topics.
2. Good studies on the cost and efficiency of using multiple sources versus collecting original data in which the policy relevant outcomes of various strategies are compared would help a great deal. Exploration of the potential of meta-analyses to address this kind of question might be a good starting point.
3. As can be seen, several of the themes raised in the Session 3 summary are also appropriate here, cost being a primary one.

References

Kovar, M. G., Chyba, M., & Fitti, J. E. (1992). The Longitudinal Study of Aging: 1984–1990. Vital and Health Statistics, 1(28).

Stearns, S. C., Hayes, K., Koch, G. G., & Kovar, M. G. (1993). Reconciling respondent reports and medicare claims for national estimates of hospital use. American Statistical Association 1993 Proceedings of the Section on Survey Research Methodology: Vol. 1. 232–237.

Conference Conclusions and Wrap-Up

As with each of the preceding conferences, there are a number of themes that emerged from the discussion that seem to cut across all the sessions. These overarching themes center around three primary issues: measurement, survey designs for reaching rare or hard-to-interview populations, and mode effects. These themes seem to bring together issues of nonsampling error and respondent characteristics. From these themes, 11 broad areas where further research is required can be identified. They are listed in the sections below.

Measurement

Day 1 of the conference addressed measurement, first in the contexts of health status and patient satisfaction and then as a general theme.

1. There was a general feeling that theoretical work is needed that addresses wording effects and is capable of separating them from mode effects, interviewer effects, and other sources of error. New work by Sudman, Bradburn, and Schwarz (1996), for example, offers an important new view of the cognitive dimensions that affect interpretation of and response to survey questions. However, the conference participants also expressed the sense that the theoretical framework needs to be expanded to include what has been learned from the work on behavioral coding that focuses less on the respondent and more on the interviewer. Mode effects are also important for a full understanding of the quality of survey measurement. Mode effects are the major theme in Session 4 but also arose as part of the discussion of measurement in Session 2. The need to integrate these various aspects of the measurement process was a major concern expressed at various points throughout the conference. The publications that have focused on total survey error (Groves, 1989; Dillman, 1978) have attempted to address how these issues contribute to nonsampling error in surveys. However, there was a consensus that the effects of new interviewing technology and the introduction of the cognitive theories of response are not integrated into the existing standard works that treat nonsampling error and need to be.
2. Another theme that arose repeatedly in the measurement sessions is the recommendation that the psychometric literature be reintroduced into the literature on survey questionnaire design. Particularly during Session 1, which dealt with issues of patient satisfaction and health status, the discussion seemed to focus on the theme that measurement models need more explication than they have been given in the survey design literature, particularly with regard to measurement reliability, validity, and stability over time.
3. Based on the discussion of patient satisfaction, it is clear that cognitive theories could potentially contribute to an understanding of the issues related to patient satisfaction. Here, conversational norms, order effects, and other cognitive themes clearly help us understand how individuals respond to questions about their illnesses and satisfaction with treatment. Nevertheless, it was the consensus of the conference participants that "satisfaction" as a construct is not well understood and caution should be employed in its adoption as a standard for defining quality performance in health care delivery. While many thought satisfaction should not be employed at all as a dimension of quality, others expressed the opinion that it is a valid dimension of quality health care delivery but may be poorly measured.
4. The issues related to measuring health status, however, seem most related to psychometric issues, particularly to validity, reliability, and the time interval across which the measures were obtained.

Survey Design

Survey design occupies Sessions 3 and 5 of the conference. The Session 3 summary addresses sampling and respondent cooperation, and the Session 5 summary concerns the integration of survey and other data. The major issues in the discussion of sampling address approaches to sampling that would enhance the ability to estimate characteristics of hard-to-locate or otherwise difficult-to-access populations. On the one hand, the papers and discussion focus on obtaining access and valid samples; on the other hand, they address issues related to cost and

the trade-offs between cost and efficiency, although more needs to be said about these trade-offs. These themes recur in Session 5 but focus on what amounts to alternative approaches to estimating population characteristics when sampling strategies or response effects make direct estimation from samples costly or problematic.

1. A major theme of both session summaries is the need for studies that address cost and the relative trade-offs between cost and the quality of the resulting estimates that can be made. There is clearly a need for more studies that specifically address the cost-quality trade-offs of these various sampling strategies or the use of multiple data sets to arrive at estimates.
2. Another common point emerging from the two sessions is that with increasing sophistication in the use of multiple sampling frames and supplementing sample data with administrative data, there is a great deal more to be learned about techniques for matching data from diverse sources and dealing with duplication in multiple frames, managing the data quality in different data sets, and dealing with gatekeepers who control access to data that might allow for more efficient sampling designs and greater efficiency in data integration across sampling frames.
3. There was a general feeling that more needs to be learned about the ways of obtaining access to administrative data sets and to respondents who are not directly accessible for interview. Gatekeeping is a continuing problem, particularly when using list or other nonhousehold sampling strategies.

Mode Effects

The final general theme of the conference concerns mode effects. As with the sessions on sampling and measurement, the discussion of mode effects also focused on the interaction between these effects and respondent characteristics. In the session on question design, three papers discuss respondent characteristics and patterns of response to certain kinds of questions. In the sampling session, much of the need for special sampling frames was dictated by the characteristics of the respondents who were to be interviewed. Similarly, the discussion of mode effects focused on special populations and the need for confidentiality during data collection. Almost entirely, this discussion focused on automated strategies for data collection and particularly on audio computer-assisted self-interviewing (ACASI).

1. A major point that emerged from the discussion of this topic is how poorly our knowledge of these techniques is integrated into our general theories of survey research. In particular, there was a recommendation that these techniques ought to be evaluated using the questionnaire design methodology discussed in the Session 2 summary.
2. It is also clear that we need to know more about how the successful use of these techniques is related to respondent characteristics, especially age and ethnicity. The interpretation of some of the data that were presented is confusing because these effects were not clearly sorted out in the evaluation designs or data analysis.
3. Small sample sizes, pilot research, and uncontrolled evaluations make interpretation of the results of this work difficult, and there was a general consensus that more funding needs to be allocated by agencies that are supporting the development of these techniques to better evaluate their efficacy and their contribution to total survey error. As noted above, the overall theory on response effects also needs to take into account the effects of computerized interviewing using various modalities on survey error.
4. These research studies of innovative computerized data collection techniques need to take into account both variable and systematic error in comparing them with more traditional modes of data collection.

References

Dillman, D. (1978). Mail and telephone surveys: The total design method. New York: Wiley.

Groves, R. M. (1989). Survey errors and costs. New York: Wiley.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Conference Participants

Lu Ann Aday, Ph.D.
Professor
School of Public Health
University of Texas
P.O. Box 20186
Houston, TX 77225
713-792-4471
laday@utsph.sph.uth.tmc.edu

Ronald Andersen, Ph.D.
Professor
Department of Health Services
UCLA School of Public Health
10833 LeConte Ave.
Los Angeles, CA 90095
310-206-1810

Barbara Bailar, Ph.D.
Vice President for Survey Research
NORC
1155 E. 60th St.
Chicago, IL 60637
312-753-7550
bailar-b@norcmail.uchicago.edu

Robert F. Belli
Assistant Research Scientist
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106-1248
313-763-6020
bbelli@umich.edu

Sandra H. Berry
Director
Survey Research Group
RAND
1700 Main St.
Santa Monica, CA 90407-2138
310-393-0411 ext. 7779
sandra_berry@rand.org

Paul Biemer
Chief Scientist
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
919-541-6056
ppb@rti.org

Johnny Blair
Associate Director
Survey Research Center
Art-Sociology Bldg., Rm. 1103
Campus Drive
University of Maryland
College Park, MD 20742
301-314-7831
jb155@umail.umd.edu

Norman M. Bradburn, Ph.D.
Senior Vice President for Research
NORC
1155 E. 60th St.
Chicago, IL 60637
312-702-1066
bradburn@norcmail.uchicago.edu

Robert M. Bray, Ph.D.
Senior Research Psychologist
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709-2194
919-541-6433
rmb@rti.org

Mary C. Burich
Project Director
Senior Associate
Abt Associates, Inc.
101 N. Wacker Dr., Ste. 400
Chicago, IL 60606-7301
312-621-2663
mary_cay_burich@abtassoc.com

Catharine W. Burt, Ed.D.
National Center for Health Statistics
6525 Belcrest Rd., Rm. 952
Hyattsville, MD 20782
301-436-7132 *8175#
cwb2@nch09a.em.cdc.gov

Richard Campbell, Ph.D.
Professor
Department of Sociology
University of Illinois at Chicago
1007 W. Harrison St., M/C 312
Chicago, IL 60607
312-413-3759
dcamp@uic.edu

Joel C. Cantor, Sc.D.
Robert Wood Johnson Foundation
Subsequently at
United Hospital Fund
Empire State Building
350th Ave., 23rd Floor
New York, NY 10118
212-494-0751

James Chromy, Ph.D.
Chief Scientist
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
919-541-6228
jrc@rti.org

Steven Cohen, Ph.D.
Director, Division of Statistics
Agency for Health Care Policy
and Research
Executive Office Building
2101 E. Jefferson, #500
Rockville, MD 20852
301-594-1406

Timothy Cuerdon, Ph.D.
Statistician
Health Standards and Quality
Health Care Financing Administration
7500 Security Blvd.
Mail Stop S1-15-18
Baltimore, MD 21244
410-786-9465
tcuerdon@hcfa.gov

Marcie Cynamon
Division of Health Interview Statistics
National Center for Health Statistics
6525 Belcrest Rd., Rm. 850
Hyattsville, MD 20782
301-436-7085 ext. 118
mlc6@nch08a.em.cdc.gov

Ronald Czaja, Ph.D.
Associate Professor
North Carolina State University
Department of Sociology and Anthropology
Box 8107
Raleigh, NC 27695-8107
919-515-3291
ronald_czaja@ncsu.edu

Don Dillman, Ph.D.
Director
Social and Economic Research Center
Washington State University
133 Wilson Hall
Pullman, WA 99164-4014
509-335-1511
dillman@wsuvm1.csc.wsu.edu

Allen P. Duffer, B.A.
Senior Survey Director/Department Manager
Survey Research Division
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
919-541-7356
apd@rti.org

Patricia Ebener
Behavioral Scientist
RAND
P.O. Box 2138
1700 Main St.
Santa Monica, CA 90407-2138
310-393-0411 ext. 7905
patricia_ebener@rand.org

Brad Edwards
Vice President
Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3128
301-294-2021
edwardb1%westat@mcimail.com

Franklin J. Eppig Jr.
Health Care Financing Administration
N3-02-02
7500 Security Blvd.
Baltimore, MD 21244-1850
410-786-7950

Susan A. Flocke, M.A.
Research Analyst
Department of Family Medicine
Case Western Reserve University
UCRC2, Rm. 306
11001 Cedar
Cleveland, OH 44106
216-368-3887
saf6@po.cwru.edu

Barbara H. Forsyth, Ph.D.
Senior Measurement Methodologist
Survey Research Division
Research Triangle Institute
6101 Executive Blvd.
Rockville, MD 20852
301-230-4696
bhf@rti.org

Floyd J. Fowler Jr., Ph.D.
Senior Research Fellow
Center for Survey Research
University of Massachusetts–Boston
100 Morrissey Blvd.
Healey Library, 10th Floor
Boston, MA 02125-3393
617-287-7200

Joseph Gfroerer
Office of Applied Studies
Substance Abuse and Mental
Health Services Administration
Rm. 16C-06, 5600 Fishers Lane
Rockville, MD 20857
301-443-7977
jgfroere@aoa2.ssw.dhhs.gov

Robert Groves, Ph.D.
Professor
Joint Program on Survey Methodology
Can be reached at
Survey Research Center
University of Michigan
426 Thomson St.
Ann Arbor, MI 48109
313-763-2359
bgroves@survey.umd.edu

John W. Hall
Senior Sampling Statistician
Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543
609-275-2357
jhh@mprnj.com

Lisa E. Harris, M.D.
Clinical Assistant Professor
Indiana University School of Medicine
Regenstrief Institute for Health Care
Fifth Floor, RHC
1001 W. Tenth St.
Indianapolis, IN 46202
317-630-6312
lharris@vax1.iupui.edu

Betty Havens, D. Litt.
Professor and Research Fellow
Department Community Health Sciences
Faculty of Medicine
University of Manitoba
S110B-750 Bannatyne Ave.
Winnipeg, Manitoba R3E 0W3 Canada
204-789-3427
havens@umanitoba.ca

Tabitha P. Hendershot
Survey Manager
Research Triangle Institute
1615 M St., N.W., Ste. 740
Washington, DC 20036
202-728-2075
tph@rti.org

Daniel H. Hill, Ph.D.
Associate Research Scientist
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106-1248
313-763-6866
dhhill@umich.edu

Michael Hilton
National Institute on Alcohol
Abuse and Alcoholism
6000 Executive Blvd., Ste. 505
Bethesda, MD 20892-7003
301-443-8753
mhilton@willco.niaaa.nih.gov

John Horm
Division of Health Interview Statistics
National Center for Health Statistics
6525 Belcrest Rd., Rm. 850
Hyattsville, MD 20782
301-436-7085
jdh3@nch08a.em.cdc.gov

Timothy P. Johnson, Ph.D.
Associate Director
Survey Research Laboratory
University of Illinois at Chicago
910 W. Van Buren, Ste. 500
M/C 336
Chicago, IL 60607
312-996-5310
u09146@uicvm.uic.edu

Graham Kalton, Ph.D.
Senior Statistician and
Senior Vice President
Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3129
301-251-8253
kaltong1@westat.com

Larry Kessler
National Cancer Institute
Subsequently Director of the Office of
Surveillance and Biometrics at
Food and Drug Administration
1350 Piccard Dr.
HFZ-500
Rockville, MD 20850
301-594-2812
lgk@fdadr.cdrh.fda.gov

Bärbel Knäuper, Dr. Phil.
Visiting Scholar
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106-1248
313-936-0550
bknauper@umich.edu

Carol Kosary
Surveillance Program
National Cancer Institute
6130 Executive Blvd. MSC 7352
Executive Plaza North, Rm. 343J
Rockville, MD 20892-7352
301-496-8510
kosaryc@dcpcepn.nci.nih.gov

Mary Grace Kovar
Senior Health Scientist
NORC
1350 Connecticut Ave, NW,
Ste. 500
Washington, DC 20036
202-223-6040
kovar@norcmail.uchicago.edu

Richard A. Kulka, Ph.D.
Research Vice President
Statistics, Health, and Social Policy Unit
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
919-541-7008
rak@rti.org

James M. Lepkowski, Ph.D.
Senior Study Director
Institute for Social Research
University of Michigan
426 Thompson St.
Ann Arbor, MI 48105
313-936-0021
jimlep@umich.edu

Karin A. Mack, Ph.D.
Assistant Professor
Mississippi State University
Department of Sociology, Anthropology,
and Social Work
P.O. Drawer C
Mississippi State, MS 39762
601-325-7874
mack@soc.msstate.edu

Katherine Marconi
Bureau of Health Resources Development
Health Resources Services Administration
5600 Fisher Lane, Rm. 7A07
Rockville, MD 20857
301-443-6560
marc102w@wonder.em.cdc.gov

Nancy Mathiowetz, Ph.D.
Agency for Health Care Policy
and Research
Subsequently Professor at
Joint Program in Survey Methodology
1218 LeFrak Hall
College Park, MD 20742
301-405-0933
nmathiow@survey.umd.edu

Catherine A. Melfi, Ph.D.
Adjunct Scientist
School of Medicine
Indiana University
melfi@ucs.indiana.edu

Also, Research Scientist at
Lilly Research Laboratories
Lilly Corporate Center—Drop Code 1850
Indianapolis, IN 46285

Lorraine Midanik, Ph.D.
Associate Professor
School of Social Welfare
University of California, Berkeley
120 Haviland Hall
Berkeley, CA 94720
510-642-7974
lmidanik@violet.berkeley.edu

Jeff Moore
Center for Survey Methods Research
U.S. Bureau of the Census
Washington, DC 20233-9150
301-457-4719
jeffrey_c_moore@ccmail.census.gov

Steve Niemcryk, Ph.D.
Chief, Resources Analysis Branch
Office of Science and Epidemiology
Bureau of Health Resources Development
Health Resources Services Administration
5600 Fishers Lane, Rm. 7a-08
Bethesda, MD 20867
301-443-6560
sniemcr@hrsa.ssw.dhhs.gov

Mary Utne O'Brien, Ph.D.
Associate Professor
School of Public Health
Epidemiology and Biostatistics Division
University of Illinois at Chicago
2121 W. Taylor St.
539 SPHW, M/C 922
Chicago, IL 60612
312-996-7978
u30273@uicvm.cc.uic.edu

Mary Beth Ofstedal
National Center for Health Statistics
6525 Belcrest Rd., Rm. 730
Hyattsville, MD 20782
301-436-5979, 145
mbo0@nch07a.em.cdc.gov

Diane O'Rourke
Coordinator of Research Programs
Survey Research Laboratory
University of Illinois
909 W. Oregon, Ste. 300
Urbana, IL 61801
217-333-7170
dorourke@srl.uic.edu

Jennifer Parsons, M.A.
Project Coordinator
Survey Research Laboratory
University of Illinois at Chicago
910 W. Van Buren, Ste. 500
M/C 336
Chicago, IL 60607
312-413-0492
jparsons@uic.edu

Willard Rodgers, Ph.D.
Research Scientist
Survey Research Center
University of Michigan
1306 ISR
P.O. Box 1248
Ann Arbor, MI 48106-1248
313-763-6623
wrodgers@umich.edu

Norbert Schwarz, Dr. Phil.
Professor
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106-1248
313-747-3616
nschwarz@umich.edu

Jane D. Shepherd, Ph.D.
Associate Director
Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3129
301-294-3967
shephej1%westat@mcimail.com

Kurt Stange, M.D., Ph.D.
Assistant Professor of Family Medicine
Epidemiology and Biostatistics and Sociology
Department of Family Medicine
Case Western Reserve University
UCRC2, Rm. 306
11001 Cedar
Cleveland, OH 44106-7124
216-368-6297

Seymour Sudman, Ph.D.
Deputy Director
Survey Research Laboratory
University of Illinois
909 W. Oregon, Ste. 300
Urbana, IL 6180l
217-333-4276
seymour@vmd.cso.uiuc.edu

Roger Tourangeau, Ph.D.
Senior Scientist
NORC
1350 Connecticut Ave., N.W., Ste. 500
Washington, DC 20036
202-223-6327
tourange@norcmail.uchicago.edu

Charles F. Turner, Ph.D.
Director
Program in Health and Behavior
Measurement
Research Triangle Institute
6101 Executive Blvd.
Rockville, MD 20852
301-230-4640
cft@rti.org

Lois M. Verbrugge, Ph.D.
Distinguished Research Scientist
Institute of Gerontology
300 North Ingalls
University of Michigan
Ann Arbor, MI 48109-2007
313-936-2103
verbrugg@umich.edu

Daniel C. Walden, Ph.D.
Agency for Health Care Policy
and Research
2101 E. Jefferson, Ste. 500
Rockville, MD 20852
301-594-1400
dwalden@cghsir.ahcpr.gov

Elinor Walker
Agency for Health Care Policy
and Research
2101 E. Jefferson St., Ste. 502
Rockville, MD 20852-4908
301-594-1352 ext. 108
ewalker@po3.ahcpr.gov

Richard B. Warnecke, Ph.D.
Director
Survey Research Laboratory
University of Illinois at Chicago
910 W. Van Buren, Ste. 500
M/C 336
Chicago, IL 60607
312-996-6130
dickw@srl.uic.edu