CONFERENCE PROCEEDINGS

HEALTH SURVEY RESEARCH METHODS

Edited by
Floyd J. Fowler, Jr., Ph.D.

September 1989

NATIONAL CENTER FOR HEALTH SERVICES RESEARCH
AND HEALTH CARE TECHNOLOGY ASSESSMENT
Public Health Service
U.S. Department of Health and Human Services

DHHS Publication No. (PHS) 89–3447

Foreword

The Fifth Conference on Health Survey Research Methods represents another advance on the continuum of methodological improvements in health survey research stimulated through the preceding four conferences. This series of conferences has two underlying goals. The long-range objective is to improve the quality of health survey data that are collected and to enhance their value and use by decisionmakers responsible for shaping health practices, policies, and programs. The immediate goal is to provide a forum in which methodologists can discuss the state-of-the-art in data collection, identify major problems and needs, and define issues, hypotheses, and priorities to guide future research.

The conferences to date have been successful in assisting researchers who are not survey methodologists with the limitations and problems inherent in survey research. The need to acquaint the research community with these issues is especially critical today, given the necessities and uncertainties of obtaining data on persons infected with the human immunodeficiency virus (HIV). Providing a vehicle for addressing these issues and improving our data gathering tools at this time is of heightened importance.

The National Center for Health Services Research and Health Care Technology Assessment (NCHSR) and the National Center for Health Statistics (NCHS) are pleased to have cosponsored these five conferences. We appreciate the assistance of the Milbank Memorial Fund and the Commonwealth Fund for making this fifth conference possible.

Using the past as prologue, we have a firm foundation on which to build. We anticipate that the proceedings of this fifth conference will make a significant contribution to the field.

J. Michael Fitzmaurice, Ph.D.
Director, NCHSR

Manning Feinleib, M.D., Dr. P.H.
Director, NCHS

Preface

Background

In 1975, a group of researchers representing academic institutions and government research agencies met informally to discuss what was known about health survey methods. The group concluded that a conference specifically devoted to this topic was needed for several reasons:

- Researchers whose work bore on health survey methods were widely dispersed, not only geographically, but also in their work settings and disciplines. Sociologists, economists, psychologists, and political scientists were involved in methodological studies of direct relevance to traditional health services researchers and statisticians. However, because of the traditional alignment of professional activity and journal publication by discipline, knowledge from these diverse sources often was not shared and readily accessible.
- The problem was exacerbated by the fact that regular professional meetings do not provide opportunities for discussing methodological issues. On the one hand, concerns about survey methodology are only a small part of the spectrum of any individual traditional academic discipline. Developing a methodology agenda at professional meetings is difficult because such meetings tend to be very large and the concentrations of methodologists too low for effective discussion and serious interchange of ideas.
- Methodological findings often are not reported in traditional channels. Many important methodological findings are incidental to the main thrust of research studies and hence are relegated to appendixes. Some journals are inconsistent in reporting methodological results. In particular, negative findings and studies of methodologies that either have problems or do not work seldom find their way into the professional literature, though they may have important implications for those engaged in health services research.

Responding to such concerns, the First Conference on Health Survey Research Methods was held at Airlie House in Virginia in 1975. Sponsored by the National Center for Health Services Research and the National Center for Health Statistics, a group of about 40 researchers active in methodological studies were brought together to discuss what was known about sources of error in surveys and strategies for reducing error. The proceedings of those discussions were published as a resource for health services researchers.

Those who participated in the Airlie House conference and readers of the resulting report thought the value of such a meeting was evident. Three similar conferences followed, each developed by a planning committee of academic and government researchers. The meetings differed in the extent to which they involved formal papers and in the number of participants. However, in each case, the result was an important contribution to health survey methodology.

The last such conference was held in 1982. Yet the problems of communication that led the original planning group to initiate the first conference are as real today as they were in 1975. Moreover, survey research has changed substantively in some rather important ways since 1982.

Perhaps the most marked development in survey research in the past 5 years is the increased demand for data that are hard to collect. Clearly, no health issue since polio has had the public health attention and importance of acquired immunodeficiency syndrome (AIDS). Concern about AIDS has led health survey investigators to develop and ask questions previously unimagined by researchers. Moreover, getting good estimates not only of infection but also of risks—a process which involves very sensitive questioning—is of great importance in projecting the impact of AIDS and in forming policy with respect to AIDS. Health researchers are increasingly asked to make estimates of events and problems that are extraordinarily difficult to estimate reliably and validly, not only with respect to AIDS but

also to other public health problems like alcohol abuse, teenage pregnancy, drug use and abuse, homelessness, and sexual abuse. The need for data about subsets of the population affected by these problems is great. Identifying and reaching such groups poses great challenges to standard methodologies.

Finally, there has been a good deal of new research in the past 7 years on topics such as the design and evaluation of survey questions, interviewing techniques, and collecting data by telephone. The primary goal of the fifth conference, therefore, was to present what was known and, equally important, what was unknown about sources of survey error and how to minimize it in a single forum accessible to health researchers.

Developing the Conference

Responsibility for putting together the format and content of the conference lay with a six-person planning committee. This committee decided that only participants knowledgeable on some aspect of survey methodology would attend the conference and participate in the program. The purpose of this guideline was to maximize the opportunity for and quality of the group discussion. Also, there would be five basic sessions, each devoted to a single integrated topic. Each session would consist of five presentations of data; one or two discussants who would place the presentations in a broader perspective; and a focused floor discussion in which the issues and problems presented would be amended, amplified, and refined to produce a summary of current knowledge about the topic.

The planning committee outlined a set of broad methodological areas, including question design and evaluation, data collection procedures, sampling, nonresponse, and interviewing techniques. Special note was made of the importance of methodological knowledge about surveys pertaining to sensitive behaviors, such as those relevant to risk of AIDS. Over 150 researchers responded with a proposal to participate in the conference. From these, presenters were selected for the five sessions based on the extent to which their proposed presentations would potentially make a distinctive methodological contribution not currently well covered in the existing literature. In addition, an important effort was made to produce integrated sessions that blended together to address a general set of methodological issues. The conference was held May 2-4, 1989, in Keystone, Colorado. Fifty-nine participants contributed to the program as presenters, discussants, chairs, recorders, or speakers at one of the special sessions.

About This Volume

The prepared presentations for each session are here designated as *feature papers*. They are presented by session topic, followed by the accompanying *discussion papers*. The session discussants addressed the generalizations that emerged from the feature papers and how they related to current methodological knowledge.

Each session was followed by an open floor discussion, in which the comments of presenters and discussants were themselves the focus of discussion. The comments were recorded by the session chair and recorder who wrote a summary of the main points of the floor discussion as well as of the main methodological issues, generalizations, conclusions, and future research considerations that emerged. These papers are designated as *session summaries* in this volume.

In addition to the five sessions outlined, the conference featured speakers at four special sessions who discussed issues related to the themes of the conference. These papers, called *conference addresses* in this volume, appear in the order in which they were presented.

The *concluding discussion paper* examines issues relevant to the study of AIDS from a total survey design perspective. Finally, major methodological problems that emerged from the conference as a whole as issues in need of further research are summarized in the *conference summary*.

List of Acronyms

ADL	Activities of Daily Living
AIDS	Acquired Immunodeficiency Syndrome
AIMS	Arthritis Impact Measurement Scale
APA	American Psychological Association
ARC	AIDS-related Complex
ASSIST	Automated System for Survey Information and Statistical Tools
BMI	Body-Mass Index
BRR	Balanced Repeated Replication
CAPI	Computer-assisted Personal Interviewing
CATI	Computer-assisted Telephone Interviewing
CDC	Centers for Disease Control
CBO	Community-based Organizations
COPD	Coronary Obstructive Pulmonary Disease
CUSP	Correction for Unequal Selection Probabilities
DHHS	Department of Health and Human Services
DIS	Diagnostic Interview Schedule
D-PTSD	Modified DIS-type PTSD Module
DRG	Diagnosis Related Groups
DSM	Diagnostic and Statistical Manual
DSM III	Diagnostic and Statistical Manual, 3rd ed.
ECA	Epidemiologic Catchment Area
FN	False Negative
FP	False Positive
GAO	Government Accounting Office
HCFA	Health Care Financing Administration
HCR	Handwritten Character Recognition
HDS	Hospital Discharge Survey
HER	Health Economics Research, Inc.
HHS	Household Survey
HIC	Health Insurance Claim
HIV	Human Immunodeficiency Virus
HMO	Health Maintenance Organization
HRH	High Risk Heterosexual
HTLV-1	Human t-cell Leukemia Virus-1
IES	Impact of Event Scale
IPA	Independent Practice Association
IPC-CR	Institutional Population Component—Current Residents
IPC-NA	Institutional Population Component—New Admissions
IRB	Institutional Review Board
IRPCC	Intermountain Regional Poison Control Center
IV	Intravenous
IVDU	Intravenous Drug Use
KABB	Knowledge, Attitudes, Beliefs, Behavior
LOS	Length of Stay
LSOA	Longitudinal Study of Aging
MACS	Multi-center AIDS Cohort Study
MADRS	Medicare Automated Data Retrieval System
MBRFS	Michigan Behavior Risk Factor Survey
MCA	Multiple Classification Analysis
MIS	Management Information System
MKA	Most Knowledgeable Adult
MMPI	Minnesota Multiphase Personality Index
MOB	Mobility
M-PTSD	Mississippi Scale for Combat-related PTSD
MRC	Medical Record Component
MSA	Metropolitan Statistical Area
MSQ	Mental Status Questionnaire
MSU	Michigan State University
NCHS	National Center for Health Statistics
NCHSR	National Center for Health Services Research and Health Care Technology Assessment
NDI	National Death Index
NHANES	National Health and Nutrition Examination Survey
NHCS	National Health Care Survey
NHIS	National Health Interview Survey
NHSB	National Survey of Health and Sexual Behavior
NHSS	National Household Seroprevalence Survey
NLS/Y	National Longitudinal Survey of Labor Market Experiences—Youth Cohort
NMES	National Medical Expenditure Survey
NORC	National Opinion Research Center
NSFG	National Survey of Family Growth
NVVRA	National Vietnam Veterans Readjustment Study
OA	Osteoarthritis
OMB	Office of Management and Budget
PAC	Physical Activity
PHS	Public Health Service
PSU	Primary Sampling Unit
PTSD	Post-Traumatic Stress Disorder
PWA	Person with AIDS
QALY	Quality Adjusted Life Years
QDL	Questionnaire Design Laboratory
RDD	Random Digit Dialing
RDU	Recreational Drug Use
RFA	Request for Application
RFP	Request for Proposal
ROC	Receiver Operator Characteristic
RSOP	Richmond Street Outreach Project
RTI	Research Triangle Institute
RUHBC	Research Unit on Health and Behavioral Change
SAC	Social Activity
SAQ	Self-administered Questionnaire
SCID	Structured Clinical Interview for DSM-III-R
SEER	Surveillance, Epidemiology, and End Result
SIP	Sickness Impact Profile
SOA	Supplement on Aging
SRC	Survey Research Center
SRL	Survey Research Laboratory
QDRL	Questionnaire Design Research Laboratory
QWB	Quality of Well-being
TN	True Negative
TP	True Positive
VA	Veterans Administration
YPLLS	Years of Potential Life Lost

Contents

SESSION 1:
STRATEGIES FOR EVALUATING QUESTIONS

SESSION 2:
VALIDITY OF REPORTING IN SURVEYS

SESSION 3:
COLLECTING DATA FROM SAMPLES OF OLDER ADULTS AND NURSING HOME POPULATIONS

Strategies for Evaluating Questions

Introduction by Floyd J. Fowler, Jr.

How to evaluate survey questions is one of the least developed methodological areas in survey research. Almost four decades ago, Stanley Payne (1951) wrote a book enunciating some common sense standards for survey questions. His advice included such appropriate principles as making questions clear, making them short, and asking one question at a time. Unfortunately, the title of his book, *The Art of Asking Questions*, also implied that an appropriate standard for questions was the taste or judgment of the individual researcher.

Even today, too often no standards are applied to survey questions except whether or not the researcher believes a question is good. Questions are used because they were used before and they worked, with no specification of what "worked" means.

One of the important current developments in survey methodology is research to make the evaluation of questions more systematic, objective, and useful. One focus of such research is on the steps that a researcher should take to evaluate questions before using them. Standard pretests, in which experienced interviewers conduct a few practice interviews with typical respondents and provide feedback to researchers, are inadequate. Researchers are experimenting with better ways to identify questions that respondents do not consistently understand, that interviewers cannot ask as worded, and that respondents cannot or do not answer in ways that meet the study objectives.

Researchers have also been examining possible ways to analyze survey questions after they have been used to determine whether they fulfilled their purpose. Researchers are not interested in answers to questions per se; rather, they are interested in answers because they correspond with some reality the researcher wants to measure. If questions are doing what they are supposed to be doing, a predictable set of patterns should emerge from the data. Questions purported to measure the same thing should produce the same results and be highly correlated. In some cases, different groups of the population should produce predictably different answers. Needed is a set of guidelines for analyzing questions to evaluate whether they are good measures.

This session's five feature papers are devoted to the general problem of how to evaluate questions and the criteria that should be used to identify good questions. Two papers deal with strategies for testing questions before they are included in a survey; the other three provide examples of how survey results can be analyzed to evaluate individual questions or a series of questions. Together, the five presentations provide a good basis for one vitally important area of survey research methodology.

Using Intensive Interviews to Evaluate Questions

Patricia N. Royston

Introduction

For the past few years, the National Center for Health Statistics (NCHS) has been experimenting with the concept that the methods and theories of cognitive psychology could be used in intensive interviews with respondents to help identify, understand, and correct sources of response errors in surveys. This concept was first discussed in 1983 at the Advanced Research Seminar on the Cognitive Aspects of Survey Methodology, conducted by the Committee on National Statistics under a grant from the National Science Foundation (Jabine & associates, 1984). In 1985, NCHS established the Questionnaire Design Research Laboratory (QDRL), which was the first permanent laboratory to use "cognitive-style" intensive interviews to test draft questionnaires. During the past 4 years, the Questionnaire Design Research Laboratory has progressed from being perceived by most NCHS data systems as a "nicety" to being perceived as a necessity. This paper presents a critique of this approach to pretesting questionnaires, based on the QDRL experience.

What Is an Intensive Interview?

The intensive interview is a procedure for determining whether questions are working as intended and, if not, for obtaining clues as to why the questions fail. The interviewer employs a variety of techniques, depending on the type of questionnaire (for example, personal interview versus self-administered, knowledge and attitudes versus autobiographical) and the type of problems expected (for example, recall, comprehension, sensitivity). These techniques have been described in detail in an earlier paper (Royston & associates, 1986), and are summarized briefly in Figure 1. The techniques are designed to provide information about the respondent's ability to interpret the question as intended and to produce an accurate answer. Interviews are conducted with paid volunteers, usually in a laboratory or other controlled setting, by specially trained interviewers. Usually the interviewer administers a draft questionnaire, but supplements it with probes and other cognitive techniques (for example, paraphrasing, confidence ratings) as needed to identify the sources of response errors. For example, with the concurrent protocol procedure, more commonly known as the "Think-aloud," respondents are asked to think aloud as they try to answer the questions. The interviewer watches for verbal and nonverbal clues that the respondent misunderstood some aspect of the question or is having difficulty answering. Because few respondents can report on their thought processes as they occur, the interviewer must compensate by probing extensively for problems. Probes are formulated by the interviewers as needed to elicit information about the respondent's interpretation of the question and formulation of the response.

Does the Interviewer Need Special Training or Qualifications?

For the intensive interview to be most effective, the interviewer must:

1. be familiar with the questionnaire objectives;
2. be knowledgeable about the wide variety of questionnaire flaws that can occur, to know when to probe;
3. be able to recognize both verbal and nonverbal clues provided by the respondent that there are problems, and to follow up with effective probes;
4. be able to probe effectively even if there are no overt problems;

Patricia Royston is with the Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control, Hyattsville, Maryland.

Figure 1. Laboratory methods

1. Focus interviews
   Unstructured discussion of a survey topic or draft questionnaire with small groups of volunteers
2. Concurrent "think-aloud" interviews
   Respondent thinks aloud when answering questions—answers are probed extensively
3. Retrospective "think-aloud" interviews
   Respondent answers all questions first, then is asked how he/she arrived at the answers
4. Paraphrasing
   Respondent repeats question in his/her own words
5. Confidence ratings
   Respondent rates degree of confidence in the accuracy of the answer
6. Response latency measurements
   The time between the question being asked and the respondent answering is noted
7. Free and dimensional sorts
   Respondent sorts lists of similar items into groups that go together or ranks the items according to specified scales

5. be able to spot mechanical problems (skips, order, and so on) while watching for respondent clues and probing.

As this list indicates, "intensive" is the appropriate word for these interviews, especially from the viewpoint of the interviewer. This is not a production operation, conducted automatically, in which the respondents spontaneously provide insights into their interpretation of questions and their strategies for recalling information. A great deal can be gained from each interview, but only if the interviewer has the qualifications listed above.

Why Conduct Intensive Interviews?

There are four main reasons to conduct intensive interviews:

1. *To determine whether questions satisfy the survey objectives*: At NCHS the interviewers are the questionnaire designers, so they have a frame of reference that permits them to evaluate whether objectives are being met, in terms of both data quality and what is being measured. They are often able to identify questions that have no flaw other than that they will not provide the data needed by the sponsor. It is infeasible to provide a field staff with the detailed knowledge of the study goals to permit identification of these problems in a field pretest.

The results of the intensive interview can also be used to help subject matter experts clarify their objectives. Often the subject matter experts have only a general idea of their objectives in the early stages of questionnaire testing. During the intensive interview phase, several meetings are usually held in which survey designers and lab staff try to elicit more definitive statements of question objectives from the subject matter experts by pointing out questions that respondents found to be vague or ambiguous. These meetings often result in more focused survey objectives.

2. *To identify the kinds of problems usually found in pretests, but identify them before the pretest*: Pretests effectively identify *overt* question flaws, that is, flaws that interfere with the flow of the interview for a variety of reasons. For example, there are instances in which the respondent cannot understand the question, cannot or will not answer the question, or is asked an obviously irrelevant question. The intensive interview can identify these same problems, often with a handful of interviews. There are many instances in which conducting only two or three lab interviews revealed a surprising number of problems with redundancy, awkward wordings, missing skips, excessive wordiness, and similar problems that were missed in a preliminary review of the questionnaire. Sometimes problems detected in the initial review are not revised until confirmed in the interview, because there is lack of consensus among the reviewers about the seriousness of the problem, and/or correcting the problem would lengthen or otherwise complicate the questionnaire. After a series of intensive interviews, questionnaires can be revised so that a field pretest can concentrate on the kinds of problems that cannot be identified in the lab.

Improving the questionnaire before the field pretest is especially useful for NCHS surveys, which operate under severe time constraints. In the National Health Interview Survey, for example, new supplements are developed and tested each year for implementation in the following calendar year. Because of Office of Management and Budget clearance requirements and lead time required by the Census Bureau for preparation of interviewer materials and for printing, there is very little time to make changes in the questionnaires after the pretest.

3. *To identify hidden sources of response error often missed in field pretests*: For years, problems that did not interfere with the flow of the interview were ignored. Little attention was given to the questions with hidden flaws which resulted in answers that were not measures of what was intended or which contained substantial response error. Research consisted of validation studies or split ballot studies, which often revealed problems but provided little insight into the reasons for these problems.

Intensive interviews identify many of the hidden problems that often affect data quality by providing information on how the respondents interpret the questions, how they recall and/or estimate the answer, and other factors that affect the final response. These sources of error are likely to be missed in pretests, as long as the pretest respondents provide a response with little or no hesitation.

These methods also provide information about the root of question flaws, so that effective solutions can be

found. Knowing the reason that a question is not working can make the solution obvious, or at least can provide clues as to what might solve the problem. For example, knowing that a word or phrase is being interpreted in two or more ways usually makes it easy to write a clearer, unambiguous question that meets the objectives. Examples of problems often missed in pretests include:

a. Questions that use vague or ambiguous words or phrases that are interpreted differently by different respondents:
"Example: How long have you used the (device)?"
This was included in a series of questions on the use of assistive devices, such as wheelchairs, canes, hearing and vision aids. Respondents who used these kinds of devices were recruited to test the questions. Many of them reported that they had used the devices intermittently, often with breaks of months or years. Some would try to add up the periods of time when the device was in use, while others would report the number of years since first use.

b. Questions with unfamiliar terms that result in the "If I don't know what it is I do not/did not have it" response strategy:
"Example: (From National Health Interview Survey checklists)
During the past 12 months, did anyone have:
Enteritis? Diverticulitis? Colitis? Gastritis? Fatty liver?"
Usually the strategy, "If I don't know what it is, I don't have it" is a reasonable heuristic; however, sometimes it results in response error when technical terms are used in place of a familiar lay term, for example, otitis media instead of ear infection. The intensive interview can help to identify those instances where that response strategy is not appropriate and results in response error.

c. Questions that ask for information which respondents do not usually have:
"Example: Why haven't you tested your house for radon?"
Most respondents simply had not given it any thought, and gave a variety of socially acceptable excuses for not testing. "Never really thought about it" was rarely mentioned, but was clearly the case for many people.

d. Questions that are out of line with current practice and the real world, so that, if interpreted literally, errors occur:
"Example: Please tell me if this house is equipped with any of the following features:
Widened doors or passages? A lowered counter? Any other changes? (specify)."
These questions were part of the supplement on assistive devices mentioned earlier. The objective of the question was to elicit reports of household features designed for the disabled. As worded, however, the questions implied that only house modifications were to be reported, when in fact many people have whole additions and even houses built to accommodate their disabilities. The root question and the "any other changes" also failed to mention that only modifications for disabled persons were to be reported.

e. Knowledge questions that result in excessive guessing:
"Example: Which of the following are early signs of mouth and throat cancer?
Sore in the mouth that does not heal;
Persistent sore throat;
Clicking or popping of the jaw."
Respondents are unwilling to admit that they do not know the answers, and use general knowledge to help them to guess the correct ones. This is easily detected in intensive interviews, because respondents are very committed to helping find flaws in the questionnaire and readily admit that they are guessing.

4. *To identify problems that would be missed in pretests because the pretest sample does not happen to contain any/enough of a specific population:* The laboratory interview provides the opportunity to recruit special target populations to test specific sets of questions. In the past 3 years, we have recruited persons with diabetes, medical device implants, digestive disease, foot problems, and disabilities that required mechanical aids, and relatives of persons with chronic mental illness. This provided us with invaluable experience with condition-specific questions that were often virtually untested in the field pretest because so few households had the target characteristic.

Does the Intensive Interview Make the Field Pretest Unnecessary?

The intensive interview does not identify all of the problems with questionnaires. The intensive interview often misses:

1. *Problems arising from respondent fatigue, distractions, hostility to the survey, or lack of motivation:* In the lab, the subjects are paid volunteers, so they are more interested than household respondents, and they are attentive, relaxed, and undisturbed by the usual household distractions. As a result, they are more often able to follow a complex line of questioning and to carry over information from one question to another. The following is an example of a question series that worked in the lab but failed in the field:

> "a. About how much did you weigh when you were 25 years old?
> b. What is the most you have ever weighed (except when you were pregnant)?
> c. How old were you when you first weighed that much?"

Respondents in the pretest were uncertain what "that much" in question c referred to; they were unable to see the connection to the answer they had just provided to question b. This had been a problem with only one respondent in the lab, so that we ignored it when making revisions, thinking it was just a fluke.

2. *Problems with questions targeted toward a segment of the population not represented by the lab volunteers:*

A small number of questions are not tested in the lab for some reason; either because (1) the questions are targeted toward the type of person that rarely volunteers for our interviews (for example, people who are not interested in or knowledgeable about health), or (2) they are targeted toward a group rare enough that none happened to volunteer at the right moment, and a targeted recruitment was not conducted because only a very few questions applied exclusively to them. For example, the following questions were part of a lengthy questionnaire tested in the Questionnaire Design Research Laboratory:

> "Have you made any LASTING and MAJOR changes in what you eat and drink for health reasons? "(If no) People have many reasons for not changing what they eat and drink. I am going to read some of those reasons. For each one, please tell me if it is a major reason why you have not made changes, just part of the reason, or not a reason. "It seems that everything you eat is bad for you, so why bother changing. "I enjoy the things I eat and drink and don't want to change. "The things I eat and drink are healthy so there is no reason for me to make changes." (. . .)

Our lab respondents are usually very interested in disease prevention and in health in general, so all lab respondents reported making changes in diet. As a result, the follow-up question on reasons for not changing the diet was never tested in the intensive interviews. We did not learn until the field pretest that the list of reasons provided to the respondents did not work well. We had been warned by a cognitive psychologist consulting on an earlier study that it was usually pointless to ask people why they did not do something, but we were unable to demonstrate in the laboratory that the conclusion held for these questions.

What Kinds of Questionnaires Can Benefit from the Laboratory Approach to Testing?

We believe that the intensive interview can and should be used to test all questionnaires, irrespective of

- previous use of the questionnaire in another survey,
- plans to field pretest the questions, or
- the expertise of the questionnaire designer.

In other words, all questionnaires can benefit from this approach. We have had experience with a wide variety of situations (some of which are described below) and feel strongly that the use of the intensive interview is indicated in all of them.

1. *Questions that were used in previous studies:* Occasionally we are asked to test questions that have been used in other surveys, sometimes even our own surveys. The questions had been thoroughly reviewed, field pretested, and revised. Yet significant sources of response error were found by intensive interviewing, sometimes after only two or three interviews. In retrospect, many of these problems seem obvious; it seems that the flaws should have been identified during the traditional questionnaire testing process. However, the fact remains that the flawed questions survived the traditional process and were used in at least one survey.

For example, the following questions were included in a series on stress in the 1985 Health Promotion Disease Prevention Supplement to the National Health Interview Survey:

> "1a. In the past year, did you think about seeking help for any personal or emotional problems from family or friends?
> 1b. From a helping professional or a self-help group?
> 2a. Did you actually seek any help?"

These questions proved to have multiple problems. First, laboratory respondents reported that they did not differentiate between thinking about seeking help and actually seeking help, so 2a seemed redundant. Second, "from family or friends" is misplaced; it sounds as though they were the cause of the problems, rather than the source of help. Third, the question does not refer specifically to stress-related problems, as was intended by the sponsor. Fourth, 1b cannot be understood unless at least part of the question stem is repeated.

2. *The intensive interview should also be used routinely on questionnaires that will be field tested:* This is the situation most often encountered at NCHS. Intensive interviews help at two points in the NCHS questionnaire design process. They reduce the number of questionnaire problems that arise in the field pretest, so that major revisions are less likely to be needed subsequent to the pretest. They also provide a way to test the revised post-field test questions, and thereby reduce the likelihood that new problems are introduced and left undetected.

3. *Finally, intensive interviews are especially useful for new questions that will not be field pretested:* Although NCHS policy is to field pretest all questionnaires, we are aware that some situations, such as epidemiological outbreak investigations of food poisonings or contagious diseases, require the use of untested questions. In such emergencies, even a very small number of intensive interviews (possibly with the first cases interviewed) will often greatly improve the questions. Our favorite example of an outbreak question that could have benefited from a few intensive interviews is:

> Have you smelled anything unusual since September 1, 1987?

We are not certain what was intended by this question; the context gave no clue. We were told that this question, administered to a group of prisoners, was meant to elicit reports of contamination from toilets that are sometimes purposely stopped up so that they overflow and the prisoners have to be moved out. If so, it clearly missed the mark.

After nearly 4 years of experience with intensive interviews, we have concluded that much more attention should be given to testing questionnaires; response errors are much more common than we had expected, and many of the questionnaire problems that lead to response error can be successfully detected and corrected by intensive interviews. We have used the intensive ap-

proach primarily for questionnaires designed for face-to-face interviews, but it has also proved useful for self-administered forms and telephone survey instruments. Virtually every intensive interview, properly conducted, provides valuable information about respondent reactions, interpretations, ability to recall, and other factors affecting response quality. In fact, it is difficult to imagine a questionnaire that would not benefit from this kind of testing.

With Such Small Sample Sizes, How Do You Judge Which Problems Must Be Addressed and Which Problems Can Be Ignored?

It is important to evaluate several aspects of each problem observed in intensive interviews before taking action. We generally follow these steps:

1. Determine how easily the problem can be solved. If there seems to be a quick, easy fix, we simply make the revision and go on to the next question.

2. If there is agreement that the problem can be solved, but there is a price to be paid in that the questionnaire must be made longer or more complex to solve the problem, then the problem is evaluated further. If the problem is likely to happen fairly often and/or it is likely to interfere with the flow of the interview or result in significant response error, then the necessary changes are made. If it is impossible to tell how often a problem will occur or how it will affect the interview or the response error, the problem is noted and a decision is deferred until after the pretest.

3. If the problem seems to be serious but there is no clear solution, we will continue testing until we find an approach that seems promising; often decisive information is only obtained from the field pretest. Ideally, two or more "promising" approaches should be pretested in this case, but because of the extremely tight schedules for the surveys, this has not been possible to date.

4. Sometimes the problem seems to be with some aspect of the *concept* being measured, which is a more fundamental difficulty than question wording. These problems must always be resolved with the subject matter experts so that there is agreement on what is to be measured in the survey and whether it can be measured.

Some Say That with Enough Probing, Flaws Can Be Found in Any Question. How Do You Decide That the Questions Have Been Adequately Tested?

At NCHS the limited amount of time available for lab testing makes it necessary to concentrate on the most flagrant problems first; there is little time for nitpicking. This is difficult to measure objectively, but there are some subjective measures supporting this statement. One supplement tested recently, which was developed by one of the most experienced questionnaire designers in the Center, was analyzed to determine the type and frequency of problems that were found. Although the questionnaire seemed reasonable in the initial review, intensive interviews revealed 7 serious problems in only 21 questions. There was complete agreement among the lab staff and the questionnaire designer that the problems compromised data quality and had to be corrected. A similar analysis of a longer supplement designed by a less experienced person had a much higher problem rate: 39 problems were detected in 68 questions. These questionnaires were tested only briefly in the lab; each supplement was administered to approximately 20 volunteers, with periodic breaks in interviewing for revisions. As this indicates, we generally find so many flagrant problems in the few weeks we are able to work on the supplements that our limited resources are completely expended in arriving at a questionnaire that everyone thinks is marginally acceptable.

References

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

Royston, P., Bercini, D., Sirken, M., & associates. (1986). Questionnaire Design Research Laboratory. *1986 Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 703-707). Washington, DC: American Statistical Association.

FEATURE PAPER

Coding Behavior in Pretests to Identify Unclear Questions

Floyd J. Fowler, Jr.

Introduction

It is axiomatic that survey questions should be clear; they should mean the same thing to all respondents. In practice, however, it is not easy for researchers to know when their questions include unclear terms. Standard pretest procedures, whereby experienced interviewers take from 10 to 50 interviews, then report back to the researchers about problems encountered, clearly are an important part of the question development process. However, the standards used by pretest interviewers to decide when a question is unclear are necessarily subjective and difficult to define.

This research evaluates strategies for making the pretesting of survey questions more objective, scientific, and useful. Coding interviewer behavior has been used to evaluate interviewer performance (Cannell, Fowler, Marquis, 1968; Cannell & Oksenberg, 1988). Morton-Williams & Sykes (1984) explored the use of interaction coding to evaluate comprehension and, based on our work to date, we believe that coding the interaction between respondents and interviewers during pretests represents an important technique for evaluating survey questions.

Coding pretest behavior is most easily handled by tape recording interviews, and that is how we have done our studies. The procedures for adding behavior coding to standard pretests are comparatively easy. It can be done for personal interview pretests or for those by telephone. Interviewers simply ask respondents if they can tape record the pretest interview; respondents seldom fail to give their permission.

Floyd J. Fowler, Jr. is with the Center for Survey Research, University of Massachusetts at Boston.

This research was supported by Grant No. 5 RO1 HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment. Senior staff on the project included Charles F. Cannell, Graham Kalton, Lois Oksenberg, and Katherine Bischoping at the University of Michigan, Ann Arbor.

The interaction coding schemes focus on the behaviors of the interviewers and the respondents. Each question produces an interaction, or turn: the interviewer asks a question, the respondent says something, then the interviewer says something, then the respondent says something until the respondent finally gives an adequate answer or the interviewer gives up. The ideal question would always be read exactly as written and answered adequately by the respondent on the first try; that would be one interviewer turn and one respondent turn. Deviations from this ideal are likely to be meaningful indications of an imperfect question. The interaction can be captured by coding either the respondent or the interviewer side; coding both sides of the interaction is probably not necessary.

This paper discusses the value of such coding to identify unclear concepts and is aimed at demonstrating three points:

1. Interaction coding is a useful way of identifying problems.
2. Questions can be rewritten to clarify terms and reduce those problems.
3. Unclear terms are significant sources of error in estimates based on surveys.

This research focused on a health interview survey consisting of 60 questions drawn from instruments used in several national health surveys mainly by academic or government organizations. Questions were drawn to include examples from the various subjects typically covered in health services research studies, including use of health services, health beliefs and attitudes, health behaviors, and health status. Questions were not chosen because they were judged to be problematic. Most important, all of these items had been subjected to standard pretest procedures by professional survey organizations or government agencies and hence were representative of the kinds of items used in good quality survey research.

The result of this process was a 20-minute survey instrument with 60 items. A pretest was carried out by 11

interviewers with a sample of 110 respondents drawn from directory listings in southeastern Michigan. Before the interview began, interviewers asked respondent for permission to tape record the telephone interviews. If respondents would not agree, interviews were not taken.

When the interviews were completed, the tape recordings were coded by trained coders at the University of Michigan. For each question asked in each interview they coded whether the interviewer read the question exactly as worded or made changes and then coded the respondent's behavior after the question was read.

The results of this interaction coding were studied for evidence of problems with questions. One indication of a question problem was whether respondents asked for clarification in at least 15 percent of the pretest interviews; another was whether or not at least one inadequate answer was given in at least 15 percent of the interviews. Although the selection of cutpoints was somewhat arbitrary, it proved a reasonably easy task to identify ambiguities and problems with questions that met the above criteria.

Several kinds of question problems led to comprehension problems. In a few cases the problem seemed to be primarily with the order of the words, so that respondents had difficulty retaining all the parts of the question that they needed to remember to answer the question. The solution to such problems usually was to reorder the question.

In other questions the reason for requests for clarification and/or inadequate answers seemed to be that at least one key term in the question was ambiguous. In those cases the solution was to change the question to clarify the meaning of the key terms. Five such questions are addressed here. It should be noted that the basic objectives were not changed; hence, questions that posed a task for respondents that they could not perform easily would not be improved by these changes.

The revised survey instrument was readministered by new groups of interviewers using procedures identical to those in the initial phase of the project. Respondent samples were drawn from the same sample frame. The interviews again were tape recorded and the interactions were coded.

The results presented focus on the effect of the revised wording on the frequency at which respondents ask for clarification, the frequency at which inadequate answers were given, and the distribution of responses to the questions.

Results

There is not an easy way to discuss the questions in aggregate because the problems posed by each of the original questions were different. Therefore, the issues and results for the five questions studied, are presented one at a time.

The first question deals with the consumption of eggs, and the results are shown in Table 1. People were asked how many servings of eggs they ate on typical days. Clarification was requested in almost a third of the interviews. The main ambiguity lay in what constituted a serving of eggs, although "typical day" may also have been unclear.

Table 1. Responses to questions related to consumption of eggs

Original
What is the number of servings in a typical day?

Revision
On days when you eat eggs, how many eggs do you usually have?

Servings	Original	Revision
1	80%	33%
2	15	61
3 or more	5	6
	100%	100%
Interviews with requests for clarification	33%	0.0%
Interviews with inadequate answers	12%	0.0%

When the question was revised to ask people how many eggs they ate on days when they ate eggs, there were major effects both on the distribution of answers and on the interaction. It is clear that many, but not all, people thought that a serving equaled two eggs; others thought it equaled only one. In any case, one gets a very different distribution and one would guess a more interpretable distribution of egg consumption with the revised question. At the same time, requests for clarification and inadequate answers dropped to 0 (Table 1). This is a clear example of how an unclear concept shows up in coding pretest behavior, and how clarifying a term affects the respondent's behavior and the distribution of answers.

The next question, shown in Table 2, asked about the consumption of butter. One ambiguity in that question was whether or not margarine counts as butter. In the original question clarification was requested in over 15

Table 2. Responses to questions related to butter consumption

Original
What is the average number of days each week you have *butter*?

Revision
The next question is just about butter. Not including margarine, what is the average number of days each week you have butter?

Days	Original	Revision
None	33%	55%
1	13	14
2–6	29	22
7	23	9
	100%	100%
Interviews with requests for clarification	18%	13%
Interviews with inadequate answers	15%	12%

Table 3. Responses to questions related to physical activities

Original
Do you exercise or play sports regularly?

Revision
Do you do any sports or hobbies involving physical activities, or any exercise, including walking, on a regular basis?

Exercise Regularly?	Original	Revision
Yes	48%	60%
No	52	40
	100%	100%
Interviews with requests for clarification	5%	0%
Interviews with inadequate answers	20%	12%

percent of the interviews, and inadequate answers were given in 15 percent of the interviews. When the question was revised to specifically exclude margarine, there was a very significant decrease in the number of days that respondents said they had any butter at all.

The changes also may have affected requests for clarification and inadequate answers, although the effects were very small at best. This pattern illustrates that unclear concepts are only one cause of these behaviors. A major cause of inadequate answers to this question was that people were supposed to come up with an exact number of days, and in some cases they found that a hard task to do. The clarification of what was included in butter did not affect the difficulty of the response task itself.

Table 3 presents a question that pertains to reported exercise, an important focus of health behavior surveys. In fact, the rate of requests for clarification was not high, but the rate of inadequate answers made the cutpoint. One issue seemed to be what counted as exercise and whether or not walking counted.

The revised question seemed to have an effect on the distribution of answers. The change from 48 to 60 percent who said they exercised regularly is nearly statistically significant. From an interaction perspective, there also was a reduction both in the rates of requests for clarification and in the rates of inadequate answers.

The question in Table 4 asked whether the last visit to a doctor occurred at a health maintenance organization (HMO). Both requests for clarification and the frequency of inadequate answers suggested a problem with the question, and part of the problem seemed to be understanding what constituted a Health Maintenance Organization.

The question was revised to clarify that. Also, it was broken into two questions, the first pertaining to whether respondents belonged to a Health Maintenance Organization, the second, whether the last visit to a doctor was through the Health Maintenance Organization plan.

This change seems to have had a significant effect on the responses, because fewer people reported that their most recent visit was through a Health Maintenance

Table 4. Responses to questions related to health care costs

Original
Was that place a health maintenance organization or health care plan (that is, a place you go for all or most medical care, which is paid for by a fixed monthly or annual amount)?

Revision
Do you belong to an HMO or health plan that has a list of people or places you go to, in order for the plan to cover your health care costs?

Was your last visit to a medical doctor covered by your health plan?

Last see doctor at HMO?	Original	Revision
Yes	39%	23%
No	61	
	100%	77
		100%
Interviews with requests for clarification	17%	2%
Interviews with inadequate answers	27%	18%[a]

[a]12% and 6% respectively for the two revised questions

Organization. In addition, there was a marked decrease in the rate of requests for clarification of this question, and there probably were fewer inadequate answers, despite the fact that people were now answering two questions, which gave them twice as many chances to give inadequate answers.

Finally, Table 5 presents the results for a standard question regarding disability days over the preceding

Table 5. Responses to questions related to disability days

Original
During the past 12 months, that is, since January 1, 1987, *about* how many days did illness or injury keep you in bed more than half of the day? (Include days while an overnight patient in a hospital.)

Revision
The next question is about extra time you have spent in bed because of illness or injury (including time spent in the hospital). During the past 12 months since July 1, 1987, on about how many *days* did you spend several extra hours in bed because you were sick, injured, or just not feeling well?

Days	Original	Revision
None	57%	48%
1–7	36	33
8 or more	7	19
	100%	100%
Mean	2.6	4.0[a]
Interviews with requests for clarification	15%	17%
Interviews with inadequate answers	7%	30%

[a]One person reported 90 days in the second sample, almost twice as many as the next person in either sample. Removing that person reduces the mean to 3.4. In either case, the difference is not statistically significant.

year. In its initial form, 15 percent of the pretest respondents requested clarification. One problem seemed to be ambiguity about what was meant by half a day; there also seemed to be confusion about whether or not extra time in bed for vague maladies (rather than specific conditions) should be counted.

The revised question, designed to clarify those two points, seems to have changed the answers. In particular, more people reported 8 or more disability days. The mean was higher, too, although it did not reach the .05 level of significance, given the sample size and variance. The distribution change is evidence that important ambiguity existed, since the two questions should be equivalent in meaning. However, it is not clear that the new version is a better question. Based on the coding of pretest behavior, it may well be a worse question. Alternatively, the real problem with the question, that it poses a virtually impossible recall task, may be much more apparent to respondents, and show up more clearly in the coding, when the meaning of what is really wanted is clarified.

Discussion

We think coding behavior should be a routine part of the pretest process. Not only is it a way to identify unclear questions, it also identifies questions that are difficult to read as worded, that are confusing, or that pose a difficult task for respondents.

Better pretesting procedures are needed. Unstructured feedback from pretest interviewers is useful but it is not systematic, and the standards they use are not well defined. We studied health questions that had been used in major health surveys and, presumably, had been subjected to state-of-the-art pretests. At least half of the questions tested had a significant problem that was apparent in the coding of pretest behaviors. Such coding seems to give reliable, objective information about some important problems with the questions, including the presence of unclear terms.

There can be little doubt that a common respondent understanding of key terms is important to good measurement in surveys. The examples presented show that ambiguity can produce major errors, and that clarifying the meaning of key terms can both change the distribution of answers and effect measurable change in behavior during interviews.

All problems of comprehension do not show up in behavior coding. Sometimes respondents will answer questions they do not understand without asking for clarification. We think focused group discussion and laboratory studies should be done before a survey instrument is subjected to formal pretesting. Intensive reinterviews and "think aloud" interviews are particularly good ways to identify comprehension problems. However, once the developmental work is done and the instrument is ready to be tested in a realistic interview setting, coding behavior is an objective, effective, and reasonably low-cost way to identify significant remaining problems with questions.

In at least three areas we need to learn more about how to use these procedures.

1. *Cutpoints*: At what rate do problems that show up in interactions in fact constitute problems that need to be solved? In all probability, the choice of cutpoint will depend on the role of a measure in a research project. In any case, a strength of the procedure is that it is reliable and scientific, so that we can study the significance of different rates of problems in a consistent, objective way.

2. *Diagnosing problems*: How do we go from the indications in the interactions that a problem exists to identifying the nature of the problem? The input of the pretest interviewers and the coders who code the behaviors can be used to help diagnose the problems, and we are working on a protocol to use efficiently what they have to offer.

3. *Solving problem questions and how to write good survey questions*: The behavior coding only indicates a problem that requires attention, but it is left to the ingenuity of the researcher to figure out the solution. Over time, we hope to produce a more useful set of generalizations about what a good question should look like.

In short, there is still work to be done on how best to use this technique, but it is clear that more systematic, objective standards for survey questions are very much needed. Coding behavior in pretests is likely to be one important part of an increasingly scientific process of evaluating questions as measures.

References

Cannell, C. F., Fowler, F. J. & Marquis, K. H. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. *Vital and Health Statistics* (Series 2, No. 26). Washington. DC: U.S. Government Printing Office.

Cannell, C. F. & Oksenberg, L. (1988). Observation of behavior in telephone interviewers. In R. Groves, P. P. Biemer, L. E. Lyberg & associates (Eds.) *Telephone survey methodology.* New York: Wiley & Sons.

Morton-Williams, J. & Sykes, W. (1984). The use of interaction coding and follow-up interviews to investigate comprehension of survey questions. *Journal of the Market Research Society,* 26, (2), 109-127.

Comparison of Responses to Similar Questions in Health Surveys

John P. Anderson, Robert M. Kaplan, and Margaret DeBon

Introduction

Over the last several decades, recognition of the need for sensitive indicators of health status and quality of life has increased. This need is apparent for several reasons. First, current health indicators are inadequate for capturing many of the health status variables that are associated with the need for health care. Measures of mortality provide hard end points but ignore all of those who are alive. Fries and associates (1989) emphasize that the likelihood of extending current life expectancy for adults is very small. Thus, there is remarkably little evidence that major medical and preventive interventions that apply to those who have survived their first years of life actually make people live longer. Yet, as Fries and colleagues (1989) have argued, substantial public health benefits may be achieved by compressing morbidity toward the end of the life cycle. Evaluating these interventions will require more sensitive measures of health outcome. Current data from the National Health Interview Survey (NHIS) provide information that only a minority of the U.S. population are, by their standards, in ill health. In 1985, for example, 90 percent of the U.S. population was reported to be in excellent, very good, or good health. A substantial majority (86 percent) reported no activity limitations (Dawson and Adams, 1987).

This paper suggests that many current techniques for evaluating health status and quality of life are insensitive for detecting important variations in health status. Specifically, it is argued that variations in the experience of what we have come to call Symptom/Problem Complexes (CPX) are, by patient-citizen preference standards, highly important to how they come to evaluate their health status. This implies that approaches that rely exclusively on dysfunction are seriously deficient in their sensitivity to important dimensions in measuring health status. Data from several studies are presented to suggest that seemingly minor variations in the wording of survey questions can produce significant differences in the estimates of the extent of dysfunction in and overall health status of populations.

Although a growing number of studies now incorporate health-related quality of life measures, there has been a strong emphasis on cost savings and time efficiency. Self-administered questionnaires are frequently assumed to be the better alternative because they are cheap and easy. Over the last two decades, our group has worked toward the development of a General Health Policy Model (Kaplan and Anderson, 1988). One of the objectives of this line of research is the development of a valid and reliable questionnaire for assessing health-related quality of life. Several studies have identified problems, particularly in the underreporting of dysfunction (Reynolds & associates, 1974; Stewart & associates, 1981). In several of our studies, both self-administered and interviewer-administered questionnaires were given to the same respondents. The results are of interest not only because of mode of administration but because they provide information on type of question. This paper summarizes three studies from our current research program. All of these studies use the Quality of Well-being (QWB) scale and instrument, which will now be briefly described.

Quality of Well-being Scale

The QWB scale combines preference-weighted measures of symptoms and functioning to provide a numerical point-in-time expression of well-being, which ranges from zero (0) for death, to one (1.0) for asymptomatic

John P. Anderson, Robert M. Kaplan, and Margaret DeBon are with Division of Health Care Sciences, Department of Community and Family Medicine, University of California, San Diego.

This research was supported in part by Grant No. R18 HS 05617 from the National Center for Health Services Research and Health Care Technology Assessment and Grant AR 33489 from the National Institutes of Health.

Table 1. List of Quality of Well-being Scale Symptom/Problem Complexes (CPX) with calculating weights

CPX no.	CPX description	Weights	CPX no.	CPX description	Weights
1	Death (not on respondent's card)	−0.727	13	Headache, or dizziness, or ringing in ears, or spells of feeling hot, or nervous, or shaky	−.244
2	Loss of consciousness such as seizure (fits), fainting, or coma (out cold or knocked out)	−.407	14	Burning or itching rash on large areas of face, body, arms, or legs	−.240
3	Burn over large areas of face, body, arms, or legs	−.367	15	Trouble talking, such as lisp, stuttering, hoarseness, or inability to speak	−.237
4	Pain, bleeding, itching, or discharge (drainage) from sexual organs—does not include normal menstrual (monthly) bleeding	−.349	16	Pain or discomfort in one or both eyes (such as burning or itching) or any trouble seeing after correction	−.230
5	Trouble learning, remembering, or thinking clearly	−.340	×17	Overweight *or underweight* for age and height or skin defect of face, body, arms or legs, such as scars, pimples, warts, bruises, or changes in color	−.186
6	Any combination of one or more hands, feet, arms, or legs either missing, deformed (crooked), paralyzed (unable to move) or broken—includes wearing artificial limbs or braces	−.333	18	Pain in ear, tooth, jaw, throat, lips, tongue; missing or crooked permanent teeth—includes wearing bridges or false teeth; stuffy, runny nose; any trouble hearing—includes wearing a hearing aid	−.170
7	Pain, stiffness, weakness, numbness, or other discomfort in chest, stomach (including hernia or rupture), side, neck, back, hips, or any joints of hands, feet, arms or legs	−.299	19	Taking medication or staying on a prescribed diet for health reasons	−.144
8	Pain, burning, bleeding, itching, or other difficulty with rectum, bowel movements, or urination (passing water)	−.292	20	Wore eyeglasses or contact lenses	−.101
9	Sick or upset stomach, vomiting or loose bowel movements, with or without fever, chills, or aching all over	−.290	21	Breathing smog or unpleasant air	−.101
10	General tiredness, weakness, or weight loss	−.259	22	No symptoms or problem (not on respondent's card)	−.000
11	Cough, wheezing, or shortness of breath with or without fever, chills, or aching all over	−.257	23	Standard symptom/problem (not on respondent's card)	−.257
12	Spells of feeling upset, being depressed, or of crying	−.257	×24	*Trouble sleeping*	−.257
			×25	*Intoxication*	−.257
			×26	*Problems with sexual interest or performance*	−.257
			×27	*Excessive worry or anxiety*	−.257

optimum functioning. Table 1 presents 25 Symptom/Problem Complexes along with their preference weights. Use of this CPX list does not require any assumptions about the intensity or duration of symptoms and problems nor the underlying pathology, if any. This measure simply indicates that symptoms are present or absent on a given day.

Quality of Well-being also involves three scales of function: Mobility (MOB), Physical Activity (PAC), and Social Activity (SAC). Each step on these scales has its own associated preference weight. These are reported in Table 2, along with the single-day QWB calculating formula (formula 1). In the General Health Policy Model, QWB inputs are integrated with terms for the number of people affected and the duration of time affected to produce the output expression of Well-years (formula 2).

Study I: Evaluation of Self-Administered QWB Items

Method

Data from this analysis come from a household interview survey of a sample of 1,324 subjects. These subjects included 866 randomly selected respondents, 369 randomly selected children, and 89 persons with a physical dysfunction who were selected on the basis of responses to screening questions. Seventy-seven percent of those initially contacted completed the study.

Figure 1 characterizes the types of data available. Each respondent answered all questions relevant to functioning on the three scales in a self-report mode. In addition, respondents were assessed by a trained interviewer. The order of presentation was counterbalanced to control for order effects. In each case, questions were

Table 2. Quality of Well-being General Health Policy Model elements and calculating formulas (function scales, with step definitions and calculating weights)

Step No.	Step definition	Weight
	Mobility scale (MOB)	
5	No limitation for health reasons	−0.000
4	Did not drive a car, health related; did not ride in a car as usual for age (younger than 15 yr), health related, and/or did not use public transportation, health related; or had or would have used more help than usual for age to use public transportation, health related	−.062
2	In hospital, health related	−.090
	Physical activity scale (PAC)	
4	No limitations for health reasons	−.000
3	In wheelchair, moved or controlled movement of wheelchair without help from someone else; or had trouble or did not try to lift, stoop, bend over, or use stairs or inclines, health related; and/or limped, used a cane, crutches, or walker, health related; and/or had any other physical limitation in walking, or did not try to walk as far or as fast as others the same age are able, health related	−.060
1	In wheelchair, did not move or control the movement of wheelchair without help from someone else, or in bed, chair, or couch for most or all of the day, health related	−.077
	Social activity scale (SAC)	
5	No limitations for health reasons	−.000
4	Limited in other (for example, recreational) role activity, health related	−.061
3	Limited in major (primary) role activity, health related, but did perform self-care activities	−.061
2	Performed no major role activity, health related, but did perform self-care activities	−.061
1	Performed no major role actvity, health related, and did not perform or had more help than usual in performance of one or more self-care activities, health related	−.106

Calculating formulas

Formula 1. Point-in-time well-being score for an individual (W):

$$W = 1 + (\text{CPX}wt) + (\text{MOB}wt) + (\text{PAC}wt) + (\text{SAC}wt)$$

where wt is the preference-weighted measure for each factor and CPX is Symptom/Problem complex. For example, the W score for a person with the following description profile may be calculated for one day as:

CPX-11	Cough, wheezing, or shortness of breath, with or without fever, chills, or aching all over	−0.257
MOB-5	No limitations	−.000
PAC-1	In bed, chair, or couch for more or all of the day, health related	−.077
SAC-2	Performed no major role activity, health related, but did perform self-care	−.061

$$W = 1 + (-.257) + (-.000) + (-.077) + (-.061) = .605$$

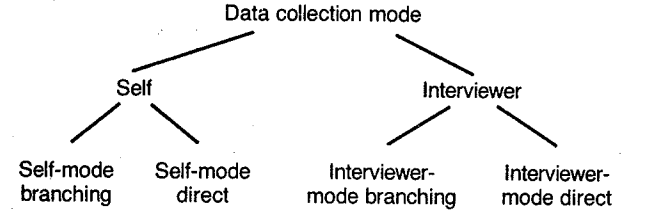
Formula 2. Well-years (WY) as an output measure:

$$WY = \text{No. of Person} \times (\text{CPX}wt + \text{MOB}wt + \text{PAC}wt + \text{SAC}wt) \times \text{Time}$$

presented in both a branching and direct mode. In the branching mode, the respondents answered an algorithmic series of closed questions and branching follow-up probes. First, questions asked whether the subjects actually performed a specific activity. If they did not, a probe question was used to determine the reasons for nonperformance. Both yes and no answers were probed in fuller detail. Strict criteria were used to code whether or not reasons for nonperformance were related to health. The questions were designed for either interviewer or self administration. Examples of the branching and direct questions for the self-administered questions for the mobility portion of the Quality of Well-being are given in Table 3.

In the self mode, the respondents were directed to read definitions for all steps in the scales. The respondent then reported to the interviewer the number of the step on each scale that best described themselves and/or the other subjects for whom they reported. The self-read definitions required the respondent to interpret whether any nonperformance of activities was due to

Figure 1. Categories of evidence



SOURCE: Anderson & associates (1986)

Table 3. Self-mode direct and branching question patterns by study from mobility scale

Initial Survey, Direct Mode Card B (Mobility Scale)	Follow-up Survey, Branching Mode Card II (Mobility Scale, Over 16)
1. In a special unit of a hospital such as an operating or recovery room, intensive care unit, incubator, isolation ward, for any part of a day	In each category choose the numbers* that A. Spent any part of the day or night as a bed patient in a hospital, nursing home, mental institution, home for the retarded, or similar place. A1. Yes A2. No
2. In a hospital, nursing home, mental hospital, home for retarded as a patient	
3. Needed help to go outside, or stayed inside all day for health reasons	B. Driving B1. Drove car (or motor vehicle) B2. Did not drive, for health reasons B3. Did not drive, for reasons not related to health
4. Could go outside without help, but could not drive and/or could not use public transportation without help from another person. (For a child: needed more help to travel than usual for age.)	C. Public Transportation Use bus, train, plane or subway — C1 Without help from anyone else; C2 With help from another person for health reasons Did not use bus, train, plane, or, subway. — C3 For health reasons; C4 For reasons not related to health.
5. Able to both drive and use public transportation (bus, train, etc.) without help. (For a child: able to travel as usual for age.)	

*Numbers in self-mode branching do not correspond to scale steps.
SOURCE: Anderson and associates, 1986

health-related reasons. Although the respondents were required to read the items, they were not requested to record the information on their own. Thus, test of the self-mode should have created the most favorable conditions for self-administration.

During long interviews a variety of other questions and observations were made. For example, the interviewers also engaged the respondent in open-ended discussion, completed interviewer notes, and tape recorded each interview. This information was used to estimate (1) how well the questions were being understood; (2) how closely the respondents understanding matched the purpose of the question; and (3) how closely the categoric answers in both types of administration matched the actual situation. When a discrepancy between the two sets of categoric responses was observed, all of this information was systematically studied to estimate the most likely true classification for the respondent.

Results

Initial analysis suggested that correlations between two different modes of administration were very high. Indeed, they tended to be .98 or higher! Even for those respondents who were highly dysfunctional, correlation between modes of administration tended to be .90 or above. Despite high correlations between overall QWB scores for the different modes of administration, there were a substantial number of inconsistencies in function classification between these modes. To evaluate these, validity analysis was conducted (Anderson and associates 1986; 1988). Table 4 summarizes the method used to assess sensitivity and specificity of the different forms. Although these methods are common, the table includes a few uncommon terms. Each of the respondents are classified into one of five categories. These include (a) report of dysfunction when there is, indeed, dysfunction; (b) reports of dysfunction when there is no dysfunction; (c) reports of no dysfunction when, indeed, there is dysfunction; and (d) report of no dysfunction when there is no true dysfunction. The final category (e) is for people who correctly report they are dysfunctional but are placed in the wrong dysfunctional category. In addition to calculating sensitivity and specificity by standard methods, we offer new concepts for predictive value of dysfunctional and predictive value of function. The predictive value of dysfunction is the ratio of those who report the correct dysfunctional category over all those reporting dysfunction. The predictive value of functional reporting is the ratio accurately reporting function over all reports of functioning. Table 5 displays the validity characteristics for both modes of administration. The two modes of administration differ dramatically in accurately classifying dysfunction when the dysfunction state is compared to actual dysfunction. These errors are reflected in the sensitivity of the measure. Analysis suggested that the sensitivity of the PAC scale was .45 in the self-administered version. In other words, only 45 percent of the actual dysfunction was captured. In contrast, the interviewer-administered version accurately classified 86 percent. The predictive value of dysfunction was also low in the self-administered versions and considerably higher in the interviewing administered version. Specificity was high for both modes of administration.

In the early days of development of the Quality of Well-being scale, a self-administered questionnaire was seen as highly desirable. The interviewer mode was cho-

Table 4. Measurement categories and validity characteristics modified for multiple state analysis

Measurement categories for multiple states			
	ACTUAL DYSFUNCTION	ACTUAL (FULL) FUNCTION	
Reported dysfunction	(a) Correctly classified dysfunction (e) Misclassified dysfunction	(b) False dysfunction	Total reported dysfunction (= a + b + e)
Reported (full) function	(c) False function	(d) Full function	Total reported (full) function (= c + d)
	Total dysfunction (= a + c + e)	Total actual (full) function (= b + d)	

Validity characteristics modified for multiple states

$$\text{Sensitivity} = \frac{\text{Correctly classified dysfunction}}{\text{Total actual dysfunction}} = \frac{a}{a + c + e}$$

$$\text{Predictive value dysfunctional} = \frac{\text{Correctly classified dysfunction}}{\text{Total reported dysfunction}} = \frac{a}{a + b + e}$$

$$\text{Specificity} = \frac{\text{Full function}}{\text{Total actual (full) function}} = \frac{d}{b + d}$$

$$\text{Predictive value functional} = \frac{\text{Full function}}{\text{Total reported (full) function}} = \frac{d}{c + d}$$

SOURCE: Anderson & associates, 1986

sen as a gold standard against which to evaluate the less expensive self-administered mode. However, as these data suggest, there may be serious problems with the sensitivity of the self-administered mode. Other studies also have reported problems in detecting limitations with self-administered scales. For example, the inability to detect limitations with single closed-ended questions, with later acknowledgment of limitations in the same interview, was also reported in data from the National Health Interview Survey (Cannell and others 1977). In the Rand Health Insurance Study, there was missing or inconsistent information on 37 percent of the functional limitations reported. We observed problems with the self-reporting of functional limitations in 28.7 percent of the cases with self-administered questionnaires.

There are many potential explanations for these findings. One is that respondents misunderstand questions in the self-administered forms. This problem becomes worse as the complexity of the questions increases. Thus, for branching questions, sensitivity will decrease. The availability of a trained interviewer allows the determination of actual performance versus nonperformance of activities. In addition, nonperformance of activities can be evaluated as related to health or for nonhealth reasons. Further, an interviewer can assess the specific days for which there was a problem. Evidence was developed that sequential branching questions that require an interviewer can reliably penetrate the complexity of quality of life and dysfunctional states. It was also noted that a very high proportion of the population experiences at least some minor dysfunction or symptom on a particular day. In the community survey, for example, interviewer-administered questionnaires with branching patterns identify only about 11 percent of the population as completely functional and asymptomatic, by comparison to the NHIS identification of 86 percent of the population as completely functional, without regard to experience of symptoms. Clearly, this wide difference means variations in what has been called

Table 5. Validity characteristics by data collection method and format

Scale Validity characteristics	Initial survey direct	Follow-up survey combined days branching
Self mode, direct and branching		
Mobility		
Sensitivity	0.68	0.66
Predictive value dysfunctional	.56	.41
Specificity	.98	.95
Predictive value functional	.99	.98
Physical and social activity		
Sensitivity	.45	.61
Predictive value dysfunctional	.59	.73
Specificity	.99	.99
Predictive value functional	.94	.96
Interviewer mode, branching		
All scales combined		
Sensitivity	.89	.86
Predictive value dysfunctional	.93	.91
Specificity	.99	.99
Predictive value functional	.99	.99

SOURCE: Adapted from Anderson and associates, 1986

high-level wellness are absolutely critical for sensitive and accurate measurement of health status.

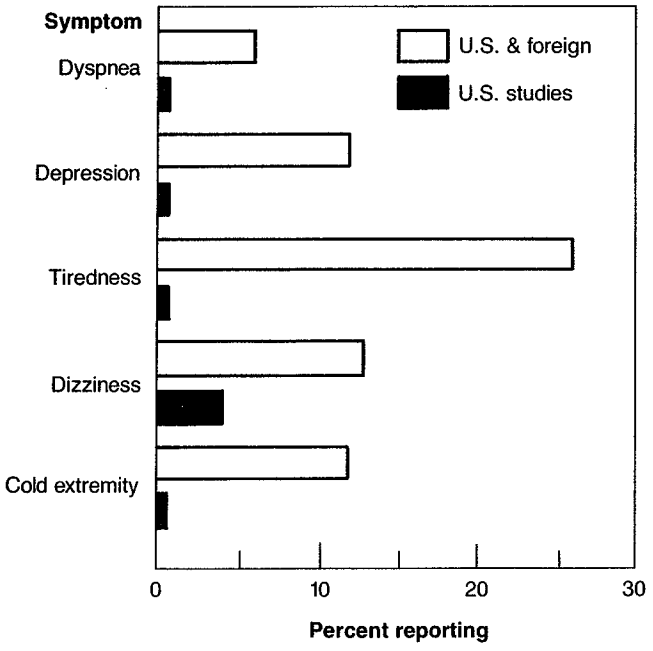
This work on structured interviews about function led to questions about the value of self-report symptom inventories. That issue was investigated in Study II.

Study II: An Experiment on Symptom Reporting

Clinical studies with the same experimental design show considerable variability in the effects of treatment on symptoms. This may be of particular concern in studies of drug side effects. Consider, for example, Figure 2. The data for this figure were taken from an advertisement for Atenolol that was published in the *Journal of the American Medical Association*. In the very small print of the advertisement, side effects of the medication were reported. The ad separated data from U.S. studies and U.S. plus foreign studies. As the figure suggests, side effects in the American studies are quite rare. Yet the very same side effects are actually quite common in U.S. plus foreign studies. Consider, for example, tiredness which occurs in 0.6 percent of U.S. studies, but 27 percent of U.S. plus foreign studies. It is presumed that U.S. plus foreign studies are combined in order to dilute what may be very common side effects in the foreign studies. Similar results are apparent for dyspnea, depression, and other symptoms.

Why do these results from studies of the same product produce such different results? One explanation is in the way that symptoms were assessed. Typically, U.S. drug studies ask about only a small number of symptoms. Then patients are asked in a free format if they have any other symptoms. In European studies, there is a systematic symptom-by-symptom inquiry. In our work on the Quality of Well-being scale, respondents are presented with a list of symptoms that is meant to be exhaustive. Then they are asked to identify which symptom complexes they have experienced for each day over the last 6 days. An alternative procedure would be to have the interviewer read each individual symptom and ask whether that symptom had been experienced (Eakin, Kaplan, & Ganiats, 1989). These formats may lead to differential report rates.

Figure 2. Symptoms associated with use of atenolol in U.S. and U.S. and foreign studies



Subjects

The participants in the study were 82 adults who were being cared for by the family medicine practice at the University of California, San Diego. All were followed by their physicians for routine health problems or other conditions that do not require the attention of a specialist.

Procedure

The patients were randomly assigned to one of two groups. Group 1 was given the standard instruction which is:

> For most of these questions, I'll be asking about the past six days, that is, from (day/date) through (day/date). First, I would like to ask you about any health problems you might have had. Please look at this list one at a time and tell me the number of all the items that you had at any time during the past six days. Don't worry about how important or serious the problem was; if it was present at all in the last six days, please give me the number. Were there any health problems not on the list that you had at any time during the past six days?

For Group 2, the interviewer proceeded through the symptom problem list and requested the patients to report whether they experienced each item. The data analysis involved a *t*-test comparing the mean number of symptoms reported for each of the two conditions.

Results

The group receiving the standard instruction reported an average of 2.64 symptoms per day while the group receiving the item by item instruction reported an average of 2.86 symptoms per day. These differences were not statistically significant ($p = 0.55$).

Study III: Comparison of Similar Items on Different Standardized Questionnaires

The third study considers a somewhat different question. In this, we compared responses to very similar items that were developed for different standardized questionnaires. Specifically, we compared responses to questions on the Quality of Well-being scale with items on an arthritis-specific measure known as the Arthritis Impact Measurement Scale (AIMS). There were several reasons for these comparisons. First, the Arthritis Impact Measurement Scale is commonly used in arthritis

Table 6. Comparison of similar items in Quality of Well-being and Arthritis Impact Measurement Scale

Items	Percent agreement	Percent QWB dysfunction	Percent AIMS dysfunction
AIMS 1. When you travel around your community, does someone have to assist you because of your health? MOB 2. On (day/date) were there reasons related in any way to your health that you did not (drive a car/ride in a car)? What were the reasons? On (day/date) (did you/would you) use more help from someone else than usual for your age?	98	0	2
AIMS 2. Are you able to use public transportation? MOB 3. On which of the past 6 days, if any, did you use public transportation, such as a bus, plane, train, or trolley? On (day/date) were there reasons related in any way to your health that you did not use public transportation? On (day/date) did you use, or would you have used, more help from someone else than usual for your age to take public transportation?	92	0/10*	8
AIMS 4. Are you in bed or a chair for most or all of the day because of your health? PAC 3. On which of the past 6 days, if any, did you spend most or all of the day in any type of chair or couch?	85	11	4
AIMS 6. Do you have any trouble either walking several blocks or climbing a few flights of stairs because of your health? PAC 6. On which of the past 6 days, if any, did you have any other physical limitation or not try to walk as far or as fast as most persons your age are able?	69	10	21
AIMS 7. Do you have trouble bending, lifting, or stooping because of your health? PAC 4. On which of the past 6 days, if any, did you have trouble, or not try, to lift, stoop, bend over, or use stairs or inclines?	76	9	15
AIMS 8. Do you have any trouble either walking one block or climbing one flight of stairs because of your health? PAC 6. On which of the past 6 days, if any, did you have other physical limitation or not try to walk as far or as fast as most persons your age are able?	67	25	8
AIMS 9. Are you unable to walk unless you are assisted by another person or by a cane, crutches, artificial limbs, or braces? PAC 5. On which of the past 6 days, if any, did you limp or use a cane, crutches, or walker?	64	36	0
AIMS 15. If you had the necessary transportation, could you go shopping for groceries or clothes? SAC 1B. If you had worked (or did work) on (day/date), were you limited in the amount or kind of work done, such as using special working aids, not doing certain tasks, taking special rest periods, or working only part of the day?	87	9	4
AIMS 26. When you bathe, either a sponge bath, tub, or shower, how much help do you need? SFC 4. Did not take bath for health reasons or had help to take bath (getting in or out of tub or shower, washing all parts of the body, etc.)	95	2	3
AIMS 27. How much help do you need in getting dressed? SFC 1. Did not dress for health reasons, or had help to dress (tying shoes, buttoning shirt, blouse, coat, etc.).	89	7	4
AIMS 28. How much help do you need to use the toilet? SFC 3. Did not use toilet for health reasons (e.g., bedpan) or had help to use toilet (getting on or off the seat, cleaning with tissues, etc.)	99	1	0
AIMS 31. During the past month how often have you had severe pain from your arthritis? CPX 7. Pain, stiffness, weakness, numbness, or other discomfort in chest, stomach, side, neck, back, hips, or any joints of hand, feet, arms, or legs.	91	9	0
AIMS 38. During the past month, how much of the time have you been in low or very low spirits? CPX 12. Spells of feeling upset, depressed, or crying.	47	1	52

research. It is believed to be more sensitive to clinical changes in arthritis patients because the items are arthritis specific. However, the Arthritis Impact Measurement Scale is often self-administered and therefore includes many of the same potential difficulties as do other self-administered questionnaires.

A second reason for conducting this analysis is that there is growing interest in imputing scores for one mea-

sure retrospectively from data collected using a different questionnaire (Erickson & associates, 1988, 1989). For example, the National Health Interview Survey does not include sensitive measures that can be used for quality-of-life evaluations. In addition, many policy analyses require data that are not available in the standard National Center for Health Statistics (NCHS) questionnaires. Nevertheless, items on the national survey are quite similar to those used in some quality-of-life measures. Thus, there is interest in imputing the more sensitive quality-of-life measures from responses given in the national surveys. These imputations make the assumption that responses from one measure can be accurately predicted from responses on another measure. Study III tests this assumption.

Method

The subjects were 92 adults with musculoskeletal diseases treated by the Scripps Clinic and Research Foundation. The rationale for selecting only patients with musculoskeletal disorders was that the Arthritis Impact Measurement Scale instrument was only appropriate to them. Using a nonhealthy population maximizes the number of estimated dysfunctional states in the population. The Quality of Well-being and Arthritis Impact Measurement Scale questionnaires were both administered by a trained interviewer during regular clinic visits. Table 6 shows the items in the two scales that are used for comparison. In addition, the table shows the percentage of patients for which there was agreement, defined as reporting a problem on both or neither of the items. Table 6 also shows the percentage of cases where only the Quality of Well-being questionnaire or only the Arthritis Impact Measurement Scale questionnaire detected health problems. As Table 6 suggests, there tended to be high agreement between the two measures for most items. Among 13 items with similar wording, the average agreement score was 82 percent. The Quality of Well-being detected more problems in eight items whereas the Arthritis Impact Measurement Scale detected more problems in 5 cases.

The cases of large discrepancy between the Arthritis Impact Measurement Scale and Quality of Well-being typically compared questions in which there were subtle differences in wording. For example, there was a large difference between Arthritis Impact Measurement Scale 9 and Physical Activity 5 from the Quality of Well-being. One difference in these questions is that the Quality of Well-being items ask about limping, and 25 patients reported a limp. The Arthritis Impact Measurement Scale does not inquire about limping.

Another disturbing discrepancy is between the AIMS question on depression and the Quality of Well-being symptom-problem for depression. A remarkable 78 percent of the arthritis patients reported depression on at least one measure. Among the 47 percent for which there was agreement, 55 percent reported depression on both scales whereas 45 percent reported it on neither. However, the Arthritis Impact Measurement Scale was much more likely to pick up depression than was the Quality of Well-being. Although it is assumed that both items will capture depression, the Arthritis Impact Measurement Scale item assumes that people experience depression and asks for how much time in the last month they were depressed. The Quality of Well-being item asks about the last 6 days only and imbeds depression within a list of physical symptoms and problems.

Discussion

This paper reviews three different studies on alternative methods for posing the same issues to survey respondents. In all three studies, trained interviewers administered different forms of similar questions, so the interviewer factor was held constant. However, in each study one form of the question was designed for self-administration. On the basis of these studies, some general conclusions might be offered. These include:

1. Interviewer-administered questions typically detect higher rates of dysfunction. There is reason to believe that these higher rates are indeed true rates of dysfunction.
2. Although correlations between self-administered and interviewer-administered questionnaires may be high, these high correlations are dominated by variability in dysfunction within the population. The issue of sensitivity is often overlooked. Highly sensitive instruments are required to capture minor variation within specific subpopulations.
3. Embedding mental health symptoms, such as depression, within the context of physical health questions may lead to underreporting. This issue needs further study.
4. The consequences of failing to have adequate sensitivity are that health status is overestimated for a population. A related problem relevant to clinical trials is that side effects of treatments are often overlooked. In fact, there may be incentives in some trials to ignore adverse drug effects. This can be accomplished most easily by using insensitive measures of health outcome.

Establishment of a laboratory for methodological studies in health-status assessment is just beginning. These studies are very preliminary. They have small sample sizes with insufficient statistical power to answer many questions. However, this is a promising line of research that will ultimately produce more valid and reliable measures of health status and health-related quality of life. These measures may have significant ben efits for health services research, policy analysis, and assessment of outcomes in clinical trials.

References

Anderson, J. P., Bush, J. W., & Berry, C. C. (1986). Classifying function for health outcome and quality-of-life evaluation: Self-versus interviewer modes. *Medical Care,* 24, 454.

Anderson J. P., Bush J. W., & Berry C. C. (1988). Internal consistency analysis: A method for studying the accuracy of function assessment for health outcome and quality-of-life evaluation. *Journal of Epidemiology,* 41, 127.

Cannell, C. F., Marquis, K. H., & Laurent, A. (1977). The summary of studies on interviewing methodology. *Vital and Health Statistics* (Series 2, No. 69, DHEW Publication RHA77-1343). Washington, DC: U.S. Government Printing Office.

Dawson, D. A., & Adams, P. F. (1987). Current estimates from the National Health Interview Survey, United States, 1986. *Vital and Health Statistics,* Series 10. Hyattsville, MD. National Center for Health Statistics.

Eakin, E., Kaplan, R. M., & Ganiats, T. G. (1989). Methods for inquiring about symptoms in the Quality of Well-being scale. Unpublished manuscript, University of California, San Diego.

Erickson, P, Anderson, J. P., Kendall, E. A., & associates. (1988). Using retrospective data for measuring quality of life: National Health Interview Survey Data and the Quality of Well-being Scale. *Quality of Life and Cardiovascular Care,* 4, 179-184.

Erickson, P., Kendall, E. A., Anderson, J. P., & associates. (1989). Using composite health status measures to assess the nation's health. *Medical Care,* 27, 566-576.

Fries, J. F., Green, L. W., & Levine, S. (1989). Health promotion and the compression of morbidity. *Lancet* II, 481-483.

Kaplan, R. M., & Anderson, J. P. (1988). A general health policy model: Update and applications. *Health Services Research,* 23, 204-235.

Reynolds, W. J., Rushing, W. A., & Miles, D. L. (1974). The validation of a function status index. *Journal of Health and Social Behavior,* 15, 271.

Stewart, A. L., Ware, J. E., Jr., & Brook, R. H. (1981). Advances in the measurement of functional status: Construction of aggregate indexes. *Medical Care,* 19, 473.

FEATURE PAPER

Health Index Validation Through Convergence of Alternative Index Construction Approaches

Larry A. Hembroff, Susan C. Zonia, and Harry Perlstadt

Introduction

This paper describes the development of a variety of health indices for a mailed survey being used to evaluate a large university-based health promotion program. Two different, approaches were taken independently in constructing the indices. One approach is based initially on assumed conceptually meaningful groupings of questionnaire items, hereafter referred to as the "C-based" approach (conceptual). The other is based initially on *ad hoc* empirically based groupings of items, hereafter referred to as "M-based" (empirical). In some instances, the results of the two approaches converge on the same or very similar indices, whereas in some cases they do not. The convergence seems to represent additional evidence regarding index validity.

In doing this analysis, something about health and American culture was discovered that would have been obscured had only one of the approaches in constructing the indices been followed. The discovery concerns perceived health status and the clarity of information regarding the constellation of attitudes and behaviors that lead to that health status. This discovery would suggest new strategies for promoting health.

Background

These indices were developed as part of the evaluation of Healthy U, a multifaceted health promotion program targeting students, faculty, staff, and retirees at Michigan State University (MSU) and funded in part by the W. K. Kellogg Foundation. The program is currently beginning the last year of its 3-year demonstration phase.

Larry A. Hembroff, Susan C. Zonia, and Harry Perlstadt are with the Department of Sociology, Michigan State University, East Lansing.

This research was funded by a grant from the Kellogg Foundation in support of the Healthy U Program at Michigan State University.

The goals of Healthy U are to increase the health knowledge, and change the attitudes and behaviors of the MSU community in five areas: exercise and fitness, stress, nutrition, substance abuse, and safety. The evaluation plan called for a baseline survey to be followed by annual surveys over the next three years.

The baseline survey was conducted in 1986. The second survey was conducted in early 1988 after the 1st year of the 3-year demonstration phase. Data collection for the third survey has recently been completed. The analysis reported here is based on the responses of the 1,830 persons who responded to the first survey and the 1,515 who responded to the second.

The questionnaire was a 13-page booklet containing approximately 200 items. Most of the health-related status, attitude, and behavior questions were written with either 5-point Likert scale or 7-point behavior frequency response options. Several of the items on key health issues were borrowed from other health surveys, such as the Health and Nutrition Examination Survey and the Survey of Personal Health Practices and Consequences. However, the number of such items that could be used was limited and many of these required some modification for the Healthy U population. Still other issues of concern to Healthy U were not addressed in other surveys, requiring many items to be constructed specifically for the Healthy U surveys. Assessing the validity and reliability of the indices became even more paramount in this context.

Construction of Indices

Our approach to testing the validity of the indices involves amassing a preponderance of evidence based on face, concurrent, and construct validity methods (Bohrnstedt, 1983; Carmines & Zeller, 1979). In constructing the various health indices, we have followed two different approaches. The C-based index construction started with an analysis of items in the 1986 baseline

Table 1. Rotated factors and factor loadings for Healthy U exercise and fitness items, 1986 (N = 1832)

Questionnaire item	Factors		
	Affective attitude	Purposive exercise	Integrative exercise
Regular exercise is extremely important	0.711	0.088	0.037
Exercise is generally a waste of valuable time	.689	.118	−.029
If you eat a highly nutritious and low-fat diet, you don't need to exercise much.	.661	−.014	.046
Engage in physically demanding individual sports (tennis, skiing, etc.)	.115	.789	.090
Engage in at least 20 minutes of jogging, swimming, cross-country skiing, or aerobic dance	.249	.719	.137
Engage in physically demanding team sports (basketball, volleyball, etc.)	−.074	.712	−.048
Practice muscle strengthening exercises	.227	.705	.133
Do flexibility or stretching exercises	.341	.598	.228
Lift or move heavy objects or materials	−.128	.539	.067
Walk briskly 20 minutes or more	−.071	.210	.716
Exercise is not effective unless it is done until it hurts	.184	−.387	.587
Use stairs instead of an elevator	.024	.303	.533
Eigenvalue	1.529	3.738	1.094
Percent variance explained	12.7	28.6	9.1

Healthy U survey. Subsequently, the identical items in the 1988 Healthy U survey were subjected to the same analysis to determine whether the initial results would be replicated. The M-based index construction started with the 1988 survey and was checked against the C-based results.

The C-Based Approach

First, all items intended to measure a particular aspect of health, such as fitness or stress were grouped into sets based on their apparent face validity as individual items and as sets. Items comparable in wording to items included in the Michigan Behavior Risk Factor Survey (MBRFS) were examined further. For these the distributions of responses on the Healthy U survey items were compared to the corresponding distributions on similar items in the MBRFS. After adjusting for differences in the demographic compositions of the samples, the distributions seemed very similar on all items compared between the two studies. Since the results of one measure of a variable are, in a sense, being correlated with the results of another measure of the same variable (although in different samples), this comparison represents a form of concurrent validity (Bohrnstedt, 1983).

The assessment of *construct validity* of the indices was based on factor analysis of the items. Each of the sets of items was factor analyzed using a principal components method with a varimax rotation. The central concern was whether items in each set covaried sufficiently to be determined by a common latent construct, presumably the concept of which they were intended indicators (Kim & Mueller, 1978a).

In most cases the factor analysis identified two or more subscales within the initial sets. For example, the exercise and fitness items seemed to represent three latent factors. One subscale was comprised of only attitude items—an affective dimension of fitness; a second subscale was comprised of participation frequency items in various kinds of purposive exercise; and a third subscale was comprised of participation frequency items representing fitness activities that can be integrated into other daily routines, such as walking from place to place

Table 2. Summary of numbers of items, percent variance explained, and reliabilities for factors generated, by health topic: 1986

Health topic Factor	Numbers of items	Percent variance explained	Alpha coefficient
Exercise and Fitness	10	50.4	0.76
Affective attitude	3	12.7	.52
Purposive exercise	5	28.6	.81
Integrative exercise	3	9.1	.23
Stress	7	59.2	.79
Status-cognitive	3	44.4	.75
Symptom frequency	4	14.8	.68
Substance Abuse	7	48.1	.54
Illegal addictive use	3	31.5	.44
Legal habitual use	4	16.6	.47
Alcohol Use	8	52.7	.73
Social drinking	6	38.3	.73
Problem drinking	3	14.4	.65
Smoking	3	78.1	.85
Safety	13	54.8	.68
Status-orientation	4	10.7	.50
Household safety	3	20.3	.45
Walking safety	2	9.2	.57
Auto safety	2	7.4	.56
Product use safety	2	7.2	.38
Nutrition	9		.44
Attitudes	3	43.7	.34
Behaviors	6	45.7	.39
Proactive behaviors	4	27.7	.32
Avoidance behaviors	2	18.0	.33

Table 3. Correlations among various Healthy U indices: 1986 (N = 1832)

	Substance abuse	Alcohol use	Safety	Nutrition	Exercise	BMI
Stress	0.216*	0.212*	0.246*	0.108*	−0.103*	0.021
Substance abuse		.646*	.348*	.226*	−.052	−.050
Alcohol use			.383*	.147*	−.112*	−.079*
Safety				.288*	−.236*	−.030
Nutrition					.248*	.085*
Exercise						.177*

*$p(r) < .001$
NOTE: Correlations are not corrected for attenuation

or using stairs rather than elevators. Table 1 illustrates the items and presents the results of this analysis.

Similarly, the analysis extracted two factors from the stress-related items—one a stress status factor and the other a symptom frequency factor. And the substance abuse items loaded on two separate factors, one the frequency of using illegal or addictive drugs and the other the frequency of using legal but unhealthy substances. Table 2 summarizes the results of the factor analyses for all the various sets of items. The analysis provided evidence of each index's construct validity.

The indices and subscales were tested for reliability using Cronbach's alpha. Some items were deleted if doing so improved the index's reliability. The resulting reliability coefficients are also presented in Table 2. In most instances, the reliabilities were at least marginally acceptable by conventional standards. Some indices seemed to have unacceptably low alpha coefficients. The nutrition indices had particularly low alphas.

Once the set of items to constitute each of the various indices was determined, unweighted, summated scores were calculated on each of the indices and subscales for the entire sample with adjustments made for missing data. That is, factor-based scores were calculated rather than factor scores (Kim and Mueller, 1978b). In each case, high index scores represented unhealthy responses to the items and low scores represented healthy responses. The distributions of scores on each of the indices were intuitively reasonable, even within and across the three primary target populations.

To assess further the concurrent validity of the indices, we examined the correlation of each index with the other health indices using Pearson's *r*. Almost without exception the indices correlated to each other in theoretically predictable ways. For example, we expected that individuals who were under much stress would be more likely in trying to cope with the stress to be involved in activities regarded as stress reducing, some of which are healthy such as exercising and some of which are unhealthy such as drinking.

The analyses showed the indices did correlate as expected. For example, as Table 3 indicates, stress was directly related with substance abuse generally and alcohol use in particular, while also being inversely correlated with exercise. Individuals who tended to be concerned about safety were less likely to abuse substances, more likely to have better nutrition, and less likely to be involved in vigorous physical exercise. Even the nutrition indices correlated as expected with respect to exercise and fitness, stress, and Body-Mass Index (BMI). That is, individuals who had unhealthy nutrition behaviors were less likely to have appropriate height-weight combination scores.

Although the correlations among the indices were weak to moderate, nearly all predicted correlations were statistically significant at the 0.001 level, and persisted with only slight weakening even after the effect of the different population groupings was partialed out.

As a final step in the C-based method, we tried to replicate these analyses with the data collected in the 1988 Healthy U survey. The attempt to replicate had to take into account modifications made between the 1986 and the 1988 versions of the questionnaire. Some new items had been added, some items had been deleted, and some items or their responses were revised. Consequently, not all the same items for each of the indices were available in the 1988 data set.

All the items in 1988 corresponding to 1986 items were grouped and coded as before. Each set of items corresponding to an index constructed previously was factor analyzed to see if the analysis would be replicated with the same factors being extracted, the same items loading on the same factors, and the factor loading of each item on its factor being nearly the same as before. The results were then compared to the reanalyzed set of only those 1986 items also appearing in the 1988 questionnaire.

The results were remarkably consistent. Table 4 summarizes the results of these comparisons. In all cases except one, the same number of factors was extracted, and, of the 47 items being analyzed, 85.1 percent loaded on the same factors across the two samples. Six of the 14 indices were replicated exactly with very similar factor loadings of items on factors.

Similarly, the test of each index for reliability resulted in nearly identical alpha coefficients for each index across time and across the two samples. This was true not only of those indices with high reliability coefficients, but those with low alpha's as well.

To this point the approach to establishing the validity of the indices has been quite conventional. Evidence from multiple means of assessment indicated the indices were valid although their reliabilities seemed problematic.

Existing rules of thumb would have called for dropping those items and indices with apparent validity but low overall reliability, but because we were committed

Table 4. Comparisons of reliability coefficients for indices in 1986 and 1988 samples using only items common to both

Index Factor	Number of items	Alpha 1986	Alpha 1988
Exercise and Fitness	7	0.704	0.711
Affective and attitude	0	—	—
Purposive exercise	5	.815	.809
Integrative exercise	2	.086	.067
Stress	6	.757	.537
Status-cognitive	5	.756	.534
Symptom frequency	1	—	—
Substance abuse	7	.544	.528
Illegal and addidictive use	4	.453	.401
Legal/Habitual Use	3	.467	.436
Alcohol Use	6	.692	.688
Social Drinking	4	.647	.671
Problem Drinking	3	.549	.509
Smoking	3	.849	.806
Safety	11	.620	.628
Status-orientation	5	.556	.550
Household safety	3	.452	.627
Walking safety	3	.342	.410
Auto safety	0	—	—
Product use safety	0	—	—
Nutrition	9	.438	.453
Attitudes	3	.336	.340
Proactive behaviors	4	.316	.303
Avoidance behaviors	2	.333	.319

to evaluating the program we felt obligated to retain them. The results using the M-based approach on the follow-up survey suggest that ignoring these indices in subsequent analyses because they lacked reliability would have been a mistake.

The M-Based Approach

The M-based approach began with the 1988 Healthy U survey data. Whereas the C-based approach formed groupings of items based on conceptual grounds then refined the indices based on empirical considerations, the M-based approach grouped items initially based purely on empirical grounds then filtered groupings based on conceptual considerations.

The first step in the M-based approach involved gathering all perceived health status, attitude, and behavior items for all of the health areas and entering them simultaneously into one large exploratory factor analysis. A principal components analysis with a varimax rotation was conducted for the set of all items. Since factor analysis extracts factors based on the common covariances in the inter-item correlations, it is quite possible in this circumstance that highly correlated measures of different health dimensions could load together on a single factor. Thus, a set of items grouped based on factor analytic criteria need not represent a single valid underlying concept.

The factor analysis sorted out a number of clusters of items, all of which were examined in detail. If, based on prima facie grounds, all items in the cluster seemed to be measuring the same underlying health concern, the cluster was retained for construction as an index. If

Table 5. Convergence of indices based on C-based and M-based approaches

C-based indices			M-based indices		
Health area	*N* of items	Number of common items	*N* of items	Alpha	Health area
Exercise and Fitness	7				
Purposive exercise	5	4	6	0.81	Fitness
Integrative exercise	2				
Stress	6				
Status-cognitive	2	2	5	.81	Stress
Symptom frequency	4		5	.73	Relaxation
Substance abuse	7				
Illegal addictive use	4				
Legal habitual use	3				
Alcohol use	6				
Social drinking	4	1 2	4	.78	Drinking
Problem drinking	4	1			
Smoking	3				
Safety	11				
Status-orientation	3		5	.53	Safety
Household safety	3	3 5			
Walking safety	2	2			
Product use safety	2				
Nutrition	9				
Attitudes	3		4	.41	Nutrition
Proactive behaviors	4	2 4			
Avoidance behaviors	2	2			

some items in a cluster seemed to measure one health concern while others measured another, the cluster was dropped from subsequent analysis.

Since the C-based index construction activity did not attempt to construct all possible indices from all the items on the questionnaire, the M-based approach produced several factors not previously identified. Nevertheless, most of those produced, which seemed on their face to be measures of singular health concerns, represented the same health dimensions as set out in the C-based approach. The common health dimensions included exercise and fitness behaviors, safety, alcohol use, stress, and nutrition. Another factor extracted but which had not been constructed in the previous approach concerned relaxation behaviors. These 6 indices are summarized in Table 5.

Each of these M-based indices was tested for reliability. As Table 5 indicates, four of the six indices had reliability coefficients adequately high by conventional standards. The reliability coefficients for the household safety and nutrition behavior indices were again "unacceptably low."

Convergence of Methods on Indices

The two methods, C-based and M-based, generated similar health indices for most of the health concerns of interest in the project. Since one method constructs indices on conceptual grounds while the other does so on empirical grounds, we believe that when the two approaches converge on a single index, the evidence regarding validity is enhanced. Table 5 summarizes the comparisons between resulting indices.

The exercise and fitness index formed by the M-based method contained six items, four of which were identical to the items in the proactive fitness index generated by the C-based method. The other two items concerned perceived fitness status and had not been included in the C-based analysis.

The five items in the M-based safety index were the same as the five items comprising the C-based household and walking safety subscales. That is, the M-based index was identical to the combination of two C-based indices.

The four consumption frequency items forming the M-based nutrition index were the same as four of the six items comprising the two C-based nutrition behavior indices. One of those had focused on seeking out healthy foods while the other focused on avoiding unhealthy foods. Two items from each of these formed the M-based index.

Of the four items forming the M-based alcohol use index, two items also appeared in the two C-based alcohol indices. The other items in the M-based index did not appear in the 1986 questionnaire at all.

Finally, of the five items in the M-based stress index, two were also among the three comprising the C-based stress status index. The other three in the M-based stress index had been newly added to the 1988 questionnaire.

Summary

Convergence on measures of the validity and identification of items using two alternative index construction approaches was extensive. The preponderance of evidence is consistent and persuasive that the indices are valid measures of the health concerns they were intended to represent.

- The items in each index seemed to be valid indicators of their respective health concepts based on face validity criteria.
- The items produce similar results as other measures of the same health concepts (a form of concurrent validity).
- As sets of indicators, the items correlate sufficiently to produce a number of unidimensional underlying factors (construct validity).
- Within sets, the items tend to be more or less reliable predictors of other items in the same set.
- Indices tend to correlate with other health indices in theoretically predictable ways (another form of concurrent validity).
- And, two very different approaches to identifying indices, one driven by conceptual considerations (C-based) and the other driven by empirical considerations (M-based), produced very similar sets of items for each health index.

Discussion

Throughout the analysis, the evidence for validity seemed quite clear for each of the indices, and yet the reliabilities, although strikingly consistent across time, seemed undesirably low for some of the indices. Although it is possible to dismiss those indices as being inadequate measures, other projects, such as the Health and Nutrition Examination Survey, encountered the same problem with measures of the same areas of health, nutrition in particular. It seems that the results of the M-based index construction approach may suggest an explanation for this otherwise frustrating outcome. In this case perhaps the reliability coefficients mean something other than the measurement consistency of items.

In the results of the M-based analysis, some of the clusters of items, which as indices had high alpha coefficients, also included perceived health status items. For example, six items loaded together on the fitness and exercise factor had an alpha coefficient of 0.81. Of the 6 items, 2 were intended to measure the respondent's subjective assessment of fitness status. Similarly, of the three items in the smoking index (C-based), one item was essentially a smoking status item (alpha = 0.85 and 0.81 in 1986 and 1988, respectively).

In each of these cases the other items in each set were behavior items. But the status and behavior items in each instance were sufficiently intercorrelated as to appear to be measures of the same thing, as if the link between a particular behavior pattern and a health status was strong enough to define it. That is, the individuals

seemed to perceive that by their behavior their status on the issue is determined.

These two health areas, smoking and exercise, seem to represent areas in which the information and messages in the media have become remarkably consistent as to which behaviors are healthy and which not. It seems reasonable that in such a context, not only would individuals easily attribute a particular health status to themselves based on their behaviors, but also would develop some consistency in the healthfulness of their behaviors. This would explain both the high reliability coefficients and the high intercorrelation of status and behavior items.

Looking across the indices, as the reliabilities decrease the links to self-perceived health status items disappear, as in the case of the drinking index. Although the somewhat lower reliability of the drinking index was still adequate, it was comprised of behavioral items but no status items. That is, while there is consistency among the measures of individuals' drinking behaviors—or perhaps their behaviors are consistent—their behaviors are not highly correlated with a health status measure. This suggests that individuals, although knowing what their behaviors are, are not certain as to how much drinking or what kind of drinking represents unhealthy behavior.

It also seems that the use of alcohol is an area in which the messages in the culture and the media are less consistent. Individuals are encouraged to consume alcohol by advertisers, are provided ample instances where consumption is normal and legitimate, while also being told of alcohol's link to a host of serious health and social problems in the case of excess consumption. But excess and problem drinking patterns are not clearly defined. In this context, it seems reasonable that individuals might develop consistent patterns of behavior but not correlate them with a particular health status, hence the fairly high reliability coefficient but no correlated status items in the index.

Still other indices such as nutrition had low reliability—that is, low consistency of responses across behavioral items—and were unconnected to perceived health status. This may be one area where the messages and information in the culture and the media are most contradictory and confusing. Not only are individuals being directed in their nutrition behaviors by fads, cultural heritage, and advertisers but also by nutrition research. Many of these directly compete with one another as to what individuals should be consuming. Even research reports seem frequently to contradict each other as to which nutrition behaviors are healthy and which are not. In this context, it is not at all surprising that individuals not only do not correlate perceived health status with their nutrition behaviors but that they are not even consistent in their behaviors.

In the case of these indices, the reliability coefficients may not really tell us much about the adequacy of the measures. Rather, it is possible that the reliability coefficients are telling us something about the inconsistencies in people's behaviors and, more importantly, the inconsistencies in the health information Americans are being confronted with.

Clearly, such an interpretation has implications as to how health promotion strategists in areas such as nutrition might efficaciously proceed. It would suggest that important advances in health promotion might be possible if the media, cultural, legal, and educational messages regarding all health-related behaviors could work in concert rather than in competition. In this case, the M-based approach may have added not only an additional bit of evidence regarding index validity, but additional insight into the nature of the health promotion problems as well.

References

Bohrnstedt, G. W. (1983). Measurement. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.). *Handbook of survey research.* (pp. 69-121). New York: Academic Press.

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills: Sage Publications.

Kim, J., & Mueller, C. W. (1978a). *Introduction to factor analysis: What it is and how to do it.* Beverly Hills: Sage Publications.

Kim, J., and Mueller, C. W. (1978b). *Factor analysis: Statistical methods and practical issues.* Beverly Hills: Sage Publications.

FEATURE PAPER

Validating Questions Against Clinical Evaluations: A Recent Example Using Diagnostic Interview Schedule-Based and Other Measures of Post-traumatic Stress Disorder

Richard A. Kulka, William E. Schlenger,
John A. Fairbank, B. Kathleen Jordan, Richard L. Hough,
Charles R. Marmar, and Daniel S. Weiss

Introduction

With the recent emergence (or re-emergence) of several new strategies for the evaluation of survey questions, it is well to remind ourselves of the strengths and limitations of somewhat more "classic" or traditional approaches to this issue, especially as new lessons are learned from ongoing research on health and mental health. Perhaps the most classic of such approaches are efforts to validate survey measures against external criteria, including medical or administrative records, the results of a physical examination, and so forth. In the field of psychosocial epidemiology the usual approach to validity testing is to correlate diagnostic classifications derived from survey measures with clinical judgment (Weissman & associates, 1986). This tradition became well-established during World War II with the development of a psychosomatic "impairment" scale as the core of the Neuropsychiatry Screening Adjunct, whereby the method of seeing whether an impairment scale differentiated between a "known-ill" and a "known-well" group became the basic procedure for validating psychiatric impairment scales. Although Robins and Guze (1970), Robins (1985), Helzer and co-workers (1987), and Kraemer and associates, (1987), among others have emphasized the danger of relying predominantly on this single strategy of assessing validity, and have suggested additional methods for establishing diagnostic validity on a scientific basis, validation of research diagnostic information against clinical judgment or evaluation remains the basic norm in psychiatric epidemiology.

Richard A. Kulka is with National Opinion Research Center, University of Chicago. William E. Schlenger, John A. Fairbank, and B. Kathleen Jordan are with Research Triangle Institute, Research Triangle Park, North Carolina. Richard L. Hough is with San Diego State University. Charles R. Marmar and Daniel S. Weiss are with Langley Porter Psychiatric Institute, University of California, San Francisco.

This research was supported by the Veteran's Administration, Contract No. V101 (93) P-1040.

Evolution of a Diagnostic Interview Schedule-Type Measure of Post-traumatic Stress Disorder for the National Vietnam Veterans Readjustment Study

Such was certainly the case in the fall of 1984 when Research Triangle Institute was awarded a contract by the Veterans Administration (VA) to conduct the National Vietnam Veterans Readjustment Study (NVVRS)—a nationwide psychosocial epidemiologic survey of sufficient size, scope, and rigor to establish "the prevalence and incidence of post-traumatic stress disorder (PTSD) and other psychological problems in readjusting to civilian life" among Vietnam veterans—as mandated by the U.S. Congress in Public Law 98-160 (Kulka & associates, 1988). Before the initiation of the NVVRS, no previous research had been completed that would support the derivation of population-based, diagnostic estimates of the prevalence of PTSD among Vietnam veterans. In part, this was due to the absence of any official diagnostic criteria for PTSD before the publication of the third edition of the Diagnostic and Statistical Manual (DSM-III) by the American Psychiatric Association (1980). Thus, previous research was based largely on limited samples, and even the three major surveys conducted during this period that involved broader and more representative samples of Vietnam veterans (Card, 1983; Egendorf & associates, 1981; Fischer & others, 1980) were hampered by an inability to link measures of "mental or emotional problems" or "stress" to the diagnostic criteria of PTSD.

One of the first advances in this domain was the development of a questionnaire module to assess PTSD for the Diagnostic Interview Schedule (DIS), a highly structured survey interview instrument designed explicitly for use by lay interviewers, that is nonclinicians. The Diagnostic Interview Schedule was originally developed by Washington University in St. Louis (Robins & associates, 1981; Robins & associates, 1985; Robins, 1986; Helzer & Robins, 1988) under the auspices of the Na-

tional Institute of Mental Health for use in a landmark study of the mental health of community and institutionalized adults in the United States—the five-site Epidemiologic Catchment Area (ECA) project (Regier & associates, 1984; Eaton & Kessler, 1985; Eaton & associates, 1986). The Diagnostic Interview Schedule is comprised of multiple questionnaire modules, each designed to detect the presence of a different psychiatric disorder. The PTSD module was developed only after the ECA studies were underway and was used only during the second wave of interviewing. Slightly different versions of this module were used by the St. Louis and North Carolina sites, and the Los Angeles ECA site employed a substantially different version from the other two. Moreover, because it was added at a later date, the PTSD module was not included in the validation studies of the other DIS modules, the results of which proved to be somewhat inconsistent and controversial (Robins & associates, 1982; Anthony & co-workers, 1985; Helzer & associates, 1985; Robins, 1985; Helzer & others, 1987; Burvill, 1987).

Nevertheless, because the Diagnostic Interview Schedule was widely regarded as the state of the art for the assessment of psychiatric disorder and derivation of estimates of incidence and prevalence of diagnoses from community-based survey data, the NVVRS study team believed it vital to include a version of the Diagnostic Interview Schedule in the survey, including particularly a PTSD module. The choice of the particular DIS PTSD module was not straightforward, however, for several reasons. First, as noted earlier, there were several versions of this module at that time, and none had yet been validated. Second, the diagnostic criteria for post-traumatic stress disorder were in transition from DSM-III to DSM-III-R (American Psychiatric Association, 1987). To grapple with this choice the research team consulted with a nationally recognized panel of expert clinicians to develop detailed guidelines for a new module for the diagnostic assessment of post-traumatic stress disorder, with a style and format consistent with other DIS modules. This new "DIS-type" measure was designed to assess symptoms of post-traumatic stress disorder using either DSM-III or DSM-III-R criteria, as well as addressing concerns raised by these experts about versions of the post-traumatic stress disorder module of the Diagnostic Interview Schedule in use at that time (including the existing St. Louis and North Carolina versions). The resulting DIS-type PTSD measure was then included as one of several measures in the NVVRS Preliminary Validation Study

Preliminary Validation Study

Because at the time the National Vietnam Veterans Readjustment Study was initiated there was no evidence in the literature for the convergent or empirical validity of any of the measures available for a survey-based assessment of post-traumatic stress disorder (including the existing DIS PTSD modules), an integral part of the study design was the completion of a Preliminary Validation Study, the purpose of which was to determine whether the available survey-based measures could distinguish true PTSD cases from noncases in a clinical (helpseeking) population. This was seen as a necessary though not sufficient condition for establishing the validity of the NVVRS PTSD assessment, and represented the first step in a two-step validation procedure.

Candidate PTSD measures were administered to 225 Vietnam veterans whose mental health status with regard to post-traumatic stress disorder and other psychiatric disorders had already been determined clinically. Subjects were in treatment for PTSD or for other psychiatric disorders at VA Medical Centers or Vet Centers at eight sites across the country. The clinical diagnoses were "double determined": subjects were accepted into the validation study only if their chart diagnosis agreed with the diagnosis made by an independent expert clinician on the basis of a structured clinical interview conducted blind to the chart diagnosis.

Subjects certified for inclusion underwent a 5-hour interview by an experienced (nonclinical) survey re-

Table 1. Relative diagnostic accuracy of PTSD measures

Measure	Percent correctly classified[a]	Kappa[b]	Sensitivity[c]	Specificity[d]
M-PTSD scale	88.9	0.753	94.0	79.7
D-PTSD scale (sum of positive items)	87.5	.714	95.5	72.6
PTSD checklist	84.9	.672	88.3	78.9
D-PTSD scale (using DSM-III-R rules)	83.5	.639	87.2	72.6
MMPI (Fairbank-Keane scale)	81.5	.605	90.1	68.8
Impact of event scale	81.6	.565	91.7	61.8

[a] The percent of the entire sample (true cases and true noncases) that are correctly classified by the survey measure.
[b] A measure of the extent of agreement between two assessments corrected for the effects of chance. (Kappas above .75 are considered to indicate excellent agreement, those between .40 and .75 fair-to-good agreement, and those below .40 poor agreement.)
[c] The percent of true cases that are classified as cases by the survey measure.
[d] The percent of true noncases that are classified as noncases by the survey measure.

search interviewer, which included five measures aimed at identifying post-traumatic stress disorder: (1) the modified DIS-type PTSD module (D-PTSD); (2) a checklist of PTSD symptoms analogous to conventional mental health impairment scales (Weissman & associates, 1986); (3) the Mississippi Scale for Combat-Related Post-traumatic Stress Disorder (M-PTSD) (Keane & associates, 1988); (4) the Impact of Event Scale (IES) (Horowitz & associates, 1979); and (5) Form AX of the Minnesota Multiphasic Personality Inventory (MMPI), which provided the Keane-Fairbank PTSD scale (Keane & associates, 1984).

Table 1 shows the classification of results for the five candidate measures included in the study, using the final clinical diagnosis as the criterion. The predictive validity of each of these measures was assessed in terms of: (1) the percent of subjects correctly classified as having or not having post-traumatic stress disorder; (2) the Kappa statistic (Cohen, 1960); (3) sensitivity, or the percent of true cases classified as positive by the survey measure; and (4) specificity, or the percent of true noncases classified as negative by the survey measure. As indicated, all of the measures performed reasonably well, with the M-PTSD scale and the D-PTSD scale providing the best prediction of the certified clinical diagnosis, although the latter performed somewhat better scored as a sum of symptoms endorsed as positive than when scored in the conventional manner of the Diagnostic Interview Schedule, that is, according to DSM-III criteria. Based on these results, the M-PTSD and D-PTSD scales were both included in the survey interview for the national study. In turn, both the Minnesota Multiphasic Personality Inventory and Impact of Event Scale were included as part of the Clinical Interview Component of the NVVRS.

Clinical Interview Component

The Clinical Interview Component represented the second stage of a two-stage strategy for validating the NVVRS PTSD diagnosis. By documenting the ability of the candidate survey interview instruments to identify true cases of post-traumatic stress disorder, the validation study provided a scientific basis for selecting the PTSD instruments to be used in the national survey component of the National Vietnam Veterans Readjustment Study. However, demonstration that the instruments could distinguish true PTSD cases from noncases in a clinical setting was a necessary but not sufficient condition for establishing the validity of the measures in the community—that is, the general population of Vietnam veterans who are predominantly nontreatment seeking. The second step in the validation process was conducted subsequent to the survey interview and designed to provide information about the ability of the instruments to distinguish PTSD cases from noncases in the general population of Vietnam veterans.

For the Clinical Interview Component, a subset of over 340 Vietnam veterans was selected to undergo a

Table 2. Post-traumatic stress disorder indicators available for clinical subsample respondents

Name	Description	Type	Source
M-PTSD	Mississippi Combat-Related PTSD scale	booklet self-report	survey interview
MMPI-PTSD	MMPI PTSD scale (Fairbank-Keane Scale)	booklet self-report	clinical interview
SCIDX	PTSD diagnosis from the SCID interview	clinician judgment based on self-report	clinical interview
SXCTCURR	number of PTSD symptoms reported as having occurred within the past 6 months	interview self-report	survey interview
SRRS-INT	intrusion subscale of the Stress Response Rating Scale—assesses the presence of signs and symptoms of intrusive thoughts	clinician judgment based on observation	clinical interview
SRRS-AVD	avoidance subscale of the Stress Response Rating Scale—assesses the presence of signs and symptoms of avoidance	clinician judgment based on observation	clinical interview
SRRS-REA	reactivity subscale of the Stress Response Rating Scale—assesses the presence of signs and symptoms of psychological reactivity	clinician judgment based on observation	clinical interview
IES-INT	intrusion subscale of the Impact of Event Scale—assesses the presence of signs and symptoms of intrusive imagery during R's self-selected worst period	booklet self-report	clinical interview
IES-AVD	avoidance subscale of the Impact of Event Scale—assesses the presence of signs and symptoms of avoidance during R's self-reported worst period	booklet self-report	clinical interview
ASSES-SC	Global Assessment Scale—assesses overall level of psychosocial functioning	clinician judgment based on observation	clinical interview

followup clinical interview with an expert mental health clinician. The clinical interview sample was drawn from among respondents who lived within "reasonable commuting distance" of 28 specific geographic areas selected to cover as much of the national sample of veterans as possible. The sample included all those who appeared on the basis of their survey interview to be PTSD positive, and a sample of those who appeared to be PTSD negative. Clinical interviews were conducted by a total of 29 mental health clinicians in the 28 geographic areas.

The clinical interview assessment included a semi-structured diagnostic interview based on Form NP-V of the Structured Clinical Interview for DSM-III-R (SCID); (Spitzer & associates, 1987), as well as the Minnesota Multiphasic Personality Inventory, the Impact of Event Scale, and some other clinical ratings. The full range of post-traumatic stress disorder indicators available for clinical subsample respondents is shown in Table 2.

The initial assessment of these data involved a comparison of the relative predictive validity of the two survey-based measures carried forward from the validation study, as well as the MMPI subscale, using the main clinical interview or SCID diagnosis as the criterion. Table 3 provides a comparison of the specificity, sensitivity, and agreement of these three PTSD measures with the clinical diagnosis. A comparison with Table 1 indicates that the sensitivity and overall levels of chance-adjusted agreement declined for all three measures by comparison with the preliminary validation study. Such a decline has quite typically been observed in the validation of psychiatric impairment measures in moving from clinical to community-based samples. However, while levels of agreement for both the M-PTSD and MMPI subscale remained in the acceptable range (see notes for Table 1), this was not the case for the D-PTSD scale—the NVVRS DIS-type measure—which did not do well in distinguishing cases from noncases in the clinical followup subsample. In contrast to its sensitivity of 87.2, specificity of 72.6 and Kappa of 0.639 in the preliminary validation study, it exhibited a sensitivity of only 21.5, specificity of 97.9, and Kappa of 0.256 in the clinical followup. Thus, although this measure was still quite successful in correctly identifying noncases, it was able to identify only about one in five cases of post-traumatic stress disorder as diagnosed by expert clinicians. As a result, the modified DIS-PTSD module developed for the NVVRS was not used in the procedure developed to identify cases of post-traumatic stress disorder for the national prevalence estimates.

Table 3. A comparison of the specificity, sensitivity and agreement of three types of PTSD diagnoses with a DSM-III-R clinical (SCID) diagnosis

	With the SCID Diagnosis		
Measure	Sensitivity	Specificity	Kappa
M-PTSD scale	77.3	82.8	0.528
D-PTSD scale (using DSM-III-R rule)	21.5	97.9	.256
MMPI (Fairbank-Keane scale)	71.6	82.4	.478

Post-traumatic Stress Disorder Case Determination

The ultimate failure of the D-PTSD to distinguish cases from noncases of post-traumatic stress disorder, as well as the relative decline in accuracy of predication of the other measures, served to confirm our initial assumptions that no single post-traumatic stress disorder assessment is completely error free. Therefore, instead of relying on a single PTSD diagnostic indicator, PTSD diagnoses in the National Vietnam Veterans Readjustment Study were made on the basis of information from multiple indicators. These composite diagnoses were made on the basis of a detailed review of all the PTSD information for each clinical subsample subject. Review began by examining the study's three primary indicators: the Mississippi Scale for Combat-Related Post-traumatic Stress Disorder, the clinical interviewer's PTSD diagnosis made on the basis of the SCID interview, and the PTSD subscale of the Minnesota Multiphasic Personality Inventory. When these three indicators agreed, the diagnosis was considered settled (decided). In the event of a discrepancy in PTSD diagnosis among the three indicators, information was used from the study's several other PTSD measures to resolve the discrepancy. Information was combined from these other indicators statistically to create two additional primary indicators for use in resolving discrepancies. As a result, at least three primary indicators concurred in the composite diagnosis for every subject. In fact, for 87 percent of the subjects, four out of five primary indicators agreed on the diagnosis (including 59 percent for which all five agreed), and for the remaining 13 percent three out of five agreed. Application of this procedure resulted in a composite PTSD diagnosis for every subject in the clinical subsample. Composite diagnoses were extended to the full survey sample of 1,630 Vietnam veterans via logistic regression.

Impact on Prevalence Rate Estimates

Based on this composite diagnosis it was estimated that 15.2 percent of Vietnam veterans are current cases of post-traumatic stress disorder (Schlenger & associates, in press). This estimate stands in rather stark contrast to that recently published by The Centers for Disease Control Vietnam Experience Study (1988) based on a random sample of 2,490 Vietnam veterans. Using a slightly modified version of the PTSD module of Version IIIA of the Diagnostic Interview Schedule, the CDC research team estimated that approximately 15 percent of these veterans had experienced combat-related post-traumatic stress disorder at some time during or after their military service, but that the current prevalence of the disorder (during the one month immedi-

ately before the assessment) was only 2.2 percent. While numerous differences between the two studies could have accounted for this sevenfold difference (that is, 15.2 versus 2.2 percent)—including differences in samples, differences in instrumentation and methodology, and other factors—a detailed analysis of these various factors suggested that the discrepancy is primarily the result of differences in the measures used to assess post-traumatic stress disorder (U.S. Congress, 1988). Diagnoses derived from all of the DIS-type algorithms in the NVVRS database produced significantly lower prevalence estimates than the current NVVRS estimate of 15 percent. The low sensitivity exhibited by the DIS-type measures suggests that the lower estimates derived from these Diagnostic Interview Schedule measures are a result of a tendency to miss true cases of the disorder and thereby underestimate true prevalence.

One potential explanation of the difference between these estimates is that the CDC prevalence estimate is lower than the NVVRS estimate because the CDC instrument detected only the most severe cases. One way of assessing this possibility was to examine the relative sensitivity of our DIS-PTSD module in detecting relatively less severe and relatively more severe cases. To test this hypothesis, composite diagnosis PTSD positives in the clinical subsample were split into three groups by level of severity using the Mississippi Scale. In the less severe category, the DIS-type measure identified as PTSD positive only 1 of 42 cases, for a sensitivity of 2.4 percent. In the more severe category, only 3 of 17 cases (sensitivity = 17.6 percent) were identified as PTSD positive; and in the most severe category, only 5 of 17 (sensitivity = 29.4 percent). Thus, although the DIS-type assessment did better at identifying more severe PTSD cases than those that were less severe, even among the most severe cases the method did not detect a large majority of true cases.

Although the evidence is not complete, it implies that the low estimates derived from the CDC Vietnam Experience Study result primarily from its reliance on one instrument, a measure not sufficiently sensitive to detect true PTSD cases in a community population. Moreover, had the National Vietnam Veterans Readjustment Study relied exclusively on the modified DIS-PTSD module as the source of its prevalence estimates, a reliance that would not have appeared unreasonable based on the results of the preliminary validation study alone, it would have seriously underestimated the current prevalence of post-traumatic stress disorder among Vietnam veterans.

This case study underlines in a rather dramatic way the dangers of relying on a single survey measure to establish a psychiatric diagnosis in the general population as well as the use of a single-stage validation procedure based on a measure's ability to distinguish true cases from noncases in a clinical population as sufficient evidence of validity. Had not a second-stage assessment been done to determine whether our instruments could distinguish true cases of post-traumatic stress disorder from noncases in the community, our assessment of the validity of our DIS-type measure would obviously have been seriously flawed.

References

American Psychiatric Association. (1980). *Diagnostic and Statistical Manual of Mental Disorders*. (3rd ed.). Washington, DC: American Psychiatric Association.

American Psychiatric Association. (1987). *Diagnostic and Statistical Manual of Mental Disorders*. (3rd ed. rev.). Washington, DC: American Psychiatric Association.

Anthony, J. C., Folstein, M., Romanoski, A. J. & associates. (1985). Comparison of lay diagnostic interview schedule and a standardized psychiatric diagnosis: experience in eastern Baltimore. *Archives of General Psychiatry*, 42, 667-675.

Burvill, P. W. (1987). An appraisal of the NIMH epidemiologic catchment area program. *Australian and New Zealand Journal of Psychiatry*, 21, 175-184.

Card, J. (1983). *Lives After Vietnam: The Personal Impact of Military Service*. Lexington, MA: Lexington Books.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: psychosocial characteristics. *Journal of the American Medical Association*, 259, 2701-2707.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Eaton, W. W. & Kessler, L. G. (Eds.) (1985). *Epidemiologic Field Methods in Psychiatry: The NIMH Epidemiologic Catchment Area Program*. Orlando, FL: Academic Press, Inc.

Eaton, W. W., Regier, D. A., Locke, B. Z. & associates. (1986). The NIMH epidemiologic catchment area program. In M.M. Weissman, J.K. Meyers, & C.E. Ross (Eds). *Community Surveys of Psychiatric Disorders* (pp. 209-219). New Brunswick, NJ: Rutgers University Press.

Egendorf, A., Kadushin, C., Laufer, R. S. & associates. (1981). *Legacies of Vietnam: Comparative Adjustment of Veterans and Their Peers*. Washington, DC: U.S. Government Printing Office.

Fischer, V., Boyle, J. M., Bucuvalas, M., & associates. (1980). *Myths and Realities: A Study of Attitudes Toward Vietnam Era Veterans*. Washington, DC: U.S. Government Printing Office.

Helzer, J. E., Robins, L. N., McEvoy, L. T., & associates. (1985). A comparison of clinical and DIS diagnoses: physician reexamination of lay interviewed cases in the general population. *Archives of General Psychiatry*, 42, 657-666.

Helzer, J. E., Spitznagel, E. L., McEvoy, L. T. & associates. (1987). The predictive validity of lay diagnostic interview schedule diagnoses in the general population. *Archives of General Psychiatry*, 44, 1069-1077.

Helzer, J. E. & Robins, L. N. (1988). The diagnostic interview schedule: its development, evolution, and use. *Social Psychiatry and Psychiatric Epidemiology*, 23, 6-16.

Horowitz, M. J., Wilner, N., & Alvarez, W. (1979). Impact of event scale: A measure of subjective stress. *Psychological Medicine*, 41, 209-218.

Keane, T. M., Caddell, J. M., & Taylor, K. L. (1988). Mississippi scale for combat-related posttraumatic stress disorder: three studies in reliability and validity. *Journal of Consulting and Clinical Psychology*, 56, 85-90.

Keane, T. M., Malloy, P. F., Fairbank, J. A. (1984). Empirical development of an MMPI subscale for the assessment of combat-related post-traumatic stress disorder. *Journal of Consulting and Clinical Psychology*, 52, 888-891.

Kraemer, H. C., Pruyn, J. P., Gibbons, R. D., & associates. (1987). Methodology in psychiatric research: Report on the 1986 MacArthur Foundation Network I Methodology Institute. *Archives of General Psychiatry*, 44, 1100-1106.

Kulka, R. A., Schlenger, W. E., Fairbank, J. A. & associates. (1988). *Contractual Report of Findings from the National Vietnam Veterans Readjustment Study* (VA Contract No. V101 (93) P-1040). Research Triangle Park, NC: Research Triangle Institute.

Regier, D. A., Myers, J. K., Kramer, M., & associates. (1984). The NIMH epidemiologic catchment area program: historical context, major objectives, and study population characteristics. *Archives of General Psychiatry*, 41, 934-941.

Robins, E. & Guze, S. B. (1970). Establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry,* 126, 107-111.

Robins, L. N., Helzer, J. E., Ratcliff, K. S., & associates. (1982). Validity of the diagnostic interview schedule, version II: DSM-III diagnoses. *Psychological Medicine*, 12, 855-870.

Robins, L. N. (1985). Epidemiology: Reflections on testing the validity of psychiatric interviews. *Archives of General Psychiatry*, 42, 918-924.

Robins, L. N., Helzer, J. E., Croughan, J., & associates. (1981). National Institute of Mental Health diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry*, 38, 381-389.

Robins, L. N., Helzer, J. E., Orvaschel, H. & associates. (1985). The diagnostic interview schedule. In W. W. Eaton & L. G. Kessler (Eds.). *Epidemiologic Field Methods in Psychiatry: The NIMH Epidemiologic Catchment Area Program* (pp. 143-170). Orlando, FL: Academic Press.

Robins, L. N. (1986). The development and characteristics of the NIMH diagnostic interview schedule. In M.M. Weissman, J.K. Meyers, & C.E. Ross (Eds.). *Community Surveys of Psychiatric Disorders* (pp. 403-427). New Brunswick, NJ: Rutgers University Press.

Schlenger, W. E., Kulka, R. A., Fairbank J. A. & associates. (In press). The prevalence of post-traumatic stress disorder in the Vietnam generation: findings from the National Vietnam Veterans Readjustment Study. *New England Journal of Medicine*.

Spitzer, R., Williams, J. B. W., & Gibbon, M. (1987). *Structural Clinical Interview for DSM-III-R, Version NP-V*. New York: New York State Psychiatric Institute, Biometrics Research Department.

U.S. Congress. Senate. (July 14, 1988). Committee on Veterans' Affairs. (1988). *Oversight on Post-traumatic Stress Disorder: Hearing* (pp. 127-221). 100th Congress, 2nd Session. Washington, D.C.: U.S. Government Printing Office.

Weissman, M. M., Myers, J. K., Ross, C. E. (1986). Community studies in psychiatric epidemiology: An introduction. In M.M Weissman, J.K. Meyers, & C.E. Ross (Eds.). *Community Surveys of Psychiatric Disorders* (pp. 1-19). New Brunswick, NJ: Rutgers University Press.

DISCUSSION PAPER

Pretesting: A Neglected Aspect of Survey Research

Stanley Presser

As is vividly demonstrated by a number of these papers, to say a question has been pretested, may be saying very little. Although it is likely that pretested questions are generally better than unpretested ones, there is ever-mounting evidence that many pretested items perform poorly.

This is partly because we have almost no rigorous knowledge about how to pretest. Despite its importance, there is probably no phase of survey research that has been the subject of as little systematic research as pretesting. The typical methods textbook acknowledges the key role played by pretests, but confines its treatment to a few perfunctory paragraphs. Numerous books explain how to write questionnaires, but almost nothing is available on the critical step of how to test and evaluate them. The closest one comes to an extended treatment is the Office of Management and Budget Statistical Policy Working Paper edited by DeMaio (1983).

One of the few formal research efforts was organized by Charles Cannell almost 20 years ago (Marquis, 1971; Cannell & Robison, 1971). It grew out of a project designed to study the problems of doing surveys in inner city poverty areas in the late 1960s. Recognizing that the nature of the sample might lead to more than the usual level of questionnaire difficulties, Cannell and his colleagues applied an objective coding scheme to tape recorded interviews to identify problem questions.

Although the results of that study seemed quite promising, few researchers made use of or further developed the idea. Cannell himself used the interaction coding of interview recordings to spearhead a line of research on interviewing (Cannell & associates, 1975; Blair, 1981; Brenner, 1982), but with the notable exception of Jean Morton-Williams' work at Social and Community Planning Research in London (Morton-Williams, 1979; Morton-Williams & Sykes, 1983), we have had to wait until today and the project reported by Fowler for that original line of inquiry to be renewed. Why that is the case would make an interesting study in the sociology of science.

Just as Fowler's report is a return to Cannell's work of two decades ago, Royston's paper may be seen as a variation of work begun almost three decades ago by William Belson (reported in summary fashion in Belson, 1981). In Belson's case, the intensive interviews were designed to uncover how respondents had interpreted the questions they had been asked the day before as part of a regular survey. As you know, Belson found very high levels of misunderstanding, even on seemingly straightforward items.

Although the work reported by Royston is in many ways akin to Belson's, it owes its actual origins to the recent courtship of survey research and cognitive psychology. As Royston notes, the Questionnaire Design Research Laboratory at the National Center for Health Statistics was set up with the express purpose of using the methods and theories of cognitive psychology to identify sources of survey response error. However, the absence of a cognitive psychology imprint in Royston's paper is striking. Aside from the "concurrent protocol" jargon, there is little in either the method or the interpretation of results that comes directly from cognitive psychology.

This does not diminish the value of Royston's work. Her results are important and clearly improved the surveys she discusses. If the ferment about applying cognitive psychology to surveys does nothing more than increase the resources devoted to pretesting, it will have been all to the good.

Yet the lack of a clear cognitive psychology stamp does raise the question of whether there is something unique about intensive interviewing in the lab, or whether similar problems would have been identified if comparable effort had been invested in traditional pretesting. Royston presents several reasons for believing that the usual pretesting methods would miss the diffi-

Stanley Presser is with the Department of Sociology and the Survey Research Center, University of Maryland, College Park.

culties uncovered in the lab, but without an experimental comparison of the two approaches the point remains uncertain.

A similar question may be posed about the paper presented by Fowler. Good general reasons exist for believing that formal rules for pinpointing problems, such as those used by Fowler and his colleagues, should outperform the informal, more subjective approach of the traditional pretest. Still, it is important to demonstrate this empirically. We need experimental studies that pit the usual approach against newer approaches like intensive interviews or the analysis of coded interviews.

Even more fundamental than studies comparing the various approaches, we need investigations of the reliability of each method. Fowler and associates allude to this problem indirectly when they say "we have . . . shown that 50 interviews produce very stable results." Although the basis for this claim of stability is unclear, it gets at a key issue. To what extent does a pretesting method produce similar results if conducted on multiple occasions? Without some sort of repeated trials experiment, it is impossible to know.

Only when we have some indication of the reliabilities of the different methods does it make sense to undertake comparisons between methods. And only at that point can we begin to identify the particular aspects of a procedure that seem most sensitive to uncovering question problems.

One of the strengths of intensive lab interviews that Royston points to is the commitment of the respondents "to helping find flaws in the questionnaire." This is what Converse and Presser (1986) call a "participating" pretest, in contrast to the more typical procedure that they call an "undeclared" pretest. Participating pretests can be conducted outside the laboratory; it will require an experimental comparison, however, to demonstrate whether the lab setting itself contributes to the success of the method.

A second strength of lab interviews noted by Royston is the ability to recruit special groups. This too is possible in the field, but in either case there is a potential danger in relying on an unrepresentative sample. The paper by Kulka and his colleagues underscores the point. Questionnaire items may behave very differently in a general population than in selected subgroups. At least in the case of post-traumatic stress disorder, measurement properties of indicators were greatly affected by whether they were tested on a clinical population or a general one.

Kulka's research compared different indicators by administering them to a single set of respondents, the same tack taken in the first study by Anderson and colleagues. This strategy is subject to unwanted context or sequence effects. One form of an item may be affected by another version of the item that precedes it. The order in which the self-report and interviewer modes were administered in the first study by Anderson and co-workers is unclear. But if the interviewer mode was second it may have performed better than the self-report because respondents and interviewers had more time and information in making judgments. If this were the case, reversing the order of administration of the methods would alter conclusions about their validity.

The solution is a split-sample design of the kind used in the second study by Anderson and associates. By asking alternate question forms of separate random subsamples, the problem is avoided.

Still another approach to studying the measurement properties of questions involves multitrait-multimethod models like those described by Campbell and Fiske (1959). This seems to be the logic underlying the paper by Hembroff and his associates. Analysis of a correlation matrix can identify the multiple facets of a construct and indicate which items tap different dimensions. Of course, the results of such an analysis are totally constrained by the particular items asked, and although the analysis can identify shared variance, the meaning of that variance is open to interpretation.

Consider, for example, a test of five questions, four of which measure a construct but are contaminated by a common methods variance and the fifth of which is an error-free indicator of the same construct. A factor analytic approach might suggest that the fifth item is the rotten one. Statistical modeling can be a powerful tool, but the numbers it produces obviously have to be interpreted cautiously.

These papers encompass many, but not all, of the methods that may be used to test questionnaires. One promising strategy not covered assesses the meaning of questions by asking respondents how individuals in hypothetical vignettes ought to answer the items (Campanelli & associates). If these papers are any indication, still other methods to evaluate questions will be developed in the future. This is a welcome prospect, for if surveys are to be taken seriously, the issue of questionnaire pretesting requires vastly more systematic research than it has received until now.

References

Belson, W. (1981). *The design and understanding of survey questions*. London: Gower.

Blair, E. (1981). Interviewer variations in asking questions. In N. Bradburn & S. Sudman (Eds.). *Improving interview method and questionnaire design*. San Francisco: Jossey Bass.

Brenner, M. (1982). Response-effects of "role-restricted" characteristics of the interviewer. In W. Dijkstra & J. Van der Zouwen (Eds.). *Response behaviour in the survey-interview*. London: Academic Press.

Campanelli, P., Martin, E., & Creighton, K. (1989). Respondents' understanding of labor force concepts: Insights from debriefing studies. (In press). *Proceedings of the Fifth Annual Census Bureau Research Conference*.

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cannell, C., Lawson, S., & Hausser, D. (1975). *A technique for evaluating interviewer performance*. Ann Arbor, MI: Institute for Social Research.

Cannell, C., & Robison, S. (1971). Analysis of individual questions. In J. Lansing, S. Withey, & A. Wolfe (Eds.). *Working papers on survey research in poverty areas*. Ann Arbor, MI: Institute for Social Research.

Converse, J., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills, CA: Sage Publications.

DeMaio, T. (Ed.). (1983). *Approaches to developing questionnaires*. (Statistical Policy Working Paper 10.) Washington, DC: Office of Management and Budget.

Marquis, K. (1971). Purpose and procedure of the tape recording analysis. In J. Lansing, S. Withey, & A. Wolfe (Eds.). *Working papers on survey research in poverty areas*. Ann Arbor, MI: Institute for Social Research.

Morton-Williams, J. (1979). The use of "verbal interaction coding" for evaluating a questionnaire. *Quality and Quantity*, 13, 59-77.

Morton-Williams, J. & Sykes, W. (1983). A study of question failure through the use of interaction coding. *Bulletin of the International Statistical Institute*, 1, (pp. 479-494). Madrid, Spain: International Statistical Institute.

DISCUSSION PAPER

Developing Health Measurement Standards: Toward a Basic Science of Health Assessment

Ian McDowell

Introduction

Limitations in the quality of contemporary health measurements are due less to our ignorance of how to develop sound measurements than to our unwillingness to apply these techniques. These methodological papers illustrate the wide range of techniques that are available to improve the quality of health survey measurements. Reflecting on this breadth of methodological experience, I am drawn to comment not so much on the papers themselves, as on their potential to influence the development of health measurements in general. I approach this topic not as a survey researcher, but as someone more broadly interested in health measurements, whether for clinical trials, evaluative research, or as survey tools. Therefore my task will be interpreted rather broadly, tackling the issue of how to ensure that these technical innovations get applied in improving the quality of health measurements. The methodological work is crucial, but not sufficient on its own to enhance the overall quality of the field: we must work to ensure its application in the broader development of the health measurement field as a whole. Let us step back from the detailed discussions for a moment, and consider possible strategies to ensure a broader dissemination of these methodological innovations. I propose that these methodological developments be integrated into formal standards for health indices, that will strengthen the basic science of health measurement.

As with so many academic endeavors, the development of health measurements has occurred in a pleasantly uncoordinated manner. The demands of clinicians, of health care evaluators and of survey researchers encouraged the development of health indices and questionnaires. Imaginative people bridged the social and medical sciences to create these methods which, it must be admitted, were of varying quality. Initially, the emphasis lay on creating indices where none existed before, and to a large extent concern over the quality of the product was secondary. The field is highly applied, and no identifiable discipline (equivalent to psychometrics or econometrics) was developed to guide people's creative talents. Methodological work of the type we have heard today has not been assembled into a coherent basic science for this effort, and we should now consider developing a more formal approach to this young discipline. I argue that a practical component of this will be the formulation of explicit guidelines for the development and testing of health questionnaires and measurements, and I wish to stimulate discussion over what the overall strategy might be.

Background

The health indicators movement began with a phase in which large numbers of measurements were developed in a somewhat uncoordinated fashion. Several dozen Activities of Daily Living (ADL) scales were developed, often virtually indistinguishable from each other; the jumble of screening tests for dementia are another example of this phase. Reviewing this work indicates a lack of attention to quality control: many early measurements remained innocent of validity and reliability testing. There is more academic reward in inventing a method than in testing it subsequently, so we saw very little systematic comparison of the performance of rival scales. Relatively superficial criticisms of scales led to their replacement by others that were only marginally better. Rarely was adequate time devoted to the task of developing the questionnaires and scaling techniques; we have learned from the Sickness Impact Profile (SIP) and the Quality of Well-being Scale that this must be counted in years.

Concern grew over the adequacy of health status measurements, and recently there has been a phase of

Ian McDowell is with the Department of Epidemiology and Community Medicine, University of Ottawa, Ontario, Canada.

critically reviewing the available methods. This is useful but reactive, and will not of itself lead to any improvement in the field. We now have a choice; a laissez-faire approach would perpetuate the uncoordinated and haphazard tradition, leaving those who develop measurements to pick up what they will of the methodological developments. Alternatively, we could collate the technical advances to form the foundations of a more systematic science of health measurement.[1] The standards for tests and measurements proposed by the American Psychological Association (APA) provide a model for this, although one that is old and does not perfectly fit the case of health measurement. The APA standards formalized testing procedures, and explained how to report the results of this testing (American Psychological Association, 1974). Although we may question the precise impact of these standards, note that psychological scales and questions have in general been far better tested than other health measurements, such as indices of physical functioning. Why should we not make similar efforts in health measurement? What would be the benefits of an attempt to lay down guidelines for developing health measurements, and what might such a program look like?

Why Propose Guidelines for Developing Health Measurements?

Most reviews of the field have identified shortcomings in the current status of health measurements. The proposed guidelines are intended to suggest solutions to these. There are shortcomings in the field as a whole, including the uneven coverage of current health measurements, with gaps in some areas and large numbers of scales in others. In addition, there are shortcomings in the quality of individual scales. These refer more to a failure to apply appropriate techniques such as the validation and item development methods discussed today than to a lack of such methods. Methodological guidelines should propose suitable development and testing standards and should set standards for presenting published descriptions of the instrument. By guiding those who develop and those who use measurements, standards would foster the explicit development of a basic science of health measurement. They would seek to foster an ethic of quality control in the field. It is critical that such measurement standards are not arbitrary; they must be backed by a wide consensus over their content, and many people in the field must have the opportunity to influence their creation. They would, of course, have to be updated as methods improve.

Standards are only of value if they are followed, and although no professional body in this field can impose sanctions, we may consider strong but informal approaches to encouraging adherence. It would be valuable to have the standards supported by a respected professional body such as the International Epidemiological Association or the American Psychological Association. Such standards should be widely disseminated, especially to granting agencies and to journal editors, so that they may be used in evaluating grant proposals to develop measurement methods and in reviewing articles describing them.

Proposed Content of the Measurement Standards

Topic I: Coverage of the Field

The current problem is simple to state, more difficult to solve. In some areas of health measurement, we have a choice of acceptable techniques, whereas in others the methods are either of poor quality or lacking altogether. Few difficulties confront the person who wishes to measure depression, anxiety, general health, or perhaps pain; whereas the choice is limited if you would measure coping or child health or evaluate health promotion. In areas such as functional disability and cognitive testing, the problem is one of quality not of quantity.

If our research is to be cumulative, we must use comparable measurement techniques as long as these are of good quality. Our goal should be to consolidate the range of measurements we have available, promoting good scales and discarding inferior ones. We should not encourage researchers to develop new scales where there exist good- quality methods, but instead should direct their attention toward those areas of measurement that are underserved. Several reviews have been produced recently, and these could be combined into a meta-analysis of the current state of the field. This requires some further work to refine ways to summarize the quality of a scale or set of questions, but we have already made progress in this direction. This analysis could be used to suggest priorities for areas of further research; the information could be circulated to funding agencies to guide an RFP funding program toward the most cost-effective approach to developing the field. Evidently such a review would represent the state of the field at a certain time and would have to be updated every few years, but is intended as a powerful stimulus to rational planning.

Objective No. 1: To develop consensus summary of areas of strength and weakness in current measurements. To select areas in which further development is required.

Topic II: Conceptual Specification of Health Measurements

There is a lack of overall coordination in the development of health indices. This was seen earlier in the fashion of reinventing the ADL, and more recently in the titles given to some measurements. Currently, virtually all forms of measurement claim to indicate quality of life, and there is an urgent need to agree on standard terminology. Quality of life, for example, is too often used as a buzz word to catch the eye of granting agencies, suggesting a nobility of concern by the investigator or therapist. Like democracy, it forms a rallying cry for

[1] Call it what you will, "Salutometrics"?

people of widely different persuasions; like democracy, this effect is enhanced by avoiding precise definition of the term.

The paper by Anderson and Kaplan illustrated the need for careful definition of terms, even though their scale is based on an explicit conceptual formulation. Many of the concepts used in health measurements are not simple: well-being, functional limitations, changes in normal activities that are health related, and so forth. These require close definition, for we may expect people to interpret them in different ways. For example, it is not clear whether an activity limitation day implies a reduction in well-being. Perhaps taking a day off work and resting will enhance my subjective well-being, quality of life, and so forth. Figuring out whether my decision to take a day off work was health related (not sickness related) is equally hard. Did it constitute a problem? I may tell different stories to different people.[2] Sound interviewer training can overcome much of the uncertainty, but the principles underlying the training should be based on a specified theoretical approach to the field.

Objective No. 2: A standard vocabulary must be developed to ensure that terms are used in a consistent manner.

Defining terms is, however, only a small part of improving the conceptual specification of measurements. The conceptual definition should link the measurement to a theoretical approach to the topic: a particular theory of the pain response, or a theory of coping, for example. It may seem obvious that a health measurement should be based on a particular conceptual approach, and yet this is rarely the case. By basing a measurement on a particular theory the measurement may be used analytically rather than simply descriptively; the theory may also be tested. The conceptual description indicates which questions should be selected for the instrument and provides a guide to the appropriate procedures to use in validating it. It also addresses the problem that empirically scales such as Leavitt's Back Pain Classification Scale succeed in their stated aims (for example, of distinguishing pains of organic origins from those of emotional origins), but we lack any theoretical understanding of why the scale works.

Objective No. 3: Measurements should be based on a particular theoretical approach to the topic, the theory guiding the design of the measurement, and the measurement enabling the theory to be tested.

Guidelines. How should we foster attention to the conceptual aspect of measurement? There is, after all, a long tradition of separation between theorists and empirical researchers. Probably the impetus to establishing a dialogue between the two will come from both camps. Theorists, such as Lazarus in the coping field and Melzack in the pain field, have proposed measurements that reflect their theories. At the same time, as illustrated by Hembroff's paper, empirical analyses can contribute to the development of theory. Conceptually, how separate is health status and health performance? Hembroff's data suggest that it varies from topic to topic, with smoking showing a clear and consistent relationship between performance and health status, while stress and nutrition do not show a clear relationship.

Topic III: Methodological Standards

This brings us to the topics that are of most interest to the readers of this volume. Fields such as psychometrics or econometrics, as well as boasting a formal title, have progressed by setting standards for the development and testing of new measurement techniques. With an eclecticism born of necessity, we have borrowed from these, but often without adequately adapting the methods of validity or reliability testing to the particular requirements of the health field.

In this context, standards aim to ensure at least an adequate minimum level of rigor in development and testing for all questionnaires or measurement scales. Minimum standards should be set in areas of instrument development and of validity and reliability testing.

Instrument Development. Item writing and item analysis are topics all too commonly relegated to a distant corner of measurement texts. Questions somehow get written; perhaps if they are bad enough someone will notice during the validity testing. They may get altered, and if there is enough time, they may be retested. But this is generally done in an unsystematic manner. A sound conceptual formulation should indicate which topics are to be covered in the questionnaire, but these papers have demonstrated how difficult it can be to produce satisfactory question wordings.

Objective 4: A manual should indicate appropriate procedures for examining the quality of all questions used in health measurements.

We have heard a wide menu of approaches to examining and improving the clarity of questions; minimum recommended procedures should be selected and included in a procedural manual. The choices include intensive interviews, probing the response, paraphrasing, having the respondent think aloud while answering. Other approaches are not based on the interview: the language should be tested, perhaps using a reading-age formula or one of the complexity formulas now enshrined in our word processors. Another approach is to borrow from cross-cultural research and use translation into another language and back translation. If the back translation is the same as the original version, this suggests that it conveys a clear concept.

In short, we do not lack techniques for testing questions, although we may lack the determination to apply them routinely. As recommended by Royston, these detailed tests should be used in addition to the normal field tests, which re-examine the items in the context of the complete questionnaire.

Validity Testing. The validation procedures used in developing many familiar health measurements again indicate a need for refinement. There is little agreement that a particular type of measurement should be tested

[2] A Herman cartoon shows a man leaving his house to go fishing. He is met at the door by his boss who says, "I know you are at home with the flu, but please could you let me have the keys to the filing cabinet?"

in a particular manner. Approaches to validation seem to reflect the academic backgrounds of those who develop the tests as much as the requirements of the situation. For example, heavy emphasis is placed on criterion validation for psychological indices, whereas disability scales are often tested by comparing groups known to differ on level of disability. There are limitations to both approaches, and for health indices we require a broad-ranging set of validity tests. These should pay attention first to the internal structure of the method (internal consistency, factor structure) and then to its relationship with other scales and clinical assessments, and then to its ability to detect change. We should pay more careful attention to the validity of the criterion scores against which we rate our measurements. The difficult issue of selecting a gold standard was implicit in Anderson and Kaplan's paper: which should be taken as the truth, the interviewer's impression or the subject's declaration? Should the subject's denial be seen as error or as a component of his or her health status?

Objective No. 5: Guidelines should outline the preferred validation procedures for various types of health measurement instrument.

Because we so often lack a gold (even a gold plated) standard, frequent reference is made to construct validation. This remains ill defined and promises more than it produces. All too often an author presents a list of correlations, perhaps ranging widely, and apparently arbitrarily concludes that the results demonstrate the construct validity of the measurement. At the very least, a formal declaration should be made of the level of correlations to be accepted as indicative of validity, and why.

Objective No. 6: There is a critical need to place construct validation procedures on a more rigorous scientific basis.

Topic IV: Descriptions of the Measurement

Guidelines should cover presentation of the measurement in the literature. Frequently this is badly done, complicating the user's attempts to select among alternative methods. This field has suffered the ravages of the academic pressure to publish, which has often led to the publication of preliminary versions of a scale that is later revised. Users become confused and may continue to use outdated versions of a measurement. The Health Opinion Survey and the Life Satisfaction Index are examples. Ideally, there should be a manual of the measurement method, giving a full description of the scale, its origins and conceptual basis, showing the questions, the administration instructions, scoring procedures (including change scores), validity and reliability results, and an indication of population reference standards to indicate the range of scores to be considered normal.

Objective No. 7: Guidelines should specify standard information to be included in written descriptions of the measurement. Ideally, test developers should provide full manuals that show the questions, describe the development and testing procedures used, and give validity and reliability results.

Implementing the Guidelines

Guidelines are of no use unless followed. As we cannot impose sanctions, the guidelines will succeed or fall on their own merit. They should seek to stimulate, not stultify, critical work and should be seen in terms of guiding rather than enforcing quality. The guidelines will specify minimum standards; the format should also indicate desirable further criteria which the measurement should meet.

An initial draft of the guidelines could be prepared by a small working group, and a delphi technique could be used to refine them. To encourage adoption, the eventual guidelines should be ratified at a large meeting devoted to health measurement. Perhaps this could become an agenda for the Sixth Annual Conference? Formal backing from a respected academic society would be desirable, and the guidelines should be published. In the long term, such a publication could be used in training courses, and could form the basis for the further development of a basic science of health measurement. In the short term, the idea of circulating the guidelines to editors and funding agencies, as guides to reviewers of projects proposing new measurement techniques, was noted above. A mechanism would be needed to ensure periodic updates, perhaps through this series of publications.

Reference

American Psychological Association (1974). *Standards for Educational and Psychological Tests.* Washington, DC: American Psychological Association.

1990 Census: Counting Selected Components of the Homeless Population

Cynthia M. Taeuber

Enumerating Selected Components of the Homeless Population

Census Day for the 1990 Census of Population and Housing is April 1, 1990. This will mark the Bicentennial of census-taking in this country, as our Nation has taken a census every 10 years since 1790. Major plans for the 1990 census are complete and we have begun preparatory operations.

Taking an accurate census in a complex and mobile Nation is an enormous challenge, especially when our charge is to count every person whose usual residence is in the United States on Census Day. We must count not only those who live in housing units and group quarters, but also those who have no usual home. How we will count selected components of the homeless population is my topic today.

The Bureau of the Census is actively building a nationwide operation for the 1990 census to provide basic demographic, social, and economic data on selected components of the homeless population. The operation will cover rural as well as urban areas, and will include both adults and children. The homeless have been counted in previous censuses but this is the first time we have had such a focused and ambitious effort to improve the count and to identify separately selected components of the homeless population.

Information has been lacking about the numbers and characteristics of America's homeless population. We don't know how fast it is growing or where it is growing fastest. There are no generally agreed-upon national counts of the homeless population. The Department of Housing and Urban Development, in 1983, used six different methods to estimate that there were 250,000 to 325,000 homeless persons (Department of Housing and Urban Development, 1984), while advocates for the homeless estimated that there were as many as 3 million (Hombs & Snyder, 1983). A study by Burt and Cohen (1988) of the Urban Institute estimated that there were close to 600,000 homeless in 1987.

Behind the lack of agreement about the numbers, there is no generally agreed-upon method of defining homelessness or counting the homeless. To most, the homeless include at least those who sleep in the streets at night or who live in emergency shelters or transitional housing because they do not have regular access to conventional housing. They are obviously and literally homeless. Some widen the definition to include the precariously housed, such as "doubled-up families" and others who are at risk of becoming homeless.

Counts of the homeless have come from expert opinions, estimates based on the use of services, and surveys with significantly different methodologies. The definition, the year, the time of year, and whether the estimate is based on a single point in time or over a longer time period all affect the outcome of studies. As a result, there is a wide range in estimates of the size of the Nation's homeless population.

The Government Accounting Office (GAO) (1988) reported on nine options for counting the homeless mentally ill and evaluated the associated biases of each option. They found that some of the variability in the estimates of the homeless are related to the method used. For example, estimates based on expert judgment produced a median homeless rate of 29 per 10,000; utilization-based estimates produced a median rate of 18 per 10,000; and census-based studies, a median rate of 13 per 10,000. Studies based on expert judgment produced the most variability in rates. The GAO also rated the technical quality of the studies and found that the most highly rated had a median homeless rate of 13 per 10,000 compared with the lower-quality studies with a median homeless rate of 22 per 10,000. In addition to the method used and the quality of the study, other sources for the differences in estimates of homelessness include: (1) true differences among local areas; (2) the definition of homelessness that is used; (3) whether the estimate

Cynthia M. Taeuber is with the Age and Statistics Branch, Population Division, Bureau of the Census, Washington, D.C.

is for a point in time (incidence) or over the course of a longer period such as a year (prevalence); and (4) the year and the time of the year in which the study was conducted (Government Accounting Office, 1988, p. 29).

Of the nine options described by the GAO, four were favored for counting the "literally homeless" based on how each option addressed the sources of bias and the cost of the option. GAO's most preferred method is for a one-time representative sample survey of shelters, institutions, and the streets, a strategy that is costly but addresses the large number of sampling biases inherent in counting the homeless. For such a sample, they recommend that a "rolling" sampling strategy be employed to account for the known seasonal variation in the number of homeless. For a national survey, GAO also recommends a two-stage probability sample of cities and likely residential settings (that is, shelters, institutions, and the streets). The following explains how the 1990 decennial census compares with this option.

Census Plans for 1990

The 1990 decennial census goes beyond a sample of cities. It is nationwide and will provide demographic, social, and economic data on selected components of the homeless population. It will also enumerate persons in many types of places, including shelters for the homeless and those living on the streets and in abandoned or boarded-up buildings. Because there are different perspectives on homelessness, the Census Bureau will not provide an official definition of the homeless. Neither will we provide a total count of the homeless from the 1990 census. Rather, counts and characteristics of selected components of the literally homeless and the precariously housed population will be provided.

There will be an enumeration of people in shelters and on street locations on March 20 and 21, 1990; other components will be enumerated as part of the regular census enumeration in April. The selected components can be analyzed separately and used as benchmark data and as building blocks to construct a count of homeless appropriate for the purposes of particular programs.

The census cannot meet every data need related to the homeless population, however. Beyond its mandated uses for Congressional reapportionment and legislative redistricting, the decennial census serves as a general purpose survey. The data collected have to be applicable to everyone, not just the homeless population. The census will provide a count of the selected components of the homeless population on one night and cannot measure the number of people who move in and out of homelessness. GAO's suggestion of a rolling sample is not possible for the decennial census even though it would be desirable to account for seasonal variation in the count of the homeless. The dynamics of homelessness will not be learned from the decennial census; the data will provide a snapshot, not a movie.

The Census Bureau faces certain constraints during this enumeration. The safety of our employees and the respondents has to be considered. Because the enumeration will be done in one night in most places, with temporary employees, and because the enumerators will be inexperienced in many cases, the enumeration operations must be kept as simple as possible. Cost constraints are also a factor.

For the shelter and street enumeration on March 20 and 21, 1990, the Bureau will recruit those who work with the homeless, as well as the homeless themselves, to apply for work as enumerators and supervisors of the enumerators. All enumerators, of course, must meet the same employment criteria and swear to keep the data confidential. Enumerator training is written to prepare all persons for this unique enumeration, but persons experienced in working with the homeless can be expected to adjust quickly to the situations. The more the enumerators are familiar with the homeless, the more complete the count will be.

Preidentification of Enumeration Sites

The locations of two components, shelters and street sites (including abandoned and boarded-up buildings) will be identified before the census with the help of local officials. In September 1989, the Census Bureau will send certified letters (receipt required) to the highest elected officials of some 39,000 local governments. They will be asked to send to us, by October 16, 1989, a list of all shelters with sleeping facilities for the homeless (private and public, permanent and temporary), subsidized hotels and motels and rooms used to shelter the homeless, street locations, and abandoned and boarded-up buildings where homeless people congregate at night. These officials will also have the opportunity to update the lists if more information becomes available before the enumeration. The Census Bureau's lists will be supplemented with this local knowledge and these will be the sites where Shelter and Street Night Enumeration will be conducted. If the local governments of cities with a population of 50,000 or more do not respond, the Census Bureau will take the responsibility of getting a list of preidentified sites.

Shelter Enumeration

First, census takers will count persons in the preidentified emergency shelters and hotels and motels used to shelter the homeless on the night of March 20, 1990. The count is scheduled to take place during the hours when the population can be expected to have settled in for the night. In most cities, enumeration in shelters will occur from 6 p.m. to midnight.

In most cases, persons will complete their own questionnaires. If they are unable to do that, census enumerators will assist them in completing the questionnaire. Persons at the pre-designated sites will not be asked if they have a usual home elsewhere, and thus we cannot guarantee that every person included in the Shelter and Street Night component is actually homeless. It is likely, however, that virtually everyone enumerated at shelters is there because he or she is homeless on enumeration night.

Street Enumeration

Street enumeration will generally occur from 2 a.m. to 4 a.m. on March 21, 1990. Enumeration of predesignated abandoned and boarded-up buildings will follow from 4 a.m. to 6:30 a.m. There is less certainty that

everyone counted during the street enumeration is actually homeless. It is difficult to tell who is homeless by a person's appearance or by asking questions. For the 1990 census everyone on the street at the predesignated sites will be enumerated except persons in uniform, such as the police, and those engaged in obvious money-making activities or commerce, other than begging or panhandling. The Census Bureau thinks this decision will help to reduce the undercount of the homeless persons that could occur if inexperienced enumerators relied on their subjective judgments about who is homeless.

Of course, some persons with a home may be out on the streets from 2 a.m. to 4 a.m., but in most cities, the Census Bureau believes this number will be relatively small in the areas of the cities being covered and at that time. The 1990 street enumeration cannot include the entire universe of persons who sleep out on the streets. For example, it will be very difficult to include those who sleep outside the predesignated sites, the people who live in their cars, those who ride in busses and subways all night, or those who sleep in dumpsters, on tops of roofs, or who are otherwise well hidden. Thus, the street enumeration is best viewed as a count of visible persons in pre-identified street locations from 2 a.m. to 4 a.m. on March 21, 1990, who are not engaged in obvious money-making activities. Data users will have to review the results of the 1990 enumeration to determine if the street component seems reasonable for their area.

The homeless, and especially those living on the streets, are among the most difficult populations to count completely. It is conceivable that in some areas the count will be distorted by erroneously including people who are not homeless. The street count will probably be conservative in most areas, however, because census takers cannot count people who are moving about or who are so hidden that their whereabouts are unknown even to local people who work with them. Physical dangers are present in some of the areas where this enumeration takes place, which is why the enumerators will not search cars, enter abandoned buildings (rather, they will be stationed outside of such buildings as described below), or climb up on roofs or into dumpsters. Safety, for both the enumerators and the respondents, has to be an important consideration.

The enumerators will work in teams. Police escorts will not be used, as they are in some private surveys, because census data are confidential. In addition, the Bureau is concerned about the perception of police involvement in the census. There is the possibility that respondents will incorrectly believe that any subsequent police actions in those areas are connected with census activities or that the census is a part of some police activity.

The count of the street population will be best for areas that can preidentify night locations of the street population, areas where more of the street population stays visibly out in the open rather than hidden, and where the Census Bureau can hire persons experienced with the homeless to work as enumerators. Preidentified abandoned and boarded-up buildings will be enumerated from 4 a.m. to 6:30 a.m. Enumerator teams will go to the pre-identified sites and wait outside the buildings in their cars. They may be assigned to more than one building and may go back and forth until they see someone they can enumerate. Only basic demographic (short form) information will be collected. The enumerators will try to get the number of people inside the building and basic demographic information about them from the first person. If this is not possible, enumerators will wait until someone else comes out. As many abandoned buildings as the budget allows will be included but local officials will be asked to assign priorities to the abandoned buildings they want enumerated first because census takers may not be able to cover every building.

In addition to the enumeration of abandoned buildings on March 21, as part of the urban enumeration procedures the Census Bureau will be identifying city areas with large numbers of abandoned and boarded-up buildings. Enumerators will be in these areas during the day on March 26, 1990, to determine if the units are vacant or if they have been converted to use since the initial listing. If the building has not been renovated but people are present, they will be enumerated but cannot be included as a component of the homeless population. Rather, they will be listed as living in housing units.

Other Components of the Homeless Population

Other persons sometimes included in the count of the homeless population will be enumerated as part of the regular census operations. These include doubled-up families, persons in institutions such as local jails which may provide temporary shelter, and persons with no usual home elsewhere living in tents at commercial campgrounds or in shelters for abused women. In institutions such as local jails, people will not be asked if they have a usual home, and thus homeless persons in institutions cannot be identified separately.

Families living in emergency shelters and transitional housing will be enumerated, as will those in barrack-style dormitories or in the so-called welfare hotels in the special March 20th Shelter and Street Night Enumeration. Summary statistics will be published for persons in these types of living quarters. Others might be homeless if they do not have another family with whom they can double up. This group will be enumerated as part of the regular household enumeration. There is no agreement on when "doubled up" means "homeless," so the Census Bureau will provide tabulations of all housing units with more than one family, cross-classified by other relevant characteristics such as income, poverty status, persons per household, and the percentage of family income spent on housing. Researchers and planners can use these data as indicators of the precariously housed or homeless as they see fit.

Information Available from the 1990 Census

In the Shelter and Street Night Enumeration, all persons in shelters and persons enumerated on the streets

will be asked to answer basic demographic questions. Names will not be collected during street enumeration. Enumerators will not wake up sleeping persons to ask them questions. Rather, enumerators will estimate as best they can the sleeping person's age, sex, and race. The enumerators will do the same for refusals and persons not in a state of mind to answer questions. If a sleeping person is covered up so that characteristics cannot be determined, the person will be counted and characteristics will be assigned later by statistical methods. It is difficult to guess a person's age and so ages in three broad age ranges only will be published. For a sample of persons in shelters, the enumerator will ask additional questions about socioeconomic characteristics; these additional questions will not be asked during the street enumeration.

The count and the characteristics of the population will be tabulated in each of the selected component settings in which the homeless live:

- shelters with sleeping facilities, including those for runaway and neglected children, low-cost hotels and motels, and hotels and motels used by cities to house the homeless regardless of cost;
- nonsheltered locations such as those enumerated at street locations (including train stations, bus stations, and so forth) and in abandoned and boarded-up buildings;
- shelters for abused women (for persons who report no other usual home);
- transient sites such as commercial campgrounds (only for persons who report no other usual home);
- maternity homes (for persons who report no other usual home); and
- drug and alcohol abuse group homes and detoxification centers (for persons who report no other usual home). Note: hospitals and hospital wards in psychiatric and general hospitals for drug and alcohol abusers are considered institutional populations and cannot be included as a component of the homeless population.

The census provides general information applicable to everyone, not just the homeless. The 1990 census will provide basic demographic information for all selected components of the homeless population. Social and economic information (for example, education, veteran status, income, and labor force participation in 1989) will be available for all components except the street population. There are no questions specific to the homeless such as why they are homeless, how their time is spent, where they get services, or their history as a homeless person. Detailed, specific questions are appropriate for voluntary surveys, but not for the decennial census which everyone is required by law to answer.

There tends to be much focus on the size of the count of the homeless that will result from the 1990 census. Perhaps it is more important to view the 1990 census as an opportunity to get a clearer idea of differences among areas of the country. The data also should give the Nation a better picture of the diversity of the homeless population.

Research and Evaluation

The Census Bureau is evaluating a proposal to develop a research program as part of the 1990 census to improve the methodology for counting the homeless in later surveys. The proposed research is to determine if there are screening questions that will help us in future efforts to determine objectively who on the street is actually homeless. As part of this proposal, a taxonomy of homeless persons will probably be developed. To date, screening questions used in other surveys have not addressed the problem of how to count people who are frequently homeless for several weeks at a time (especially at the end of the month when limited earnings or benefit checks run out) or people who do not want to admit that they are homeless.

The Census Bureau is also evaluating a proposal to conduct a research and evaluation program on the homeless to test alternative methods for counting the homeless. The proposal includes the use of administrative records, interviewing at service points and congregating sites during the day, and independent evaluations by participant observers.

Summary

This is the basic plan for the 1990 census for enumerating selected components of the homeless population nationwide. There are limitations in a decennial census, and not all sources of bias in enumerating the homeless can be addressed. The Bureau believes it has developed an effective method and a practical approach that should provide useful benchmark data for use by analysts, planners, and policy makers.

The focus and extensive attention the Census Bureau is giving this issue will provide more useful data on selected components of the homeless population than has been available before. Also, the emphasis and new procedures—such as working closely with local areas to identify shelters and street locations—should improve the chances of including homeless persons in the census counts.

References

Burt, M. & Cohen, B. (1988). *Feeding the homeless.* Urban Institute, General Accounting Office, Washington, DC.

Hombs, M. E. & Snyder, M. (1983). *Homelessness in America: A forced march to nowhere.* Washington, DC: Community for Creative Non-Violence.

U.S. Department of Housing and Urban Development, Office of Policy Development and Research (1984). *A report to the Secretary on the homeless and emergency shelters.* Washington, DC.

SESSION SUMMARY

Strategies for Evaluating Questions

Lu Ann Aday, Recorder, and Judith D. Kasper, Chair

The issues addressed in the discussion went beyond those raised in papers and discussants' remarks relating to the development of formalized methods for the design and testing of survey questions, the documentation and dissemination of the results of such studies, and whether these results could be generalized across population groups.

In response to the first paper by Patricia Royston on the use of intensive interview techniques in the NCHS Questionnaire Design Laboratory, questions were raised about whether items on the NHIS—which serve as models for many researchers in their own surveys—are systematically evaluated for test/retest and internal consistency reliability. Representatives from NCHS indicated that questions are changed on supplements to the NHIS each year; and often, time needed to conduct extensive methodological analyses of these items is insufficient. This limitation should improve with the recent establishment of a new research and evaluation unit within NCHS intended to provide an ongoing evaluation of the core NHIS questionnaire. In addition, however, the selection and form of items involves considerations other than methodological ones. For instance the suggestion to eliminate some conditions on the list of chronic conditions asked of NHIS respondents has met resistance internally within NCHS, as well as from CDC and OMB, because these items have been included in the NHIS for over 25 years. This view holds that trend analyses would be seriously compromised should they be deleted.

This point led to a broader discussion about laboratory-based methods research and the utility of documenting and disseminating this work. With regard to the NCHS Questionnaire Design Laboratory, no systematic methodological reports are prepared on the intensive interview and pretest results. (Royston and others, 1986, provides a more extensive description of the QDL and of the cognitive techniques being used.) Much of the laboratory's work supports internal questionnaire development, and time and resource constraints limit the systematic dissemination of the results. In addition, selection of items for testing tends to be driven principally by practical needs for questionnaire refinement rather than methodological concerns in general. Royston also pointed out that the final form of questions recommended by the laboratory is not always accepted because other design issues enter into the final decision-making, such as maintaining comparability over time or the preferences of investigators in the agencies funding NHIS supplements. She also reported that QDL staff have started to document more systematically the results of intensive interviewing methods in the lab and these are currently available in inhouse correspondence.

Other researchers working in laboratory environments pointed out the importance of systematically evaluating the advantages and disadvantages of intensive interviewing techniques. Two common techniques, the think-aloud approach and the postinterview interview, were discussed in some detail. Some researchers seemed to regard the think-aloud approach (whereby respondents are queried about how they are answering questions as the interview proceeds) as being of less value. Others saw real advantages to this strategy for understanding how respondents arrive at answers (for example, in reporting doctor visits, whether respondents count or use an estimating or averaging method).

An alternative technique, the postinterview interview, is similar to that used in early work by Belson (1981) on experiments with question wording. Here respondents are asked to comment on certain aspects of questions after the interview is completed. This eliminates one problem that potentially accompanies the think-aloud strategy—that is, asking evaluative questions in the course of the interview may well influence how respondents answer the questions that follow. It seems likely that both methods have value, but knowledge about their effectiveness for different types of respondents or in evaluating different types of questions is lacking.

The discussion around the relative merits of think-aloud versus postinterview interviews revealed that dif-

Lu Ann Aday is with the School of Public Health, University of Texas Health Science Center at Houston. Judith D. Kasper is with the Department of Health Policy and Management, The Johns Hopkins University.

ferent survey labs use different intensive interviewing strategies. However, there appears to be little systematic documentation or communication about which approaches are being used and how well they work.

A third approach to evaluating questions is the use of focus groups. Its use in a study directed at gay and bisexual men (see Bradford, in this volume) provided the opportunity for a variety of points of view to be expressed and discussed in developing a questionnaire on this group's sexual practices. Despite the potential of focus groups, it was suggested in later discussion that one misleading aspect of them is that certain responses or response sets that may be legitimized through a group process do not apply to actual one-on-one personal interview situations. Again, communication about the value of this approach under varying circumstances is relatively informal and unsystematic.

In response to Fowler's paper on the advantages of more quantitative approaches to question evaluation in the context of pretests, Verbrugge noted that those items for which the largest numbers of response problems were observed were ones in which the cognitive load on respondents was apt to be great (primarily due to question length). She posed the question of whether to build the detailed components of a question of interest into an introductory preface or to have a series of shorter branching questions. Research on the Quality of Well-Being scale supports the use of a detailed series of questions to detect disability rather than a single lengthy question that attempts to address both the presence of a health-related disability and its relationship to need for assistance. Prevalence estimates of disability were higher using the series of questions. Another participant, however, pointed out that work on question length by Laurent (1972) and by Bradburn, Sudman, and associates (1979) does not necessarily support the argument that longer questions are less desirable. Longer questions can in fact work well if they provide cues rather than clutter; that is, built-in redundancy may allow the respondent the opportunity to think through the concept before answering. The point was also made that the advantages or disadvantages of longer questions may vary by the age and education of the respondent (for example, reporting is better on longer questions for respondents with more education).

Questions were also raised about the costs of applying the interaction coding methodology to large-scale surveys. Fowler indicated only a small number of interviews (about 50) were needed to yield stable estimates. Concerns were expressed about whether, in methodological surveys, sample sizes are adequate to test for the significance of differences between groups for which different instrument design methods have been used. There was also a call for pretest samples to reflect a representative cross section of the population to be included in the final study. However, it was pointed out that trying to link probability designs with laboratory testing of questionnaires violates the theory underlying intensive interviewing and pretest techniques. It is often the case in laboratory settings that only selective survey design issues are addressed, and there may be problems in the pretest with obtaining adequate numbers of certain population groups.

Fowler noted that one of the limits of the interaction approach was the inability to detect questions that people answer without difficulty (as indicated by requests for clarification) even though they may not understand the question or may choose not to answer honestly. And the point was made by another participant that evaluation of questions should go beyond issues of clarity and clutter to the symbolic framework respondents use in responding.

It was pointed out that question response is not simply a matter of what people can remember. Respondents may feel constrained if their opinion is not widely shared or socially desirable; or their response may be affected by the way a particular issue relates to other concerns in their lives. Researchers may frame questions in a particular context, such as mental health; but respondents may not frame their answers in the same context. Issues do not stand alone but are tied to family relationships, job responsibilities, ethnic group identification, etc., all of which may come into play in one's response. Framing cognitive issues only in terms of the clarity of the question to the respondent is too narrow a perspective for examining this issue. Tourangeau (1984) proposes several stages in the cognitive process of answering questions, a final stage being a judgment call on the part of the respondent about whether or not to provide accurate information. Noelle-Neumann (1984) also addresses the significance of the cognitive framework in respondent behavior. These issues become even more important as researchers attempt to interview populations outside the mainstream, such as persons at risk for AIDS, the frail elderly, and ethnic populations for whom questionnaires may have to be translated.

Other procedures for evaluating questions focus on more formalized validity and reliability analyses (Koons, 1973 discusses the difference between reliability and validity). The observation was made that in methodological work on health surveys, the focus has been more often on validity rather than reliability analyses. Careful thought needs to be given in reliability studies for health surveys about the appropriateness of traditional reliability analysis for the items of interest, such as the use of test-retest methods relating to acute medical care episodes. Real variance must be separated from error variance in these types of analyses. In addition, whether the focus of the research is on descriptive estimates or on the relationships among variables must be considered. More variability around point estimates may be tolerable if the direction of the relationships among the major variables of interest is not altered.

Comments about the Kulka paper pointed out that validity analyses conducted with clinic patients may not be generalizable to community populations. Kulka underlined that in the mental health measurement area, the testing of different methods with different populations at different periods of time is necessary to fully evaluate the generalizability of validity analyses of those measures. Another aspect of validity analysis raised in the discussion was that the methods of measuring con-

struct validity are not well developed. (See Validity of Reporting in Surveys section in this volume for further discussion of validity and appropriate standards against which to measure it.)

Considerations for Further Research

Future methodological research in evaluating questions revolves around four major issues which are summarized below.

1. Many of the problems in evaluating questions regarding respondent understanding or interpretation are not new. Intensive interviewing appears to offer insights into certain aspects of the cognitive process of responding to questions. However, information about approaches and techniques is ad hoc and not systematically shared, in part because the intensive interview process is often viewed simply as one of many steps internal to the questionnaire development process. In addition, a critical review of lab techniques is absent. The value of intensive interviewing for health survey methods in general would be enhanced by the following:

- Conducting systematic randomized studies that compare the costs and benefits of alternative intensive interviewing techniques.
- Documenting and disseminating the results of different intensive interviewing techniques. (One option would be to conduct a survey of labs, which approaches they use, and for what types of studies. In addition NCHS might consider a working paper series on findings from the QDL.)
- Refining methods for measuring systematic error or bias as well as variable error that use the different intensive interviewing approaches.

2. Evaluating questions in pretests which stimulate real interview conditions remains an important mechanism for evaluating how well questions work. The value of pretests may be enhanced by introducing more objective means of question evaluation that go beyond interviewer impressions of what works and what does not. Survey methodologists are aware that people may answer questions they do not really understand (sometimes with no apparent difficulty) and that the cognitive frameworks people use in assessing and answering questions affect their answers and may vary across individuals. However, the methods for identifying and dealing with these issues in practice are not well developed. Evaluating question clarity and adequacy in pretests may be improved in the following ways:

- Using the approach proposed by Fowler more widely. Given the relative ease and low cost of administration, it offers the possibility of beginning to build a knowledge base about behavior of questions across different populations.
- Exploring ways to combine intensive interviewing and the quantification of respondent behavior in study pretests.
- Attending to the issues raised by cognitive frameworks as they affect survey response and exploring empirical approaches to these issues.

3. The validity and reliability of specific scales, instruments, or theoretical constructs (such as health) remain important areas of formally evaluating survey questions (see Kaplan and others, Kulka, and Hembroff, respectively, in this volume). To date there has been more emphasis on validity than reliability analysis. In addition, the standards for determining validity—often clinical or record data—are themselves increasingly called into question. Such questions could be addressed by the following:

- Conducting more extensive reliability analyses of health survey items, particularly those that are in greatest use.
- Synthesizing and integrating the reliability and validity analyses that have been conducted to date for major health survey question items and scales. (This could serve as a first step toward establishing standards in the field as proposed by McDowell.)
- Evaluating the appropriateness of applying the results of these analyses across varying subgroups of the population.

4. The session offered a clear contrast between researchers who study measurement error properties of questions by comparing alternative question wording as opposed to those who measure error through models of the error properties of a given set of questions. (The papers by Royston, Fowler and others, and Kaplan and colleagues illustrate the former; those by Kulka and Hembroff, the latter.) The former use split samples or alternative wordings as a procedural strategy. The latter use psychometric measurement error models to estimate the magnitude and direction of error and generally do not attend to the source—linguistic, cognitive, or other—of the error. If these two groups came together, it would increase the feasibility of additional studies on model-based estimates of validity that compare different question wordings (Andrews, 1984; also see Groves, in this volume, for further amplification of these different perspectives on measurement error).

References

Andrews, F.M. (1984, Summer). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly* 48(2), pp. 409-442.

Belson W. (1981). *The design and understanding of survey questions*. London: Gower.

Bradburn, N.M., Sudman, S. and associates. (1979). *Improving interview methods and questionnaire design*. San Francisco, CA: Jossey-Bass.

Koons, D.A. (1973, March). *Quality control and measurement of nonsampling error in the health interview survey* (DHEW Publication No. (HSM) 73-1328). National Center for Health Statistics Series 2, No. 54.

Laurent, A. (1972, June). Effects of question length on reporting behavior in the survey interview. *Journal of the American Statistical Association* 67(338), pp. 298-305.

Noelle-Neumann, E. (1984). *The spiral of silence*. Chicago, IL: University of Chicago Press.

Royston P., Bercini, D., Sirken, M. and Mingay, D. (1986). Questionnaire Design Laboratory. In *American Statistical Association: Proceedings of the Section on Survey Research Methods*, American Statistical Association, (pp.703-707). Alexandria, VA: American Statistical Association.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T.B. Jabine and others (Eds.) *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: National Academy Press.

SESSION 2

Validity of Reporting in Surveys

Introduction by Floyd J. Fowler, Jr.

This session's feature papers address the topic of how reports from surveys correspond with data on the same topics derived from records. Historically such studies have been thought of as validating, or evaluating, the quality of survey data. Implicit in this concept is the notion that data from social surveys are error prone and that record data can be used as a gold standard against which to measure survey data.

Evidence has shown that there is measurement error in surveys. Questions are asked that respondents do not understand. Respondents are asked questions that require information they do not have or cannot recall in detail. Respondents are not willing to answer all questions accurately. However, records also have limitations. Record systems are maintained for some purpose which seldom includes providing the kinds of data researchers seek. Furthermore, record data are subject to sources of error parallel to those of surveys. The sample of people for whom record data are available may not constitute or represent the population the researcher wants to describe. Or there may be errors in the records from inconsistent interpretation of what was to be included, from inconsistent patterns of recording information, or from systematic distortion in the entry process.

This session's presentations emphasize that when comparisons of record and survey data are made, differences are found. However more often than not, the conclusions are not so much about the validity of survey reporting as about how the data are collected and about whom they are collected affect comparability of the estimates, whether from records or from surveys.

FEATURE PAPER

Results of the National Medical Expenditure Survey Household Survey Medicare Record Component Pretest

Kathleen A. Calore and Jiuan Lim

Introduction

The 1987 National Medical Expenditure Survey (NMES) was authorized to provide up-to-date and reliable estimates of use and expenditures for health care services as well as the extent of insurance coverage. The NMES data will be used for estimations for the civilian noninstitutionalized population as well as for the population residing in nursing homes and institutions for the mentally retarded. This study, the Medicare Record Component (MRC) of National Medical Expenditure Survey was designed to supplement the household and institutional surveys by incorporating actual reimbursement data for sampled persons who are Medicare eligible. Perhaps more importantly, the Medicare Record Component also provides a unique opportunity to validate the use and expenditure information provided by survey respondents with actual Medicare expenditure records.

In addition to evaluating respondents' recall of hospitalizations, the study illustrates how well respondents understand:

- What type of care was provided?
- How many providers were involved?
- Who billed for care?
- How much did it cost?
- Who paid for it?

The MRC's objective is to design a systematic methodology to match the NMES-reported health use with Medicare administrative data (claims) from the Medicare Automated Data Retrieval System (MADRS). The MRC pretest includes a sample of respondents from three components of the NMES survey, the Household Survey Component (HHS), the Institutionalized Population Component-Current Residents (IPC-CR), and the Institutionalized Population Component-New Admissions (IPC-NA). This paper discusses the results of inpatient hospital matches for the Household Survey Component of National Medical Expenditure Survey.

Medicare Record Component Pretest Sample

The MRC sample (Table 1) includes 1,700 beneficiaries (both users and nonusers of health services), 700 Medicare beneficiaries from Household Survey Component and 500 each from the Institutionalized Population Component-Current Residents and the Institutionalized Population Component-New Admissions. Two groups in Household Survey Component were oversampled: the very old (over age 80) and those under age 65 who are primarily the disabled Medicare eligibles.

Medicare eligibility of respondents in the HHS and IPC subsamples was verified and a tape of Health Insurance Claim (HIC) numbers submitted to the Health Care Financing Administration (HCFA), who then provided a tape with all MADRS claims for these selected beneficiaries.

Comparison of NMES and MADRS Input Files

Before discussing the details of the inpatient match, it is important to review the purpose and design of the two sources of data in order to better understand the matching results. The principal focus of National Medical Expenditure Survey is to collect information about each occasion of health care utilization. The purpose of the Medicare Automated Data Retrieval System claims, on the other hand, is to document Medicare expenditures (payments) for health care services. These divergent goals produce different definitions of health care events. For example, National Medical Expenditure Survey asks for all inpatient stays (including same-day admission and discharge), which may include same-day surgeries or tests that are billed as outpatient rather than inpatient services. The different purposes of the two matching files present more serious concerns for the

Kathleen A. Calore and Jiuan Lim are with Health Economics Research, Inc., Needham, Massachusetts.

This work was performed under subcontract to WESTAT under Contract No. 282-86-0013 from the National Center for Health Services Research and Health Care Technology Assessment.

Table 1. Sample sizes for pretest of Medicare Records Component of the National Medical Expenditure Survey

Component and age	Verified HIC numbers	Pretest sample size
Household survey component		
Under 65	254	54
65-74	2,210	334
75-84	1,556	232
85+	420	80
Total	4,440	700
Institutionalized population component—current residents		
Under 65	177	37
65-74	287	60
75-84	735	154
85+	1,189	249
Total	2,388	500
Institutionalized population component—new admissions		
Under 65	42	22
65-74	147	79
75-84	355	190
85+	391	209
Total	935	500
Grand total	7,763	1,700

SOURCE: HER NMES MCR Pretest file, 1987.

match of outpatient services than for the admission matches.

At the outset of the discussion of the inpatient match results, a modest difference in the number of inpatient stays identified within the two sources of data (Table 2) is noted. The HHS respondents identified 293[1] admissions during 1987; while in Medicare Automated Data Retrieval System, there are 305 hospital stays for our sampled beneficiaries. Of these, 258 are MADRS inpatient hospital claims, and another 47 are cases where a physician billed for at least one inpatient service but no hospital claim was found.

The objective of the MRC pretest is to define a matching algorithm that would allow for the maximum number of correct matches while minimizing the possibility of incorrect matches. If respondent recollection were perfect and claims exactly mirrored utilization, the matching of Medicare claims to survey responses would be a straightforward task. However, respondents may err in a variety of ways:

- by reporting a non-Medicare covered service;
- by failing to identify all providers involved in providing a service;
- by reporting an incorrect date (wrong day or month);
- by misidentifying the type of service provided (that is, confusing outpatient surgery with an inpatient stay);

[1]The sample included 62 non-full-year respondents (those not responding to all four rounds of National Medical Expenditure Survey), accounting for 5.1 percent of possible person rounds.

Table 2. Description of Medicare Automated Data Retrieval System and National Medical Expenditure Survey matching files—Household Survey Component

Inpatient	MADRS	NMES
Hospital stays	305	293[a]
Hospital claims	258	—
Physician claim without hospital claim	47[b]	—
(related physician claims)	1039[c]	—

[a]While NMES included 293 unique inpatient (STAZ) records, 15 of these events were 1 day (no overnight) episodes that did not match an inpatient claim. The 1-day stays were subsequently incorporated into the outpatient match file.
[b]In this case, a physician billed for inpatient services but no hospital bill was submitted.
[c]Related claims are for physician services identified as related to a hospital claim.
SOURCE: HER NMES MRC Pretest file, 1987.

- by separately reporting a service provided within a global fee; and
- by omitting a service.

The MADRS claims data also introduce several potential sources for error. Most importantly, claims for hospitalizations may be finalized well after the discharge date. For example, claims for hospitalizations during the last few months of 1987 might not be paid until March 1988. Hence, requests for claims data should specify a sufficient time after discharge to ensure that most claims have been finalized.

In addition, coding errors within the claims data may incorrectly identify services as occurring during a hospital stay. This latter problem probably accounts for some of the 47 cases in the matching file with physician claims but no hospital claim.[2] However, it is also possible that some Part A claims were not finalized until much later, and hence were not included in the file.

HHS Inpatient Matching

Rules for Household Survey Component Inpatient Matching

Three possible matching criteria are contained in some form on both matching files:

- Dates of hospitalization are included in both National Medical Expenditure Survey and Medicare Automated Data Retrieval System.
- Cost is included on National Medical Expenditure Survey in an extensive series of questions concerning the costs associated with all providers as well as the distribution of cost among payers. However, only 30 percent of the NMES Medicare hospitalizations had at least one cost record associated with it. Since the purpose of Medicare Automated Data Retrieval Service is to record Medicare expenditures, it represents a complete record of Medicare costs for all providers. Due to the lack of complete NMES expenditure information, cost was not included as a matching criterion.

[2]In constructing other merged Part A-B merged data bases, we have found a small number of similar cases.

- Condition (reason for admission) is included in National Medical Expenditure Survey as text describing what the physician told the patient. The MADRS file includes ICD-9-CM diagnoses and procedure code(s) as well as Diagnosis Related Group (DRG).
- Hospital identifiers are included in both National Medical Expenditure Survey and Medicare Automated Data Retrieval System although there is presently no crosswalk to link the two files.

For the main study the pretest matching algorithm, which is dependent on date(s), may be expanded to include condition and hospital identifier when these elements are transformed into comparable fields.

The MRC matching algorithm is comprised of seven passes at the two input files. Pairs of observations that match in one pass are not included in subsequent passes. Only observations that fail to match are included in the input file for the next pass at matching. The following describes the rules used in each of the seven passes:

Pass one: Begin (admission) date on NMES exactly matches begin (admission) date on MADRS.

Pass two: End (discharge) date on NMES exactly matches end (discharge) date on MADRS.

Pass three: Begin (admission) month and length of stay (LOS) on NMES matches begin (admission) month and LOS on MADRS.

Pass four: End (discharge) month and LOS on NMES matches end (discharge) month and LOS on MADRS.

Pass five: Begin (admission) month on NMES matches begin (admission) month on MADRS.

Pass six: End (discharge) month on NMES matches end (discharge) month on MADRS.

Pass seven: Selecting only cases where there is one NMES observation and one MADRS observation, consider a match if begin (admission) date on MADRS is within 30 days (either before or after) of begin (admission) date on NMES.

Results of the HHS Inpatient Matching

Using the seven rules described above, a total of 225 (76.8 percent) of the NMES inpatient admissions and 218 (71.5 percent) of the MADRS admissions were matched. The match rate for MADRS hospital claims (Part A) was quite high (77.9 percent), especially when compared with the rate for the group of admissions where there was a physician claim but no facility bill (41.4 percent).

A review of the match rate for each rule (Table 3) shows that the majority of NMES cases are matched using exact begin (admission) date in pass one (48.5 percent). An additional 5.1 percent of cases match on exact end (discharge) date. Using exact month of admission (begin month) matches another 12.6 percent of NMES reported admissions, while month of discharge (end month) matches another 5.5 percent. Since more admissions were identified for the sampled beneficiaries in Medicare Automated Data Retrieval System, those match rates are slightly lower. In pass one, using begin (admission) date, 45.2 percent matched, while another 4.9 percent matched using end (discharge) date. Another 11.5 percent were matched using the admission month rule and 4.9 percent matched by end (discharge) month. Pass seven matches only those unmatched respondents after pass six with a single NMES and a single MADRS claim. Of the 18 respondents meeting this criteria 5 more admissions are matched.

While Medicare Automated Data Retrieval System may not be the gold standard to assess the validity of NMES-reported utilization, it is the best available measure. Because of its limitations, Medicare Automated Data Retrieval System is offered as a gold-plated standard by which to evaluate how well National Medical Expenditure Survey reflects actual utilization. Hence, the latter match rates are the more meaningful measures of how well NMES responses reflect actual utilization. It is possible that by including admissions identified by

Table 3. Results of National Medical Expenditure Survey Household Survey Component inpatient match

		Percent NMES match rate	NMES	Percent MADRS match rate	MADRS
Pass 1:	Match begin date	48.5	142	45.2	138[a]
Pass 2:	Match end date	5.1	15	4.9	15
Pass 3:	Match begin month and LOS	1.7	5	1.6	5
Pass 4:	Match end month and LOS	1.7	5	1.6	5
Pass 5:	Match begin month	12.6	37	11.5	35
Pass 6:	Match end month	5.5	16	4.9	15
Pass 7:	Match within 30 days for cases with only one MADRS and one NMES observation	1.7	5	1.6	5
Matched		76.8	225	71.5	218
Unmatched		23.2	68	28.5	87
Total		100.0	293	100.0	305

[a]The MADRS claims have been matched to more than one NMES event where the number of MADRS claims is less than the number of NMES events. Multiple matches occur when the matching criteria are met in more than one NMES event.
SOURCE: HER NMES MRC Pretest file, 1987.

physician claims only within the matching file we have overestimated the actual number of admissions in Medicare Automated Data Retrieval System. If none matched, it would be appropriate to drop them from the matching. However, 17 of the 47 did match (36.2 percent), so it remains unclear whether they should be dropped for the main study. The event level matching results also suggest that the majority of HHS respondents who reported inpatient services recalled these episodes of hospitalization quite well (53.6 percent provided either the exact admission or discharge date).

Further analysis of the unmatched hospitalizations in both files shows the following:

- Of the 81 unmatched NMES hospital stays, 50 (61.7 percent) are for respondents with no residual MADRS inpatient claims. Another 12 (14.8 percent) with a zero length of stay were subsequently matched to an outpatient claim.
- Of the 43 people with unmatched MADRS hospital (Part A) claims (representing 60 admissions), 27 (representing 37 admissions) had no residual NMES hospitalizations.

Hence additional efforts at matching would be targeted at 19 NMES reported events and 23 MADRS hospital claims.

Analysis of the HHS Timeline

The above discussions, however, do not tell us anything about how well aggregate inpatient utilization levels based on NMES HHS data compare with the utilization observed in the MADRS claims. In addition, event level matching at the admission level does not provide a measure of agreement between the two sources.

To address these concerns, two timelines were constructed for each beneficiary; one based on the inpatient days reported during 1987 on National Medical Expenditure Survey (labeled "S") the other based on days billed to Medicare for 1987 (labeled "M"). The timeline is essentially a single variable with 365 characters (the number of days in the year). Days when the beneficiary is reportedly in the hospital take a value of 1. Days when the beneficiary is not in the hospital take the value 0. A comparison of the two timelines identifies the frequency of exact agreement in hospital days reported, and further indicates whether allowing more tolerance around hospitalization dates would substantially improve the match rate.

Two examples of the timeline are shown in Table 4. Each line represents a quarter of 1987, the line labeled "M" is the utilization reported on Medicare Automated Data Retrieval System, while that labeled "S" is the NMES reported use. In the first example there is exact agreement between the two sources, for admissions in quarters three and four. While the two sources frequently agree, this approach also allows us to see when they are at variance. The next example illustrates several scenarios: exact agreement, days where only NMES reports use, as well as days where only Medicare Automated Data Retrieval System reports use.

How can we use the picture of agreement between the two sources shown in the timeline to derive a discrete measure of agreement? Two questions are considered.

- How well do respondents recall the extent of inpatient care received? (That is, how do total inpatient days reported in the two sources compare?)
- How well do respondents recall the exact dates of hospitalization?

Comparison of HHS Utilization Levels Reported

Because National Medical Expenditure Survey is used as the basis for estimation of inpatient admission rates

Table 4. Example of timeline

	PERSID-01807012	
M	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	Day 1-90
S	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	
M	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	Day 91-181
S	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	
M	00000000000000000000000000000000000000000000000000000000000000000001111000000000000000000000	Day 182-273
S	00000000000000000000000000000000000000000000000000000000000000000001111000000000000000000000	
M	00000000000000000000000000000000000000000000000000000000000000000000000011111111111111111111	Day 274-365
S	00000000000000000000000000000000000000000000000000000000000000000000000011111111111111111111	
	PERSID-05616012	
M	111000000000000000000000000000000000000000000000000000000000000000000000000000000000000011	Day 1-90
S	111000000000000000000000000000000000000000000000000000000000000000000000000000000000000011	
M	1111100000000000000000000000111111110000000001111111111111111111100000000000000000000000000	Day 91-181
S	1111110000000000000000000000111111110000000011111000000000000000000000000000000000000000000	
M	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	Day 182-273
S	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	
M	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	Day 274-365
S	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000	

SOURCE: HER NMES MRC Pretest file, 1987.

Table 5. Comparison of National Medical Expenditure Survey and Medicare Automated Data Retrieval System inpatient utilization Household Survey

	NMES inpatient days	MADRS inpatient days	Days underreported in NMES	Percent underreported in NMES
Quarter 1				
January-March	582	664	82	12.4
Quarter 2				
April-June	787	824	37	4.5
Quarter 3				
July-September	570	677	107	15.8
Quarter 4				
October-December	408	546	138	25.3
Annual total	2347	2711	364	13.4

SOURCE: HER NMES MRC Pretest file, 1987.

and inpatient days, comparisons of reported utilization is an important aspect of the timeline analysis. In Household Survey Component there are 185 sampled persons (out of 700) who reported either NMES inpatient utilization or had MADRS claim(s) for inpatient services. The comparison of aggregate inpatient utilization reported in the two sources (Table 5) shows that NMES respondents underreport inpatient days (as measured by Medicare Automated Data Retrieval System) by 13.4 percent.

To address the issue of how well respondents recall the exact dates of hospitalization, 130 respondents were analyzed who had both NMES reported stays and MADRS inpatient claims activity during the year. A comparison of the dates reported as inpatient days between the two sources is used to derive an agreement rate. This is defined as the sum of dates which both sources identified as part of an inpatient stay as the numerator, and the total number of inpatient days reported in Medicare Automated Data Retrieval Service as the denominator. Agreement is measured by the percentage of inpatient use where the respondent correctly identified the exact dates (Table 6).

Given the stringency of the rule (that is, the dates must coincide exactly), the agreement rates are quite high. However, it is important to remember that this is not a random group of respondents. To be included, a respondent had to report at least one NMES inpatient stay and have at least one MADRS inpatient hospital claim. The results indicate that if the respondents report a hospital stay, they also recall the precise dates nearly 60 percent of the time, a finding consistent with our event level match.[3]

By quarter within the year, the agreement rates vary somewhat. In the first two quarters they are significantly higher than the rates in the later quarters. Some of the shortfall, particularly in the final quarter, is due to possible underreporting of MADRS days. For the main study the recommendation is that the data request include all claims with a date of service in 1987, whether paid during 1987 or 1988. Contrary to expectations, relaxing the exact one-to-one date rule for stays (that is, to 15 days or 30 days) does not substantially improve the measured agreement. If a respondent reports the stay, it is likely that the actual dates are reported.

Table 6. Summary of Household Survey component agreement

	Total days that agree	Reported MADRS days	Agreement in percent[a]
Quarter 1			
January-March	422	664	63.6
Quarter 2			
April-June	551	824	66.9
Quarter 3			
July-September	372	677	54.9
Quarter 4			
October-December	260	546	47.6
Annual total	1605	2711	59.2

[a]Agreement is defined as the sum of days during the year that both sources have identified as part of an inpatient stay expressed as a percent of total MADRS utilization.
SOURCE: HER NMES MRC Pretest file, 1987.

Lessons from the MRC Pretest

The MRC pretest for the household succeeded in matching 71.5 percent of the MADRS hospital claims with a NMES reported inpatient event, while finding a match for 76.8 percent of the NMES reported events. Further, respondents correctly identified either the admission or discharge date in 50.1 percent of cases. Although the dates of hospitalization were the only matching variables available for the pretest, the quality of the matches seems quite high. Comparisons by hand of selected data elements belonging to the matched pairs, including condition (reason for hospitalization) and hospital name, confirm the matches made by the date algorithm.

The detailed timeline analysis of the exact dates of hospitalization reported on National Medical Expendi-

[3]The agreement rate is affected negatively only when the respondent fails to report a day of hospitalization. Overreporting days of hospitalization does not factor into the agreement rate.

ture Survey supports the findings of the event level matching. When the respondent reports a stay and also has an inpatient claim, 60 percent of the reported days coincide with the days on the hospital claim. The timeline analysis permits us to assess whether allowing tolerance around the dates of a hospitalization reported on National Medical Expenditure Survey would improve the match. Based on our findings, expanding the interval to include up to 30 days does not improve the match rate.

FEATURE PAPER

Validating Reporting of Usual Sources of Health Care

Janet D. Perloff and Naomi M. Morris

Introduction

Having a usual source of health care has long been considered an important indicator of access. Numerous investigators have observed that individuals with a regular source of care are more likely to use care when illness arises, seek preventive services, and express greater satisfaction with health care (Aday & associates, 1980). It has also been observed that different racial and economic subgroups tend to identify different settings as usual sources of health care and that the type of place named as the usual source of care correlates with patterns of use as well. Perhaps the most notable example is the greater reliance on hospital outpatient departments and emergency rooms by the poor, nonwhites, persons with Medicaid coverage, and the uninsured (Dutton, 1978; Kasper 1987; Walden & Rossiter, 1979; Kasper & Berk, 1980; Short & associates, 1985). The interests of health services researchers have not been limited to the influence on access of having a usual source of care: the type of place and associated patterns of health care use, insurance coverage, and costs have been of great interest as well.

Health care surveys are the basis for most information about the type of places named as the usual source of care.[1] Little is known, however, about the accuracy with which survey respondents characterize the type of place they usually visit for health care. This paper reports results of an investigation into the accuracy with which survey respondents characterize these settings. It also describes the strategy used to validate survey respondents' characterizations of their usual source of care, presents the findings, and discusses some of their implications for future health care surveys.

Brief Description of the Study Methods and Sample

This research was conducted as part of a study of patterns of health care use among low-income black women in three inner-city Chicago communities. The sampling design, data collection instruments, and field work were developed and carried out in collaboration with the University of Illinois Survey Research Laboratory. A brief overview of the research methods is presented in Table 1 and a complete description of the methods is available elsewhere (Perloff & Morris, 1989). For present purposes it is important to emphasize that this validity check was carried out using survey responses supplied by a sample of 302 relatively disadvantaged inner-city women. As Table 2 indicates, most of the respondents are poor, not well educated, single, and unemployed and almost all are either covered by Medicaid or are uninsured.

Survey Items and Sample

Table 3 shows the three survey items used to conduct this validity check and the frequency distributions of

Janet D. Perloff is with the Graduate School of Public Affairs and the School of Public Health, State University of New York at Albany. Naomi M. Morris is with the School of Public Health, University of Illinois at Chicago.

This research is funded in part by a grant from The Chicago Community Trust. The authors gratefully acknowledge the research assistance of Susumu Kudo, Kathleen Thoma, and Cheryl Wiseman. They also thank Karen Burke, Gretchen Fleming, James Fossett, and Stephanie McFall for comments on an earlier draft.

[1]Items measuring the type of usual source of health care have been included in many health care surveys. See, for example, instruments used in the periodic National Center for Health Statistics Health Interview Surveys; the 1975-76 Survey of Access to Medical Care conducted by the Center for Health Administration in collaboration with the National Opinion Research Center; the National Center for Health Statistics and National Center for Health Services Research 1977 National Medical Care Expenditure Survey; and the Robert Wood Johnson Foundation 1986 Access Study.

Table 1. Overview of survey methods

Objective: to document health care use among low-income black women in three inner-city Chicago communities

Sampling design:

Three communities: Austin, Near West Side, West Garfield Park

Within communities: randomly selected blocks within randomly selected low-income census tracts

Within blocks: black women, aged 18-45, pregnant at the time of the survey or with a child under 6 in the home

Survey methods:

Door-to-door canvassing and screening

Personal interviews by interviewers of the same sex and race averaging 36 minutes in duration

Total cooperation: 86 percent

responses to each item. Question 1 was a variant of a commonly asked health survey question: "Is there a place you usually go for your health care, that is, if you are sick or want advice from a doctor or nurse about your health?" Table 3 indicates that 20 respondents had no such usual place; these cases were excluded from further analysis.[2]

In the second question, also a variant of a commonly asked health survey question, the 282 respondents reporting a usual source were asked to indicate the type of place they usually visit. "Where do you usually go? Is it to 1) a private doctor's office; 2) a hospital outpatient clinic; 3) a health department clinic; 4) an emergency room of a local hospital; 5) some other clinic not connected with a hospital, or (6) some other place (specify)?" Responses to this item represent unvalidated self-reports of the type of the usual place. Table 3 indicates that due to insufficient information, 10 of the 282 responses to Question 2 could not be validated.

Question 3, not often included in health care surveys, asked the 282 respondents with a usual source: "What is the name and address of the place you usually go?" In 275 cases a place was named, and in 265 cases an address was provided. Some respondents gave only a name; other respondents gave only an address.[3] We were able to identify the usual source of care named by 281 cases. When the 10 cases with insufficient responses to Question 2 were eliminated, 271 cases remained that could be included in the validity check.

[2]Although we had 302 respondents to our survey, we could validate only 271 cases. Twenty were excluded because the had no usual source of care. We could identify usual health care places (from Question 3) for 281 of the remaining 282 cases. However, due to insufficient responses to Question 2, another 10 cases also were eliminated. Two of these 10 could not be validated because no answer was given to Question 2, and 8 additional cases were excluded because insufficient information to validate the response had been collected in Question 3: four cases responded "other" to question 2 but there was no comparable category for Question 3, and 4 cases responded "hospital emergency room" but Question 3 did not require that the respondent specify if they specifically visited an emergency room.

[3]When respondents gave only a name, telephone books and other rosters were used to more fully identify the place. When respondents gave only an address, efforts were made to identify these places by checking various rosters and, when otherwise unidentifiable, visiting the locations specified.

Table 2. Characteristics of survey respondents (N=302)

—Average age of 27 years
—Two-thirds have never been married
—77 percent obtained some high school education
—38 percent completed grade 12
—82 percent are unemployed
—76 percent are receiving public assistance
—90 percent are covered by Medicaid
—8 percent are uninsured

Approach to Validation

We used the names and addresses supplied by respondents in Question 3 to independently determine and classify each respondent's usual source of health care as either a private doctor's office, a hospital outpatient department, a public health clinic, or some other clinic not connected with a hospital. The validity check was accomplished by comparing the category indicated by re-

Table 3. Survey items used in validity check

Question 1: Is there a place you usually go for your health care, that is, if you are sick or want advice from a doctor or nurse about your health?

Response categories	N
Yes	282
No	20
Total	302

Question 2: Where do you usually go? Is it to . . .

Response categories	N
A private physician's office	54
A hospital outpatient clinic	151
A health department clinic	22
An emergency room of a local hospital	4*
Some other clinic not connected with a hospital or	45
Some other place?	4*
No answer	2*
Total	282

Question 3: What is the name and address of the place you usually go?

Response categories	N
Named a place	275**
Did not know	6
Gave no answer	1
Total	282
Gave an address	265**
Did not know	7
Gave no answer	10
Total	282

*The 10 responses in these categories could not be validated.
**Using names and/or addresses, usual places were identified for a total of 281 cases.

Table 4. Overall results of validity check (N=271)

Results	N	Percent
Correctly classified responses	166	61
Incorrectly classified responses	105	39

spondents (that is, unvalidated Question 2) with the category independently determined and assigned by the investigators to each reported usual place (that is, validated Question 3).

Various secondary sources were used to independently determine the type of usual place named by each respondent. Hospitals were identified using the most recent *American Hospital Association Guide*. Health department clinics were identified with a roster obtained from the Chicago Department of Health. Other clinics not connected with a hospital included the 13 Chicago clinics listed on a roster obtained from the Illinois Primary Health Care Association, an association of the not-for-profit, federally funded community and neighborhood health centers in Illinois. The remaining places included private office-based practices of various types including solo practitioners, single specialty groups, and multispecialty groups. Since Illinois enrolls some Medicaid-eligibles in Health Maintenance Organizations (HMOs), these private practices included both closed panel HMOs as well as office-based physicians participating in Independent Practice Association (IPA)-type Health Maintenance Organizations.

Even in this small, geographically finite sample, there was unexpected diversity in the places identified as respondents' usual sources of health care. The literature suggests that low-income women such as these would name a small number of large hospitals and public clinics as their usual sources of care. Rather, the 281 respondents named 86 unique health care places. The 10 most frequently mentioned health care places (primarily hospital outpatient clinics and community health clinics) accounted for only 145 of the 281 responses. The remaining 136 cases named 76 additional places (which were primarily private physician's offices).

Table 5. Comparison of percent distribution based on unvalidated Question 2 with percent distribution based on validated Question 3 (N=271)

Type of Place	Unvalidated Question 2 (%)	Validated Question 3 (%)	Difference (%)
Private physicians office	19.93	45.76	−25.83
Hospital outpatient department	55.35	42.07	+12.28
Health department clinic	8.12	3.32	+ 4.80
Other clinic not connected with hospital	16.61	8.86	+ 7.75

Results

When answers to Question 2 were compared with an independent check of the places named in Question 3, 39 percent of respondents were found to have incorrectly classified their usual health care place (Table 4). Table 5 compares the distribution of respondents' classification of their usual place with the distribution resulting from the validated data. Reliance on the respondents' characterizations of their usual place would understate by 26 percent their use of private physician's offices as usual sources and overstate their use of hospital outpatient departments, health department clinics, and other clinics not connected with hospitals.

Table 6 cross-classifies answers to unvalidated Question 2 by validated responses developed from answers to Question 3. The percentage in each cell shows the relationship of the cell value to the column total, or the percent distribution of validated responses (based on Question 3) by unvalidated responses (based on Question 2).

Respondents whose usual source of care is a hospital outpatient department generally know the type of place they visit and classify it accurately: 92 percent of those visiting a hospital outpatient department correctly classified it as such. By contrast, other clinics not connected with a hospital were correctly classified by only 29 percent of the respondents visiting such places. The data in the fourth column of Table 6 indicate considerable confusion between the other clinic category, the hospital outpatient department category, and the health department clinic category: 46 percent of respondents using other clinics misclassified their usual place as a "hospital outpatient department" and 21 percent of these respondents misclassified them as a "health department clinic."

Private physicians' offices also were likely to be misclassified. The first column of Table 6 indicates that 124 respondents actually named a private physician's office as their usual source of health care. However, only 38.7 percent of these respondents correctly classified their usual place as a private physician's office. Many respondents using private physicians were likely to confuse the places they visited with hospital outpatient departments and other clinics: 26 percent of respondents using private physicians misclassified their usual place as a "hospital outpatient department" and 28 percent misclassified their usual place as a "clinic not connected with a hospital."

From these findings more can be learned about the characteristics of places for which validity appeared low. For example, were particular private practices frequently mistaken for hospital outpatient departments and, if so, why might this be? To gain this insight the extent to which each multiply named place had been placed by respondents into more than one category of Question 2 was examined, because this might indicate that this type of place was not particularly well understood. Table 7 summarizes this analysis and provides

Table 6. Percent distribution of unvalidated responses to Question 2 by validated responses to Question 3 (N=271)

Unvalidated responses to Question 2	Validated responses to Question 3			
	Private physician's office (N=124)	Hospital outpatient clinic (N=114)	Health department clinic (N=9)	Other clinic (N=24)
Private physician's office	38.7	4.4	0.0	4.2
Hospital outpatient clinic	25.8	92.1	22.2	45.8
Health department clinic	7.3	1.8	66.7	20.8
Other clinic	28.2	1.8	11.1	29.2
Total	100.0	100.1*	100.0	100.0

*Total exceeds 100 % due to rounding

examples. Three types of places were most likely to be multiply classified: neighborhood and community health centers, large group practices and IPA participants, and closed panel Health Maintenance Organizations.

Upon closer examination it is somewhat easier to understand some of the possible sources of respondents' confusion. For example, neighborhood and community health clinics such as the Mile Square Health Center were rarely mistaken for private physicians' offices (Table 7). However, perhaps because they are relatively large and institutional in appearance and operation, such clinics were likely to be mistaken for hospital outpatient departments and health department clinics.

Similarly, large multispecialty group practices such as the Liberty Medical Center were not often considered by respondents to be private physicians' offices. Instead, there was a tendency for respondents to consider these to be hospital outpatient departments or other clinics not connected with a hospital. These private practices are also large and clinic-like and tend to call themselves medical centers or clinics. Thus, it is not so surprising that respondents would consider them to be clinics. The confusion of these settings with hospital outpatient departments is somewhat more difficult to understand. The practices are not located in or especially close to hospitals and, although specific physicians may have hospital affiliations with nearby hospitals, these private practices are not noticeably identified with particular hospitals. Respondents may believe these places resemble hospitals or may perceive some hospital affiliation not readily apparent to the investigators. Alternatively, the response categories may not be well understood.

Finally, a large, closed panel Health Maintenance Organization, the Anchor HMO, was often confused with a hospital outpatient clinic. Anchor is owned and operated by and very closely identified with one of the city's large hospitals, Rush-Presbyterian-St. Luke's. Moreover, the Anchor facility used by most women in our sample is located only two blocks from the hospital. Thus, it is easy to understand why respondents may confuse the Anchor facility with a hospital outpatient clinic.

Since all respondents to the survey were inner-city, low-income women, the extent to which members of other socioeconomic groups have similar problems characterizing their usual sources of care is unknown. However, the tendency of particular categories of respondents within this homogeneous sample to misclassify the type of place they usually go for care was examined. Factors such as age, education, income level, insurance coverage status, and health status showed no significant association with correctly or incorrectly classifying the type of place they usually go to for health care.

Table 7. Types of usual health care places likely to be multiply classified

1. Neighborhood and community health clinics
   Example: Mile Square Health Center (N=12)
   - 3 classified correctly (other clinic not connected with a hospital)
   - 4 classified as a hospital outpatient clinic
   - 4 classified as a health department clinic
   - 1 classified as other
2. Large group practices and IPA participants
   Example: Liberty Medical Center (N=18)
   - 2 classified correctly (private physician's office)
   - 4 classified as a hospital outpatient clinic
   - 1 classified as a health department clinic
   - 8 classified as other clinic
   - 1 classified as other
   - 2 no answer
3. Closed panel Health Maintenance Organizations
   Example: Anchor Organization for Health Maintenance (N=7)
   - 0 classified correctly (physician's office)
   - 6 classified as a hospital outpatient clinic
   - 1 classified as other clinic

Discussion

A challenging methodological problem is suggested by the observation that 39 percent of these respondents incorrectly classified the type of place they usually go to for health care. Moreover, differences between the frequency distributions for the unvalidated and the validated responses suggest the possibility that unvalidated data may lead to erroneous conclusions about the types of places respondents rely on for health care. Since items describing the usual source of care are widely used in health care surveys, and since such data are widely used to analyze by setting issues of access, use, insurance

coverage, and costs, several strategies should be considered for improving the accuracy of survey data describing the usual source of care.

One strategy would be to abandon the practice of asking respondents to characterize the type of their usual source of care. Respondents might be asked only to provide names, addresses, and phone numbers for the places they usually visit and leave subsequent classifying of the types of these places to researchers. This approach to measurement might circumvent apparent validity problems, but it would be labor intensive. Even in this small, geographically finite sample, 281 respondents named 86 unique health care places. Such a strategy probably would not be feasible in a larger scale, more geographically dispersed, and socioeconomically diverse sample.

Alternatively, more widespread and systematic attention should be given to validating different survey items purporting to describe the usual source of care. It is surprising to find that so widely used a construct as the usual source of care—and one for which there might by now be a so-called standard measure—had, in fact, been measured and reported in the literature in many ways with little apparent attention given to the validity of different approaches. The result is that although many approaches to measurement have been taken in the past, it is not possible to say that data from these surveys were any more valid than those observed here. It seems entirely possible that various survey-derived characterizations of the U. S. population's reliance on usual sources of care (and associated patterns of use and costs) may be riddled with inaccuracy.

The 1987 National Medical Expenditure Survey seems to have within it capability for more fully investigating respondent abilities to characterize the types of health care places. The survey asks about visits to particular types of places and names, addresses, and phone numbers of places are collected for follow-up purposes. The identifying information supplied by respondents might be used to independently determine the actual type of the place, much as was done in the work reported here. This information could then be compared with the type the respondent had considered the place to be. Validation of a limited subsample of these national data might enable us to go well beyond the work done here. For example, it would be possible to examine more systematically how accurately respondents characterize the types of places they visit and whether particular subsets of respondents have trouble identifying correctly the places they visit. Such work also would enable us to develop a better understanding of possible sources of confusion about the types of places respondents go for care. This insight could contribute greatly to improving future attempts to measure the type of the usual source of care.

Third, it is entirely possible that more valid results could be achieved by improving the way items about the usual source of care are asked. For example, some confusion might be eliminated by rewording categories such as "private doctor's office" to "private doctor's office or clinic" thereby facilitating a correct choice by those visiting the large, clinic-like private practices discussed earlier. Such a strategy seems particularly important when surveying a low-income, inner-city population which may depend heavily on the so-called storefront clinics and, increasingly, on Health Maintenance Organizations—both of which are private practices, although they may not readily be identifiable as such by respondents. This measurement strategy was used in the 1975-1976 CHAS-NORC Survey of Access to Care although, as noted earlier, evidence is lacking that the results obtained were any more valid than those observed here (Aday & associates, 1980).

More detailed response categories also may be needed so that respondents are maximally aided in selecting the category that best describes their usual source of health care. This would be particularly useful in the case of clinics because the term clinic is used in the names of a variety of types of health care providers. For example, the 1982 National Survey of Family Growth includes eight clinic categories in an item measuring settings used for prenatal care (National Center for Health Statistics, 1988). The strength of this approach lies in the fact that decisions about grouping like kinds of clinics (such as those under public or private auspices) can be left to the researcher rather than expecting the respondent to be able to accurately formulate and portray such distinctions.

When more detailed categories are used, however, respondents may become overwhelmed by the dizzying number of choices. A multilevel item such as that used in the Robert Wood Johnson Foundation (1986) Access Study might therefore prove most useful. In such an item the respondent first chooses a broad category such as clinic and then subsequently selects the category best describing the particular kind of clinic. However, the validity of such a multilevel item may depend on the respondent's ability to make a correct choice at the first level.

The source of confusion in items measuring the usual source of health care may not lie in the structure of the response categories but rather in the differing definitions of settings being used by the researcher and the respondent. Furnishing detailed definitions of settings may help improve the accuracy of information gleaned from such items. However, it has been pointed out that such definitions may discourage respondents from reading or listening to instructions (Sudman, 1981).

Alternatively, we may need to confront the fact that the health care sector is growing increasingly complex, and distinctions between settings are becoming more blurred. There are fewer solo, office-based practices and there are more large, multispecialty group practices (and Health Maintenance Organizations) many of which are so closely affiliated with hospitals as to perhaps be—in the minds of respondents—indistinguishable from these hospitals. Similarly, many large private practices are now very clinic-like causing confusion between private physicians offices and all kinds of clinics. Finally, dimensions such as "auspices" or "affiliation with hospital" often can be very subtle and seem, from these results, to be a source of genuine confusion. Perhaps respondents simply are not knowledgeable about such matters and cannot be relied upon to report this infor-

mation accurately—no matter how researchers may tinker with the way the measure is constructed. Overall, then, although we may undertake to improve items characterizing the usual source of care, the possibility that respondents do not have sufficient knowledge of an increasingly complex health care system cannot be discarded.

In conclusion, Sudman and Bradburn (1982) caution that survey researchers should make every effort not to ask questions for which respondents do not know the answers. The challenge we face with regard to survey items depicting the usual source of care is to determine when invalid responses are a function of the way we asked the question and when they are, in fact, a function of respondents lacking the knowledge with which to answer our questions correctly.

References

Aday, L., Andersen, R., & Fleming, G. V. (1980). *Health care in the U. S.* Beverly Hills, CA: Sage Publications.

American Hospital Association Guide. (1988). Chicago, IL: American Hospital Association.

Dutton, D.B. (1978). Explaining the low use of health services by the poor: Costs, attitudes, or delivery systems? *American Sociological Review,* 43(June), 348-368.

Kasper, J.A. & Berk, M. L. (1980). Comparisons by age, sex, race, and insurance coverage on some indicators of access to care. Paper presented at the meeting of the American Sociological Association, New York, NY.

Kasper, J. A. (1987). The importance of type of usual source of care for children's physician access and expenditures. *Medical Care,* 25(5), 386-398.

National Center for Health Statistics. (1988). *Health aspects of pregnancy and childbirth, United States, 1982* (DHHS Publication No. (PHS) 89-1992). Data from the National Survey of Family Growth Series 23, No. 16. Washington, DC: Public Health Service.

Perloff, J., & Morris, N. M. (1989). Maternal and child health in Chicago: Overview of research methods. (Working Paper #1) University of Illinois at Chicago, Chicago, IL.

Robert Wood Johnson Foundation. (1988). Study of access to health care, 1986. Princeton, NJ: Robert Wood Johnson Foundation.

Short, P. F., Cafferta, G. L. & Berk, M. L. (1985, November). Outpatient use of hospitals by the poor and uninsured. Paper presented at the meeting of the American Public Health Association, Washington, DC.

Sudman, S. (Ed.). (1981). *Health survey research methods third biennial conference* (DHHS Publication No. (PHS) 81-3268). Research Proceedings Series, National Center for Health Services Research. Washington, DC: Public Health Service.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions.* San Francisco, CA: Jossey-Bass.

Walden, D. C. & Rossiter, L. F. (1979, November). Pediatric care: Charges, payments and the medical setting. Paper presented at the meeting of the American Public Health Association, New York, NY.

FEATURE PAPER

Recalling Pediatric Poison Events: Situational and Temporal Determinants of Accuracy

Ken R. Smith and Newell E. McElwee

Introduction

Over one million reported poisoning episodes occur each year in the United States, with an estimated 10 to 20 million poisonings going unreported (Temple, 1984). Over 750,000 reported exposures occurred to children under the age of 6 in 1987 in those states reporting such exposures, which cover roughly half of the U.S. population (Litovitz & associates, 1988). The medical and economic consequences associated with these poisonings are considerable. However, remarkably little work has been done to evaluate how best to collect information on pediatric poisonings or any other type of unintentional injury or death.

Similar to many epidemiologic studies where target individuals are unable to respond for themselves (Andersen & associates, 1979; Mosely & Wolinsky, 1986; Means & associates, forthcoming), studies of acute health conditions such as poisonings often rely on proxy respondents who must recall details of an event that occurred under potentially trying conditions. It is conceivable that stressful events such as accidental poisonings occurring to children may serve to maximize a parent's ability to recall the circumstances. On the other hand, factors may exist that act to adversely affect a parent's memory, including the passage of time (that is, memory decay) (Sudman & Bradburn, 1973; Bradburn & associates, 1987) or a sense of guilt or responsibility for the accident (that is, conscious suppression of information for reasons of social desirability).

Research Questions

The purpose of this study is to investigate factors that affect recall and knowledge of acute poisonings by proxy respondents. Attention is focused exclusively on poisonings occurring to young children. To establish whether a poisoning event is recalled accurately, official health records on the poisoning event gathered at the time of the poisoning are compared with survey data collected at various intervals after the poisoning event.

This study addresses the following questions:

1. How does the accurate recall of pediatric poisoning events change with the passage of time for proxy respondents? Are some types of pediatric poisonings recalled with greater accuracy with the passage of time than other types of poisonings?
2. How is recall accuracy affected by characteristics of the poisoned child, the proxy respondent, the poisoning episode, the respondents household, and the recall interview?

This paper examines such factors as the relationship of the proxy to the child, the proxy respondent's awareness of the poisoning event, the type of product ingested, treatment rendered, time of the recall interview, number of children living in the household, the accident proneness of the poisoned child, and whether or not the respondent tends to be at home a great deal.

Sample Selection

The sample was generated by selecting individuals, usually the parent, who telephoned the Intermountain Regional Poison Control Center (IRPCC) at the University of Utah Medical Center between September 1987 and February 1988. For this investigation, the focus is on episodes involving an acute poisoning by ingestion among children under the age of 6. These restrictions were imposed because such episodes comprise a significant portion of all calls received by the Intermountain Regional Poison Control Center and other poison control centers.

Cases were sampled from the IRPCC files. Each call received by the Intermountain Regional Poison Control Center generates a record about the poisoning at the

Ken R. Smith is with the Department of Family and Consumer Studies and Survey Research Center, University of Utah. Newell E. McElwee is with the Intermountain Regional Poison Control Center, University of Utah Medical Center.

This research was supported by a grant from the Intermountain Regional Poison Control Center.

time of its occurrence. Consequently, details about the event as reported by the caller are quite trustworthy given that most calls are made within 5 minutes of the accident. Moreover, trained individuals working for the Intermountain Regional Poison Control Center are in a position to question callers about the details of the event so that data collected about the episode are structured by the Intermountain Regional Poison Control Center. The IRPCC employees are pharmacists who have completed training in toxicology with additional training on how to take a history of a poisoning event. Information collected during these calls (hereafter called the PCC call) is then compared to responses provided during a later interview (hereafter called the recall interview) which was conducted under more typical survey conditions.

All recall interviews were conducted with a proxy respondent, but some interviews were completed by the person who made the PCC call whereas others were not. It is possible, therefore, to test for differential reporting biases between those individuals making the PCC call (called a witness respondent) and those who did not (non-witness respondent). This was possible because the recall interview was designed to imitate, as much as possible, a general telephone survey on childhood poisonings where the selected respondent could be any eligible adult in the household. For this study, the "most knowledgeable adult" respondent selection rule was used (Mathiowetz & Groves, 1985). Rather than selecting a random adult based on a household census, the adult who was most knowledgeable about the health of all household members was questioned. This was considered to be the optimal design for this investigation because all the health questions pertained only to the children under age 6 living in the household.

To assess the role that the passage of time has on the quality of recall of poisoning episodes, randomly selected IRPCC cases which had occurred 2, 4, and 20 weeks before the recall interview were used. These intervals were chosen because they represent what was considered to be the limits of reasonable recall for pediatric poisonings, particularly mild cases. Thirty cases were selected for each of the three time periods for a total of 90 cases.

Within each of the three time intervals, we divided each subsample further into three groups (ten in each group) based on the nature of the severity and treatment rendered to the poisoned child. The first group consists of children who were treated at home and where the caller, usually the mother, father, or babysitter, was asked by the Intermountain Regional Poison Control Center to observe the child for important symptoms. Children in this category suffered no effects resulting from the exposure. These cases are called MILD. Common exposures among children in this group include ingestion of cough syrup, vitamin pills, fluoride tablets, houseplants, hand soap, and cologne.

The second group consists of children who were also managed at home but who were given syrup of ipecac to induce vomiting. These children suffered no or only minor effects from the exposure. This group is referred to as the IPECAC group. Typical exposures for this group include ingestion of wild mushrooms, larger quantities of cough syrup (for example, Triaminic®), and adult aspirin. Parents of children in this group might have been asked to induce vomiting for the same product or food as in the MILD group. The difference in treatment was often dictated by the quantity ingested, symptoms of the child, and the child's age.

The last group consists of children who were managed in a health care facility (that is, hospital, clinic, or physician's office). These are clearly the most serious cases and consisted of children who ingested products such as concrete mix, bullets, rat poison, paint thinner, and prescription medication. This group is called the SEVERE group.

The Recall Interview

During the recall interview, the most knowledgeable adult (MKA) in the household was asked to provide the names and ages of all children living in the household. In three separate questions, the Most Knowledgeable Adult was then asked whether each child in their home had become sick during the past 6 months because they had eaten any (1) plants or wild mushrooms; (2) household products such as cleaning substances, cosmetics, garden supplies, or coins; or (3) pills or medicines. If the respondent said yes for a given child, he or she was then asked to provide particular information about the episode. This procedure was repeated for all children under age 6.

Questions about the episode were based primarily on the variables which were available from the IRPCC records, thereby allowing a comparison between the true and recalled circumstances of the poisoning event. Naturally, these variables are also those that are most important from a public health standpoint.

Measuring Recall Accuracy

Data collected during the recall interview was compared to the comparable variables contained in the IRPCC records. Recall accuracy was based on five criteria (Table 1 shows the coding scheme):

1. Whether the respondent correctly recalled the event to any degree.
2. The degree to which the respondent accurately recalled the substance ingested by the child.
3. The degree to which the respondent accurately recalled the symptoms experienced by the poisoned child.
4. The degree to which the respondent accurately recalled the amount ingested by the poisoned child.
5. The degree to which the respondent accurately recalled the time of the poisoning.

The basic indicator of recall accuracy is the first measure: was the respondent able to remember the event? In 30 percent of the cases, the respondent was unable to remember the event during the recall interview. The other four measures examine different aspects of recall. Each of these four outcomes are measured as ordinal

Table 1. Coding scheme for five measures which describe recall accuracy

Variable	Description	Coding structure
REMEMBER	Whether the event is recalled	1 = Recalled the event 0 = Did not recall the event
SUBMTCH	How well the substance is recalled	0 = Complete omission of the event 1 = Remembered other aspects of the poisoning but not the substance 2 = Substance was recalled by category; not exact 3 = Substance was recalled exactly
SYMMTCH	How well the respondent could remember the child's symptoms	0 = Complete omission of the event 1 = Incorrectly reported 2 = Correctly reported
AMTMTCH	How well the respondent remembers the amount of the ingested substance	0 = Complete omission of the event 1 = Off by more than 50% 2 = Within 50% 3 = Exact match
DATEMTCH	How well the date of the poisoning is recalled	0 = Complete omission of the event 1 = Recalled date is off by ±22 days or more 2 = Recalled date is off by ±8 to 21 days 3 = Recalled date is off by less than ±8 days

variables taking on either three or four levels. As shown in Table 1, the lowest level of recall for the indicators occurs when the episode is not recalled at all. As a consequence, measures 2 through 5 provide additional information about the 70 percent of the cases who recalled varying levels of detail about the event.

Method of Analysis

Each recall measure was regressed on a set of six independent variables using either probit (for REMEMBER) or ordered probit equations (for measures two through five). This baseline equation takes on the following form:

$$\Pr(\text{Outcome}_{i,j,k} = 1)\ X, B_m) = A + B1^*(\text{MILD}=1) + B2^*(\text{IPECAC}=1) + B3^*(\text{DAYSBET}) + B4^*(\text{MILD}^*\text{DAYS}) + B5^*(\text{IPECAC}^*\text{DAYS}) + B6^*(\text{DIFFRESP})$$

where X = the matrix of independent variables, A = intercept,
B_m = m probit regression coefficients, m = 6,
i = individual respondent i, $i = 1, \ldots, 90$
j = recall measure j, $j = 1, \ldots, 5$ (Table 1)
k = level of recall measure j, $k = 1,0$ or $2,1,0$ or $3,2,1,0$.

MILD and IPECAC are dummy variables; SEVERE is the omitted category. DAYSBET is treated as a continuous variable and it measures time between the PCC call and the recall interview. Coding DAYSBET, a set of dummy variables did not alter the substantive findings. The inclusion of the terms MILD*DAYS and IPECAC*DAYS allows for the effects of MILD and IPECAC (induced vomiting) on recall accuracy to co-vary with the time between the PCC call and the recall interview.

The respondent selection rules allow for the possibility that the respondent during the recall interview may be different from the person who made the PCC call. By including the term DIFFRESP in the equations the effect this design feature may have on recall accuracy can be estimated. It is possible that the respondents' recollections of the events are poor because they were not directly involved in the event or they were simply unaware that the event occurred at all. Selected descriptive statistics about the sample are provided in Table 2. Means of the independent and dependent variables used in the equations are shown in Table 3.

Results from the Baseline Model

Table 3 reports estimates for five probit equations, and begins by focusing on the variable REMEMBER. This takes on a value of one if the respondent remembers the event, and zero otherwise. Therefore, a positive regression coefficient suggests that the corresponding independent variable increases the likelihood that the episode is reported during the recall interview.

Table 2. Selected statistics on the characteristics of the sample

	Mean	SD
1. Age of poisoned child in months	27.0	12.7
2. Proportion male	0.651	0.48
3. Number of children in the household	3.02	1.49
4. Number of children in household < 6 years	1.83	0.85
4. Age of the respondent in years	30.02	5.85
5. Proportion of respondents married	.94	.025

Table 3. Effects of the passage of time and pediatric poisoning severity on recall accuracy (probit regression coefficients; t-statistics are in parentheses)

Independent variable name	Dependent Variables: Measures of recall accuracy					Mean of independent variable
	REMEMBR*	SUBMTCH	SYMMTCH	AMTMTCH	DATEMTCH	
MILD	−2.06 (3.42)	−1.43 (2.86)	−1.11 (2.27)	−1.44 (3.06)	−2.05 (4.46)	0.33
IPECAC	1.15 (1.39)	0.410 (0.72)	0.493 (0.99)	1.15 (2.18)	−0.252 (0.533)	.33
DAYSBET	−0.005 (1.26)	−.002 (0.67)	−.002 (0.54)	−0.002 (0.65)	−.010 (2.84)	62.0
DIFFRESP	−1.26 (2.70)	−.928 (2.47)	−.707 (1.89)	−.612 (1.68)	−.703 (2.01)	.19
MILD*DAYS	.013 (1.82)	.008 (1.26)	.005 (0.79)	.010 (1.60)	.011 (1.37)	13.8
IPECAC*DAYS	−.007 (1.01)	−.003 (0.59)	−.007 (0.14)	−.006 (1.21)	.004 (.67)	23.0
CONSTANT	1.53 (3.21)	1.03 (2.71)	.827 (2.14)	.721 (2.11)	1.91 (5.23)	1.0
-2*LOG-l	28.84	18.62	15.69	27.77	30.30	
df	6	6	6	6	6	
p	.0001	.0001	.016	.0001	<.0001	
Mean of dependent variable	.301	.663	.671	.653	.704	

[a] REMEMBER is a dichotomy. Regression parameters were estimated using binomial probit equations. The remaining dependent variables are ordinal variables. They were estimated using ordered probit.

For this equation, respondents in the MILD group have the greatest likelihood of completely forgetting the episode relative to respondents in the omitted SEVERE category. This indicates that milder forms of pediatric poisonings have a good chance of going undetected in a telephone survey relative to more intense episodes.

The interaction terms between time (between the PCC call and the recall interview) and the MILD and IPECAC groups show (1) that respondents from the MILD group have poorer recall when the recall period is short compared to respondents from the IPECAC and SEVERE group, but (2) this differential dissipates with lengthening recall periods. In other words, the passage of time (at least up to 5 months) seems to make it more difficult to recall pediatric poisoning episodes regardless of the severity of the poisoning.[1]

The use of the MKA rule in this study allows us to examine an interesting aspect of conducting surveys on pediatric poisonings and any other survey that seeks to assess the health status of children or other individuals who are unable to respond for themselves. When the respondent for the recall interview is not the same person who made the PCC call (DIFFRESP = 1), the chances that the poisoning event goes unreported increase significantly. These results are shown graphically in Figure 1, which shows predicted probabilities of recalling the poisoning event (based on the REMEMBER equation) for the SEVERE and the IPECAC groups.

There are at least two possible interpretations for this result. First, the self-appointed health spokespersons (via the MKA rule) may not be the optimal respondents. They may have rejected that role if they were informed about the exact nature of the questions ahead of time. The MKA technique may need some refinement so that specific knowledgeability criteria are met by the respondent before the interview is conducted. Second, there will always be the chance that the witness to a pediatric poisoning will not be the person selected for an interview, independent of any respondent selection rule. This suggests that the reporting behavior of the nonwitness respondent is probably a function of how much the witness shares information about the event with others and how well the nonwitness respondent remembers he details about the event after having been told. It will be important for future studies to determine the relative effects of these two components to increase recall accuracy in proxy surveys.

Because of this finding, those instances where this inconsistency occurs (that is, no recall of the event) was reexamined, particularly among the SEVERE group. The most interesting cases involved three (SEVERE) poisoning episodes where the father was the witness, and probably the only adult supervising the child. In all three cases, the father was instructed by the Intermoun-

[1]The sample comprising the MILD group is somewhat unusual because recall was poorer for those incidents which occurred 2-3 weeks prior to the recall interview compared to those which occurred 5 months earlier. Subsequent studies with larger samples will reveal whether or not this finding is atypical.

tain Regional Poison Control Center to take the child immediately to the hospital. These three cases involved the (possible) ingestion of .22 caliber bullets, concrete mix, and prescription drugs. During the recall interview, the mother was interviewed in each instance. In all three cases, she did not report any such poisonings. It is conceivable that the fathers never discussed the events with the mothers. Alternatively, she may have elected to suppress the information for reasons of social desirability, although this is doubtful because if respondents are motivated to provide socially desirable answers, then one would expect witness respondents to suppress information at least as often as nonwitness respondents. Since this is not the case in these data, the communication problems between the child's parents or guardians may be an important mechanism which reduces recall accuracy when a nonwitness is the respondent.

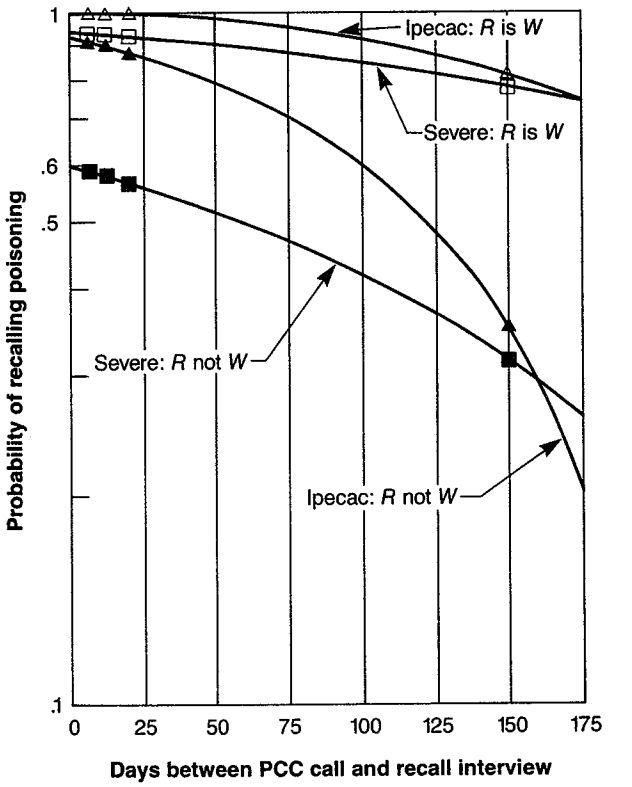
The regression results for the other five indicators of recall accuracy are also presented in Table 3. The results among these measures are similar to those found with the REMEMBER equation. Note that these regressions were estimated with an ordinal dependent variable where low values represent poor recall and high values represent higher quality recall. Again, positive regression coefficients indicate that the corresponding variable serves to enhance accurate recall.

Two relationships were detected for these latter measures which were not found in the REMEMBER regression, both of which involved the accuracy of quantitative aspects of poisonings: the amount of the ingested substance and dates of poisonings. First, respondents from the IPECAC group have more accurate recall than the omitted SEVERE group when the task calls for remembering the amount of the substance ingested by the child. This is somewhat intuitive given that a parent who must induce vomiting is more sensitized to the quantity ingested by the child. In fact, a parent from the IPECAC group may have had more interaction with IRPCC staff than the other two groups because their in-home treatment was more involved than that for the MILD group and may have required more direct treatment by the parent than the SEVERE group. In other words, members of the IPECAC group may be more directly involved in the episode in such a way that quantities of ingested poisons are more accurately recalled.

The second relationship occurs in the DATEMTCH equation. Not surprisingly, we find that shorter intervals between the PCC call and the recall interview increase the accurate reporting of the date of the poisoning. This confirms the practice of limiting the recall period for acute repeatable conditions to short periods of time.

Because of sample size considerations, more complex relationships than those reported in Table 3 were not estimated. However, other suspected factors which might affect recall accuracy were examined by adding them one at a time to the baseline model (that is, there were never more than seven independent variables in any equation). Variables examined included the child's sex and age, the parent's assessment of the child's accident proneness, household size, number of calls needed to obtain a completed interview, whether the interview occurred around dinner time, whether the respondent

Figure 1. Recall of pediatric poisonings by time and respondent status



R is W = Respondent made the PCC call
R not W = Respondent did not make the PCC call

lives alone, and whether the respondent is predominantly in the home or outside the home. Contrary to other studies on the recall of injuries in general (Larson & Pless, 1988), these characteristics were found to have no influence on any measure of recall accuracy.

However, respondents who reported other poisonings for either the sampled child (that is, they reported an event for the child but it did not happen to be the one sampled) or for other children in the household had less accurate reports of the amount ingested by the poisoned child and the nature of the substance ingested. Similar but not significant relationships were found in the other equations. This finding suggests that as the number of similar poisoning events occurring in the household increases, the less the respondent is able to accurately remember any one event for a given child. Neisser (1986) addresses this issue in his discussion of autobiographical memory. He suggests that individuals have difficulty retrieving details of specific events that happen with recurring frequency. This model of autobiographical memory contends that recall is adversely affected because similar events begin to blend into a generic pattern (or generic memory) making it difficult to recall any one event.

Conclusions

A study of factors affecting differential recall of poisoning events bears directly on studies of acute health conditions, but it also serves to identify new methods for designing and evaluating the accuracy and quality of survey questions. This study represents an initial attempt at identifying salient characteristics of the poisoning events, the respondents, and the interview which alter an individual's ability to accurately recall poisoning events.

The most interesting findings reported in this study are the relatively weak effects that the passage of time has on recall accuracy (except for recalling poison event dates) controlling for poisoning severity and the consistently strong differentials detected among the three poisoning groups. From this experimental study, it would appear that parents or guardians reporting on more severe poisonings (IPECAC and SEVERE groups) provide better reporting quality than those from the MILD group and that, overall, these differentials persist across time, at least as far back as 20 weeks. Large-scale surveillance surveys on poisonings might therefore focus in on methods that would enhance recall for the more common poisonings. On the other hand, these less severe poisonings may also represent the least worrisome episodes to survey because they may have the smallest effect on public health. The effort expended in measuring these milder episodes may depend on the objective of the survey. From a health surveillance standpoint, these milder exposures may be too difficult or too costly to measure in a survey. However, measuring the occurrence of these poisonings may be important because they alter health care utilization patterns (for example, emergency room visits) which may needlessly affect health care costs. Nonetheless, poisoning surveillance surveys need to be concerned with the omission of these types of episodes.

Perhaps the single most influential factor that will require further study is the witness-respondent relationship and the impact it has on accurate reports of poisonings. The findings from this study indicate that patterns of family communication might well serve as a focal area for survey research, particularly where respondents are reporting for those who are unable to report for themselves. Such an approach might involve doing multiple interviews within households or perhaps gauging the knowledgeability of the respondent to know how best to weight his or her responses.

References

Andersen, R., Kooper, J., Frankel, M. R. & associates (1979). *Total survey error: Applications to improve health surveys.* San Francisco, CA.: Jossey-Bass.

Bradburn, N. M., Rips, L. J., & Shevell, S.K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science,* 236, 157-161.

Larson, C. P., & Pless, B. (1988). Risk factors for injury in a 3-year-old birth cohort. *American Journal of Diseases of Children,* 142, 1052-1057.

Litovitz, T. L., Schmitz, B. F., Matyunas, N., & associate (1988). Annual report of the American Association of Poison Control Centers National Data Collection System. *American Journal of Emergency Medicine,* 6, 479-515.

Mathiowetz, N. A., & Groves, R. M. (1985). The effects of respondent rules on health survey reports. *American Journal of Public Health,* 75, 639-644.

Means, B., Mingay, D. J., Nigam, A. & associate (in press). A cognitive approach to enhancing health survey reports of medical visits. In Gruneberg, M. M., Morris, P. E. & Sykes, R. N. (Eds.). *Practical aspects of memory.* Chicester: Wiley & Sons.

Mosely, R. R. II, & Wolinsky, F. D. (1986). The use of proxies in health surveys: Substantive and policy implications. *Medical Care,* 24, 496-510.

Neisser, U. (1986). Nested structure in autobiographical memory. In Rubin, D. C. (Ed.). *Autobiographical Memory* (pp. 71-81). Cambridge: Cambridge University Press.

Sudman, S. & Bradburn, N. M. (1973). Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association,* 68, 805-815.

Temple, A. R. (1984). Foreword to symposium on medical toxicology. *Emergency Medicine Clinics of North America* 2, 1.

FEATURE PAPER

An Analysis of the Structure of the Diagnostic Interview Schedule

Mark Reiser and William W. Eaton

Introduction

This paper discusses the validity of some items from the National Institute of Mental Health Diagnostic Interview Schedule (DIS), which is an instrument used by the National Institute of Mental Health in a series of epidemiologic surveys known as the Epidemiologic Catchment Area (ECA) Program. The data to be discussed here was collected on a multistage probability sample of 3,481 respondents from the Baltimore, Maryland area. More details are given on the research design by Eaton and associates (1984), Regier and associates (1984), and Eaton and Kessler (1985).

Eaton and Bohrnstedt (1989) describe the Diagnostic Interview Schedule as a highly structured interview designed to resemble a typical psychiatric interview and to yield similar results in terms of specific mental disorders. The Diagnostic Interview Schedule represents a major advancement in epidemiologic research on mental health. In the opinion of Regier and associates (1984), the absence of such a highly structured case identification instrument had left descriptive psychiatric epidemiology in a state of relative quiescence for the 30 years before the development of the Diagnostic Interview Schedule.

The items to be discussed pertain to the areas of anxiety and depression, areas to which the Diagnostic Interview Schedule devotes 41 items. The items that occur in the results that are presented below are shown in Table 1. With anxiety and depression, an important topic is the overlap between the two syndromes. That topic has received a great deal of attention in the past in both pharmacologic and nosologic studies, and is the topic of recent studies based on the ECA data (Eaton & Bohrnstedt, 1989).

The topic of this paper is not the overlap of the two syndromes however, but rather the validity of two items in particular, relative to the responses on other items. The first item of interest is an assessment of general dysphoria and anhedonia: "Have you ever had two weeks or more during which you felt sad, blue, depressed, or when you lost all interest and pleasure in things that you usually cared about or enjoyed?" The second item asks about common anxiety: "Have you ever considered yourself a nervous person?"

These two items are particularly important in the Diagnostic Interview Survey. Nervous person represents chronic common anxiety, and dysphoria or anhedonia represents criterion A of the definition of major depressive disorder as given by the American Psychiatric Association's Diagnostic and Statistical Manual, Third Edition (DSM-III). The DSM-III definition of major depressive disorder is shown in Table 2. According to this definition, a diagnosis of major depressive disorder is made if in addition to dysphoria or anhedonia, symptoms from four of the groups in criterion B are also present.

Dysphoria or Anhedonia

The results from a factor analysis of the 41 anxiety and depression items by Muthen (1989) show that the dysphoria or anhedonia item is strongly related to other items that represent symptoms of depression. Dysphoria, "sleeping too much," "thoughts slower," "tired," and "worthless," as well as several other variables load highly a depression factor. Muthen did not perform an explicit test of fit for his results, and so such a test is applied here, using the unidimensional latent trait or "item response" model. The item response model specifies that the covariation among dichotomous manifest variables can be accounted for by a single continuous

Mark Reiser is with the College of Business, Arizona State University, Tempe, Arizona. William Eaton is with the Department of Mental Hygiene and Public Health, Johns Hopkins University, Baltimore, Maryland.

This research was supported in part by Grant No. MH41908 and in part by a grant from Arizona State University.

Table 1. Some items from the Diagnostic Interview Schedule

Nervous person
61. Have you ever considered yourself a nervous person?

Panic attack
62. Have you ever had a spell or attack when all of a sudden you felt frightened, anxious or very uneasy in situations when most people would not be afraid?

Fear of Heights
68a. Have you ever had an unreasonable fear of heights?

Fear of a crowd
68c. Have you ever had an unreasonable fear of being in a crowd?

Fear of eating in public
68h. Have you ever had an unreasonable fear of eating in front of other people (either you know or in public)?

Fear of speaking in public
68i. Have you ever had an unreasonable fear of speaking in front of a small group of people you know?

Fear of speaking to strangers
68j. Have you ever had an unreasonable fear of speaking to strangers or meeting new people?

Fear of storms
68k. Have you ever had an unreasonable fear of storms or thunder or lightning?

Fear of water
68l. Have you ever had an unreasonable fear of being in water, for instance in a swimming pool or lake?

Fear of insects, etc.
68m. Have you ever had an unreasonable fear of spiders, bugs, mice, snakes, bats, birds, or cats?

Fear of animals
68h. Have you ever had an unreasonable fear of being near any (other) harmless animal or a dangerous animal that couldn't get to you?

Dysphoria/anhedonia
72. Have you ever had two weeks or more during which you felt sad, blue, depressed, or when you lost all interest and pleasure in things that you usually cared about or enjoyed?

DIS/DSM-III Appetite Group

Lost appetite
74. Has there ever been a period of two weeks or longer when you lost your appetite?

Lost weight
75. Have you ever lost weight without trying to—as much as two pounds a week for several weeks (or as much as 10 pounds altogether?)

Eating increased
76. Have you ever had a period when your eating increased so much (Did your eating increase so much) that you gained as much as two pounds a week for several weeks (or 10 pounds altogether?)

DIS/DSM-III Sleep Group

Trouble falling asleep
77. Have you ever had a period of two weeks or longer when you had trouble falling asleep, staying asleep, or waking up too early?

Sleeping too much
78. Have you ever had a period of two weeks or longer when you were sleeping too much?

DIS/DSM-III Tired

Tired out
79. Has there ever been a period lasting two weeks or more when you felt tired out all the time?

DIS/DSM-III Psychomotor Group

Talked more slowly
80. Has there ever been a period of two weeks or more when you talked or moved more slowly than is normal for you?

Moving all the time
81. Has there ever been a period of two weeks or more when you had to be moving all the time—that is, you couldn't sit still and paced up and down?

DIS/DSM-III Lost Interest

Interest in sex
82. Was there ever a period of several weeks when your interest in sex was a lot less than usual?

DIS/DSM-III Worthless

Worthless, sinful, guilty
83. Has there ever been a period of two weeks or more when you felt worthless, sinful, or guilty?

DIS/DSM-III Trouble Thinking

Trouble concentrating
84. Has there ever been a period of two weeks or more when you had a lot more trouble concentrating than is normal for you?

Thoughts slower
85. Have you ever had a period of two weeks or more when your thoughts came much slower than usual or seemed mixed up?

DIS/DSM-III Thoughts of Death Group

Thought about death
86. Has there ever been a period of two weeks or more when you thought a lot about death—either your own, someone else's or death in general?

Wanted to die
87. Has there ever been a period of two weeks or more when you felt like you wanted to die?

Thought of suicide
88. Have you ever felt so low you thought of committing suicide?

Attempted suicide
89. Have you ever attempted suicide?

Table 2. DSM-III: major depressive disorder

Criterion A:
Dysphoric mood or loss of interest or pleasure in all or almost all usual activities and pastimes

Criterion B:
At least four of the following symptoms have each been present nearly every day for a period of at least 2 weeks

Symptoms groups

1. Poor appetite, significant weight loss, or increased appetite, weight gain
2. Insomnia or hypersomnia
3. Psychomotor agitation or retardation
4. Loss of interest or pleasure
5. Loss of energy
6. Feelings of worthlessness
7. Trouble concentrating, slowed thinking
8. Thoughts of death, suicidal ideation

latent variable. The probability, p_i of expressing a symptom is parameterized as a logistic function of the latent variable X:

$$\log[p_i/(1-p_i)] = a_i + b_i x$$

The intercept a_i and the slope b_i are the item parameters. The fit of the model can be assessed across the multinomial vector formed by the patterns of responses to the manifest variables. Table 3A shows item parameters and goodness-of-fit statistics obtained with the variables "sleeping too much," "thoughts slower," "tired," and "worthless." The likelihood ratio chi-square of 6.35 on 7 degrees of freedom (16 response patterns—8 parameter estimates—1) indicates a very good fit for the model of one latent dimension. When the dysphoria item (SAD2WK) is added to the variables, the degrees of freedom for the model increase to 21, but the chi-square statistic increases to 35.48, as shown in Table 3B, indicating that the model of a single latent variable for the five items can be rejected at the 0.025 level.

Table 3. Item response model results for depression items

	a_i (SE)*	b_i (SE)
A. Four symptoms		
Sleep too much	−1.96 (0.24)	1.65 (0.08)
Thoughts slower	−2.34 (.27)	2.30 (.13)
Tired	−1.83 (.20)	2.32 (.13)
Worthless	−2.40 (.27)	2.41 (.14)
$G^2 = 6.35, DF = 7$		
B. Symptoms and dysphoria		
Sleep too much	−3.17 (.15)	1.56 (.14)
Thoughts slower	−5.29 (.34)	2.20 (.23)
Tired	−4.07 (.25)	2.15 (.21)
Worthless	−6.14 (.48)	2.60 (.30)
Sad 2 weeks	−5.87 (.48)	2.82 (.32)
$G^2 = 35.48, DF = 21, p < 0.025$		

* Quantities in parentheses are estimated asymptotic standard errors, calculated under the assumption of simple random sampling. Since the sampling design involved clustering at the block level, the estimated standard errors are too small by a factor that is greater than 1.0 but probably less than 2.0. With the use of Taylor series linearization, the sampling design effect has been found to be quite close to 1.0 for most variables in the DIS (Eaton & associates, 1984).

Some insight into the reasons for these results can be gained by examining the cross-tabulation of the dysphoria item with a score calculated as the number of symptom groups in which a symptom is present, where the groups are as given by the DSM-III definition of major depressive disorder. This cross-tabulation is shown in Table 4. As discussed previously, the DSM-III definition prescribes a diagnosis of major depressive disorder if dysphoria, as well as symptoms from at least four of the groups in criterion B, have been present for at least 2 weeks. In Table 4 the association between dysphoria and the symptom groups is somewhat ambiguous. Less than one half of those with dysphoria also have symptoms from four or more groups, and less than one half of those with symptoms from four or more groups also indicated that dysphoria was present. Even for high scores on the group variables, that is, symptoms present from six or more groups, only 64 percent of those respondents also indicated dysphoria present. The equivocal relationship between dysphoria and the symptom groups has apparently been noticed by clinical psychiatrists, because the DSM-III definition specifically provides for it: "In a major depressive episode, dysphoria is always present to some degree, but the individual may not complain of this or even be aware of the loss . . . "

Eaton and associates (1989) give a model that captures the essence of these empirical relationships in terms of four latent classes. The first latent class, with an estimated prevalence of 81.6 percent, reflects individuals with few or no symptoms. The second class, with an estimated prevalence of 14.6 percent, is characterized by the presence of a few somatic symptoms but, at 12 percent, a relatively low prevalence of dysphoria. Eaton and associates (1989) suggest that this class may represent a "masked" depression. Class three seems very close to the syndrome of the DSM-III definition of Major Depressive Disorder, since all symptoms that appear in the definition have a relatively high prevalence for this class. Class three, itself, has an estimated prevalence of

Table 4. Frequency distribution of scores by SAD2WK

	SAD2WK		
Score	Absent	Present	Total
0	2,477	15	2,492
1	428	21	449
2	166	28	194
3	62	20	82
4	34	22	56
5	18	13	31
6	7	11	18
7	2	6	8
8	1	1	2
Total	3,195	137	3,332
	95.89%	4.11%	100.00%
	Absent	Present	Total
< 4	3,133	84	3,217
4 or more	62	53	115
Total	3,195	137	3,332

2.9 percent. Finally, class four, which is characterized primarily by suicidal symptoms, has a prevalence of 0.8 percent. Such a model goes a long way toward explaining the ambiguous relationship between dysphoria and score on the symptom groups as shown in Table 4. Individuals from class two would contribute to the count of those with a high symptom score, but absence of dysphoria; and individuals from class four would contribute to the count of those with a low symptom count and dysphoria present.

Nervous Person

As mentioned previously, "nervous person" is a DIS item that represents anxiety which is a chronic condition that might be treated pharmacologically with antianxiety agents. Phobias and panic attacks, both of which are special types of anxiety, are represented by other items, but anyone suffering from phobias or panic attacks could be expected to have a great deal of anticipatory anxiety associated with the disorders. Chronic anxiety would typically result from stress, but stress has also been implicated theoretically in the etiology of depression (Mills, 1977). A valid measure of chronic anxiety is important, therefore, in the study of the overlap between anxiety and depression.

In the results of the previously mentioned factor analysis of the 41 anxiety and depression items (Muthen, 1989), nervous person had a very low communality as shown by very low factor loadings. In other words, nervous person showed only a weak relationship to the variables in the analysis. As in the analysis of the dysphoria item, we may investigate the fit of a unidimensional latent variable model for a set of items that includes "nervous person." Table 5A presents results from Reiser (1989), showing that the items dealing with the five simple phobias constitute a very homogeneous scale, since a model with slopes constrained to equality fits the data very well. This feature implies that the items are homogeneous in their degree of association with the underlying variable of a tendency toward unreasonable fears of simple objects or places. However, as shown in Table 5B, when nervous person is included with these items the model of a single latent variable can be rejected, even if the equality constraint for the slopes is removed. The weak association with the other items is evidenced by the low slope value of 0.59 for this item.

We can also attempt to fit a one factor model for the four social phobias and nervous person. As shown in Table 6, the model fits well, with $G^2 = 23.8$ on 21 degrees of freedom, but nervous person still has a relatively weak association with the other items. The slope value of 0.76 corresponds to a factor loading on a scale of 0.0 to 1.0 of only 0.41. These results for nervous person with simple and social phobias could support an interpretation in terms of chronic anxiety. People with a simple phobia may be able to avoid the object of their unreasonable fear without much impedance to normal functioning, thus avoiding the experience that would lead them to think of themselves as nervous persons. People with social phobias, on the other hand, would have to completely disrupt their lives to avoid situations that would trigger phobia reactions, and it would be much more difficult for such people to avoid confrontations with evidence that they are nervous persons. Thus, nervous person could have a more direct relationship to social phobias than simple phobias. However, the strength of that relationship is still relatively weak.

Table 5. Results for simple phobias

	a_i (SE)	b_i (SE)
A. Phobias only		
Insects	−3.83 (0.13)	1.939 (0.11)
Heights	−4.36 (.15)	1.939 (.11)
Closed places	−5.25 (.17)	1.939 (.11)
Storms	−4.95 (.17)	1.939 (.11)
Animals	−6.40 (.22)	1.939 (.11)
	$G^2 = 28.7$, $DF = 25$	
B. Phobias and nervous person		
Insects	−3.88 (.28)	2.27 (.25)
Heights	−3.38 (.17)	1.56 (.15)
Closed places	−5.34 (.37)	2.12 (.25)
Storms	−4.69 (.30)	2.13 (.22)
Animals	−5.73 (.40)	2.09 (.26)
Nervous person	−1.27 (.48)	0.59 (.08)
	$G^2 = 76.6$, $DF = 51$, $p < .05$	

SOURCE: Adapted from Reiser, 1989

At least a partial explanation for the weak association between nervous person and the other items may be seen in the marginal proportions. Nervous person has by far the highest marginal proportion, at 23.6 percent (that is, 23.6 percent of the respondents indicated that they considered themselves nervous people). The next highest value is 9 percent, which occurs for both "fear of insects" and "trouble falling asleep." Most of the other variables have marginal proportions that are below 5 percent. In judging the high number for nervous person, we should keep in mind that most sample members are responding "no" to all or almost all of the DIS questions, and so they probably begin to feel a pressure to acquiesce—to say "yes" to something.

In results from Reiser and associates (1986), it was shown that the tendency to acquiesce can have a large effect on response rates under circumstances that are as simple as having an item worded positively instead of negatively. So, as respondents begin to feel a pressure to acquiesce, they will probably do so to questions that are the least sensitive and the least threatening to their self-images. Nervous person, in contrast to the other DIS variables, seems to be such a question. It would not be awkward for people to volunteer the statement in

Table 6. Results for social phobias

	a_i (SE)	b_i (SE)
Public speaking	−6.25 (0.49)	2.20 (0.29)
Speaking to stranger	−7.62 (.75)	3.14 (.41)
Eat in public	−9.35 (1.23)	3.55 (.57)
Crowds	−6.48 (.58)	2.81 (.35)
Nervous person	−1.32 (.58)	0.76 (.11)
	$G^2 = 23.8$, $DF = 21$	

everyday conversation that they consider themselves to be nervous people, or even that they have a "touch of claustrophobia," but it would be very unusual for people to state that they have an "unreasonable fear," or that they "think a lot about death." Although respondents are probed for severity of symptoms during the interview, the high prevalence for nervous person probably represents a tendency to acquiesce to a nonthreatening question in an environment where negative responses are being given for almost all items.

Eaton and associates (1989) give a model that succinctly portrays some of these results for nervous person with four latent classes. In their model, the first class has a probability of zero for any symptom except nervous person. This class represents normal individuals, some of whom may be responding to nervous person due to acquiescing. The second class represents phobias without panic attacks. The third class reflects individuals who suffer from phobias and also have a high probability of considering themselves to be nervous. The fourth class is also characterized by a high probability of being a nervous person, but the conditional probabilities for the other symptoms are fairly low. The prevalence rates for the classes are 72.6 percent, 21.4 percent, 3.3 percent, and 2.7 percent respectively.

Conclusions

The results reported here, as well as other reports (Eaton & Bohrnstedt, 1989; Robins & associates, 1981) show that in general the Diagnostic Interview Schedule is a good instrument for assessing psychopathology. Most of the questions on the Diagnostic Interview Schedule ask the respondent to simply report on the presence or absence of a symptom. This paper discusses at length two questions that require the respondent to do more than just report a symptom. To some extent, with both the dysphoria and nervous person items, the respondent is required to make an inference about himself or herself. In making such an inference, the evidence seems to indicate that some respondents have difficulty recognizing their own emotional status. Also, such responses seem to be influenced by extraneous factors such as acquiescing and responding desirably.

The relationship of the dysphoria item to the depression symptoms was ambiguous. It may be that respondents find it easier to admit to symptoms of depression than to the actual condition itself. Mental illness carries with it a stigma in our society, and some respondents who are depressed may not be willing to admit it to themselves. Moreover, many of the symptoms of depression can be rationalized away as due to a physical rather than psychological cause. These results imply a more complex relationship among the variables than can be captured by traditional notions of reliability and validity. Fortunately, these complex relationships can be modeled, as shown by Eaton and associates (1989).

These results also imply that users of epidemiologic data on mental health may want to modify the role of the dysphoria item in the diagnosis of major depressive disorder from that given by the DSM-III definition. As the DSM-III definition now reads, major depressive disorder is absent with probability equal to 1.0 if dysphoria is absent. In a clinical diagnosis, a psychiatrist may be able to use his or her experience and other sources of information, such as family members, when making a determination regarding criterion A of the DSM-III definition. In an epidemiologic survey, however, it may be more realistic to increase the reliability of the diagnosis by including more items in the determination of dysphoria. Eaton and others (1989) suggest that, in addition to the dysphoria item, items asking about "thoughts of death," "feeling worthless," "lost interest in sex," "sad for two weeks" and "feeling hopeless" could be included in a definition of dysphoria. With that approach, the single dysphoria item, "sad two weeks" has a status more like the other items: in the presence of major depressive disorder, it is expressed by the respondent with a probability greater than 0.0 but less than 1.0.

If the goal of the definition is to classify people as to the presence or absence of Major Depressive Disorder, then a latent class model has some especially appealing features. Since the goal is to classify, the problem may be approached with a discriminant function. If, as here, the predictor variables are discrete, then a discrete discriminant function may be represented by a contingency table (Goldstein & Dillon, 1978). If the dependent variable is unobservable, then the contingency table is called incomplete. The parameterization of that incomplete contingency table is the latent class model (Haberman, 1979). Thus, a latent class model could be expected to give the optimum classification of response patterns to diagnostic categories.

The relationship of the nervous person item to the other anxiety items was fairly weak. The marginal proportion for this item was quite high as compared to the other items, and given the colloquial wording of the question, the weak relationship to the other variables may be due to acquiescing. The Diagnostic Interview Schedule asks a lot of the respondents in the area of chronic anxiety. There are a couple of symptoms of depression that could also be indicative of anxiety, such as trouble falling asleep and appetite problems, but for the most part it is up to the respondents to tell us whether they suffer from chronic anxiety when they respond to the nervous person item. As such, the syndrome of chronic anxiety, as distinct from the syndrome of depression, phobia, and panic attacks, is not well represented in the Diagnostic Interview Schedule. In an epidemiologic study, mental health researchers might want to consider including additional items that would reflect the pathological nature of chronic anxiety, such as asking respondents if they are often so nervous that they wish they had a tranquilizer.

References

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*. (3rd ed.). Washington, DC: American Psychiatric Association.

Eaton, W. W., Holzer C. E., Von Korff, M. & associates (1984). The design of the epidemiologic catchment area surveys. *Archives of General Psychiatry*, 41, pp 942-948.

Eaton, W. W. & Kessler, L. G. (Eds.) (1985). *Epidemiologic field methods in psychiatry: The NIMH epidemiologic catchment area program*. New York: Academic Press.

Eaton, W. W. & Bohrnstedt. G. (in press) Latent variable models with dichotomous outcomes: analysis of data from the epidemiological catchment area program. *Sociological Methods and Research.*

Eaton, W. W., McCutcheon, A. & Sorenson, A. (in press). Latent class analysis of anxiety and depression. *Sociological Methods and Research.*

Goldstein, M. & Dillon, M. (1978). *Discrete discriminant analysis*. New York: Wiley & Sons.

Haberman, S. (1979). *Analysis of qualitative data (Vol. 2)*. New York: Academic Press.

Mills, I. H. (1977). Noradrenaline and the coping process in the brain. In A. Jukes (Ed.) *Depression—the biochemical and psychological role of ludiomil* (pp. 53-58). Horsham, England: Ciba Laboratories.

Muthen, B. (in press) Dichotomous factor analysis of symptom data. *Sociological Methods and Research.*

Regier, D. A., Meyers, J. K., Kramer, M. & associates (1984). The NIMH Epidemiologic Catchment Area Program. *Archives of General Psychiatry*, 41, pp. 934-941.

Reiser, M. R. (in press) An application of the item response model to psychiatric epidemiology. *Sociological Methods and Research.*

Reiser, M.R., Schuessler, K. F., & Wallace, M. (1986). Direction of wording effect in dichotomous social life feeling items. In N.B. Tuma (Ed.). *Sociological methodology 1986* (pp. 1-25). Washington DC: American Sociological Association.

Robins, L. N., Helzer, J. E., Croughan, J. & associate (1981). National Institute of Mental Health diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry*, 38, 381-389.

FEATURE PAPER

Validity of Self-Reports of Cancer Incidence in a Prospective Study

Donald J. Brambilla, Nancy L. Bifano,
Sonja M. McKinlay, and Richard W. Clapp

Introduction

In many epidemiologic investigations, especially large-scale studies, such methods of data collection as clinical examinations or abstraction of medical records are impractical owing to budgetary constraints and other limitations. Most of the investigators who face such constraints rely on self-reports of both health status and medical history. This choice of methodology has fostered interest in the quality of self-reported medical data. While the literature in this area is growing rapidly, some topics remain relatively unexplored, particularly the reliability and validity of self-reports of chronic disease status and diagnoses.

In this paper, an investigation of the quality of self-reports of cancer incidence is described. Self-reports of chronic disease status have generally been verified by comparing them to medical records or by obtaining further data in clinical examinations. These approaches are quite useful for estimating rates of false positive reporting. For example, Colditz and coworkers (1986) examined the medical records of women who reported that they had cancer, cardiovascular disease, or other conditions. These techniques, however, are much less useful for estimating rates of false negative reporting, for at least two reasons. First, incidence rates for these conditions are generally rather low. Thus, even at moderately high rates of misreporting, false negatives will comprise at most a small proportion of total negative reports and large numbers of negative reports would have to be screened to identify a few false negatives. Second, some of the false negatives may reflect deliberate misreporting caused by a desire for privacy, fear of the consequences of disclosure, or self-denial of the existence of the disease. Subjects who provide false negative reports for such reasons may also tend to refuse to provide access to medical records or to refuse clinical verification of their reports. Under these circumstances, accurate estimation of the proportion of false negatives would be very difficult.

In this report, an alternative approach is employed. Self-reports of cancer status that were obtained during a 5-year prospective study were matched against the files of a cancer registry that is maintained by the state in which the study took place. This approach avoids the difficulties just described and at the same time provides estimates of rates of both false positive and false negative reporting.

Methods

Subjects for the prospective study were recruited from the respondents to a baseline survey of mid-age Massachusetts women that took place in 1981-1982. The baseline survey employed a two-stage cluster design to obtain a random sample of women who were 45 to 55 years old (McKinlay & associates, 1985; Brambilla & McKinlay, 1987). Two mailings of a brief questionnaire, followed by telephone interviews of women who did not respond to the mailings, produced 8,050 responses (response rate: 77 percent). Respondents who had menstruated in the 3 months before the baseline survey and who had not had hysterectomies or other procedures that would permanently halt their menses were invited to participate in the prospective study. This study consisted of six telephone interviews that were conducted at 9-month intervals. In these interviews various aspects of health, medical history, and social circumstances were explored. A total of 2,569 women, or 93.7 percent of those who were eligible, were successfully recruited and 2,311 (90.0

Donald J. Brambilla, Nancy L. Bifano, and Sonja M. McKinlay are with the New England Research Institute, Watertown, Massachusetts. Richard Clapp is with the Massachusetts Cancer Registry, Boston, Massachusetts.

This research was supported by Grant No. AG03111 from the National Institute on Aging of the National Institutes of Health. We thank Diana Orenberg, Elaine Groipen, and Ralph Marple for help with the matching and Judy Pierson for her help with the preparation of this manuscript.

percent) of those who were recruited completed all six interviews. Another 215 women dropped out during the study after completing an average of three follow-up interviews. The other 43 cohort members missed one or more interviews but completed the sixth follow-up interview.

In the cross-sectional survey, each woman was asked if she had cancer and was offered a choice among three precoded possible answers: no; yes, but not receiving treatment; and yes, receiving treatment. At each follow-up interview, she was asked if, in the prior 9 months, she had been told for the first time that she had cancer. The same choice of possible answers was offered. Several women mistakenly answered affirmatively at more than one followup. Only the first affirmative answer is included in this analysis.

Responses to these questions were verified using the files of the Massachusetts Cancer Registry. The Registry was established to record incident cancers diagnosed in Massachusetts residents on or after January 1, 1982. Under State regulations, hospitals are required to report new cases within 6 months of diagnosis, although they are not required to report nonmelanotic skin cancers or in situ malignancies. Through a cooperative agreement, Massachusetts residents who have cancers diagnosed in surrounding states are also included. Listings maintained by the Registry include name, home address, date of birth, date of diagnosis, and other information that could be used to match records.

To verify the self-reports, a list of the 2,569 women in the prospective study was matched against the files maintained at the Cancer Registry. The list included multiple records for women who had moved or changed their names during the study. The entire list, consisting of over 4,000 records, was matched against the Registry's files twice, the first time using last name and year born as matching variables and the second using last name and first initial. Apparent matches were then confirmed by comparing the full names and addresses listed at the Registry with those in the cohort files.

Reports of cancer that were obtained during the prospective study but were not found on the Registry and those who were listed at the Registry but had not been reported during the prospective study were investigated further. The listing of cohort members that was sent to the Registry was reexamined to ensure that the mismatches had not been caused by misspelled names or other errors. Each of these subjects, or their proxies if necessary, were contacted for further information regarding the diagnosis and for changes of name or address that might have prevented a match. Where such changes were encountered, further runs against the Registry's files were performed using the new information. The proportion of mismatches obtained in a study such as this obviously depends on the completeness of the Registry's files. However, little is known about the completeness of these files, so the examination of mismatches was supplemented with other information. Interviewers on the prospective study had been trained to record all information volunteered by subjects during the interviews, so completed interview instruments were checked for information regarding cancer. Medical records and death certificates were available for some subjects owing to other validation studies. They were used to resolve some of the mismatches.

Some mismatches occurred because a few cancers were not diagnosed until after the final interview in the prospective study. In this paper, these cases are treated as free of cancer according to both the self-reports and the Registry.

Forty-one mismatches were excluded from this investigation primarily because the self-reports could not be verified. Twenty women reported that they had skin cancer and two others reported that they had in situ disease. These reports could not be verified because the Cancer Registry does not record cases of these forms of cancer, as noted earlier. Nineteen cancers reported at baseline or at the first followup could not be validated using the Registry's files because they were diagnosed before the Cancer Registry began operations.

Results

Excluding the cases just noted, 76 women reported diagnoses of cancer during the prospective study (Table 1). Sixty-three of these reports were confirmed using the Cancer Registry. Among the 13 self-reports of cancer that were not found on the Registry's files, one was attributed to an error by a follow-up interviewer and eight others were attributed to misreports by respondents who actually had benign tumors or other nonmalignant growths. Four (50 percent) of the women who erroneously reported nonmalignant growths as cancer indicated that they were receiving treatment, compared with 71 percent of the 63 confirmed cases of cancer. Callbacks to the respondents revealed that several of these self-reports had been obtained while the respondents were waiting for the results of biopsies that later proved to be negative. Thus, these respondents reported that they had cancer before they had obtained final diagnoses.

The remaining four affirmative self-reports that were not found on the Registry's files were confirmed using medical records or death certificates. All four were diagnosed after the Registry began operations but, in spite of an intensive search of the Registry's files, none of the four was located. One case was classified as a borderline carcinoma in a pathology report, so it may not have been reported to the Registry. No explanation could be found for the absence of the other three. These four provided some measure of the completeness of these files, a point that is considered in the discussion.

Among the 2,452 women who did not report diagnoses of cancer during the prospective study, 11 were found on the Registry's files (Table 1). One mismatch was caused by an interviewer error during the prospective study. The other 10 were women who had dropped out of the study before their cancers were diagnosed. In six of these cases, interviewers had been informed by proxies that the subjects had cancer and were too ill to be interviewed. Thus, 6 of the 10 dropped out specifically because they had cancer.

Table 1. Results of matching self-reports of cancer incidence with files of the Massachusetts Cancer Registry

	Listed on Cancer Registry?		
Prospective study	Yes	No	Total
Cancer	63	13	76
No cancer	11	2,441	2,452
Total	74	2,454	2,528

One of the major concerns in prospective epidemiologic studies is that the incidence rate of a disease of interest among subjects who drop out of the study may differ from the rate among subjects who remained in the study. In the prospective study, 4.6 percent (10 of 215) of the dropouts developed cancer after leaving the study, compared with 2.9 percent (67 of 2,338) of those who remained in the study. The difference is not statistically significant, possibly because of the sample sizes involved, but it does suggest that incidence rates for cancer may be higher among the dropouts than among those who remain.

In summary then, 67 (88 percent) of the self-reported cases of cancer were confirmed using the Cancer Registry and other means, while 11 percent were attributed to false positive reports by the respondents. Ten (13 percent) of the cancers that could have been recorded during the prospective study were missed because subjects dropped out of the study before they were diagnosed as having cancer. None of the false negative self-reports appeared to be deliberate failures to disclose information that might be caused by a desire for privacy or fear of the consequences of disclosure.

Discussion

The small proportion of false positives among the affirmative self-reports is encouraging with regard to the validity of self-reports of cancer incidence. The proportion of false positives can probably be reduced even further if information regarding the type of cancer involved, the method of diagnosis, and the methods of treatment that were offered or received is obtained from subjects who report diagnoses of cancer. Such information can be employed to eliminate probable misreports of benign tumors or other nonmalignant growths. It is probably not sufficient to ask subjects who report that they have cancer if they are receiving treatment for it: roughly half of those who erroneously reported that they have cancer also indicated that they were under treatment. Thus, more detail is needed to distinguish false positive from true positive self-reports.

It may be necessary to recontact some of the respondents to confirm their self-reports. For example, subjects should be recontacted if the additional information described above suggests that they have provided false positive reports. Even if such information is not available, a subsample of respondents should be recontacted so that the rate of false positive reporting can be estimated.

It is also important to recontact subjects who drop out of a study, or their proxies if the subjects cannot be interviewed. The results of this study indicate that the incidence rate of cancer may be higher among dropouts than among those who remain in the study. At the very least, failure to ascertain the status of dropouts would mean a loss of cases and a potentially substantial reduction in statistical power. If the probability that a subject drops out varies with risk factors for the disease of interest, then the loss of subjects could bias the results of the study. The potential magnitude of the bias will depend on the overall dropout rates. It is especially important to ascertain the status of dropouts where the dropout rate is relatively high.

These steps can substantially increase the cost of a study. Increased costs may result in smaller sample sizes, if budgets are limited, but they will also result in higher quality data if the measures described above are effective. Striking the proper balance between reduced sample size and improved data quality will require careful planning.

Finally, one of the striking features of this study is the absence of deliberately false negative reports of cancer. This does not mean that deliberately false negative self-reports do not occur. Rather, it should be interpreted to mean that they occur at a low rate. It is possible that some false negative reports went undetected, but this seems highly unlikely. Such reports would go undetected if the listing on the cohort files or the Registry's files contained errors that prevented a match. It was precisely because of potential problems with such errors that initial rounds of matching employed minimal information and that a series of attempts to match, each using somewhat different information, was employed in this investigation. These efforts should have minimized problems with errors in the listings. False negative self-reports would also go undetected if the case was missing from the Registry's files. That would imply the intersection of two events that are probably independent: failure of a respondent to report a cancer on the prospective study and failure of a hospital to report a case to the Registry. Four of 80 possible cases, or 5 percent of the total, could not be found in the Registry's files, indicating that a small proportion of incident cases was indeed missing. Thus, each of these events seems to have a low probability of occurrence and their intersection is therefore even less likely.

The absence of deliberately false negative self-reports may be as much a function of the design of the study as it is a function of the study population. All of the interviewers in this study were mid-age women. Other investigators (Hochstim & Athanasopoulos, 1970) have argued that such matching of interviewer and respondent characteristics can increase response rates and improve response quality. Usually this issue is discussed in the context of face-to-face interview surveys, but there is reason to suspect that characteristics that can be perceived over a telephone can also affect responses (Stokes & Yeh, 1988). It is also important to recognize that participation in a longitudinal study such as this implies a commitment to the study that may include a willingness to provide these data. If this is a major factor

in the absence of false negative reports, then cross-sectional investigations may produce higher rates of such misreporting. Indeed Chambers and co-workers (1976) as well as others have reported rather substantial proportions of false negatives in their cross-sectional surveys.

This is but one study of an important issue, and the extent to which the results can be generalized remains unclear. Rates of overreporting and underreporting may well vary with the respondent's characteristics, including, among others, age, education, and the respondent's level of knowledge regarding the etiology of cancer. The rate of false positive reporting will vary among studies that focus on cancers at different sites because the incidence rates of nonmalignant growths that could be misinterpreted probably varies among sites. Gender may influence rates of misreporting in studies that focus on the incidence of cancer in general. The incidence rates of benign tumors and other growths that could be misreported as malignant neoplasms are likely to vary with gender. For example, most of the false positive reports obtained in this study were provided by women who had benign tumors or cysts on their breasts, something that is much less common in men.

Nevertheless, the results indicate that valid and reliable self-reports of cancer can be obtained, especially if investigators take the time to gather from the respondent detailed information on the type of cancer, the treatment received, and methods of diagnosis.

References

Brambilla, D. J. & McKinlay, S. M. (1987). A comparison of responses to mailed questionnaires and telephone interviews in a mixed mode health survey. *American Journal of Epidemiology,* 126(5), 962-971.

Chambers, L. W., Spitzer, W. O., Hill, G. B., & associate. (1976). Underreporting of cancer in medical surveys: a source of systematic error in cancer research. *American Journal of Epidemiology,* 104(2), 141-145.

Colditz, G. A., Martin, P., Stampfer, M. J., & associates. (1986). Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *American Journal of Epidemiology,* 123(5), 894-900.

Hochstim, J. R. & Athanasopoulos, D. A. (1970). Personal follow-up in a mail survey: its contribution and its cost. *Public Opinion Quarterly,* 43, 69-81.

McKinlay, S. M., Bifano, N. L., & McKinlay, J. B. (1985). Smoking and age at menopause in women. *Annals of Internal Medicine,* 103(3), 350-356.

Stokes, L. & Yeh, M.-Y. (1988). Searching for causes of interviewer effects in telephone surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, & associates (Eds.). *Telephone survey methodology* (pp. 357-373). New York City: Wiley & Sons.

DISCUSSION PAPER

Scientific and Professional Allies in Validity Studies

Lois M. Verbrugge

Introduction

Science is a particular strategy for discovering the truth of things. Within its bounds of perspective, technique, and ethos, scientists have a persistent concern for the quality of measurements. Clinicians, too, have this concern in the context of differential diagnosis and patient evaluation.

There is a perpetual tug between the quality of scientific and clinical measurements and their price in money and effort. A key reason for validity studies is to find convenient instruments (called tests) that can substitute for established ones (called criterions). The aim is to develop inexpensive or easily administered instruments that do about as well as expensive or difficult ones. Some loss of quality typically occurs in the degree of detail or precision obtained. Whether it is tolerable or not depends on the use to which the results will be put: If scientific knowledge is the only product of a validity study, the difference between a criterion and test is measured and reported; no one suffers or is deeply bothered by a large gap. But if the test will be used for (a) decisions about diagnosis, surgery, disability program entry or exit, or other interventions, or for (b) epidemiological rates or surveillance, then just how well the test matches the criterion has immense importance.

This paper states nine issues that health survey researchers should consider when engaging in validity studies and, when relevant, uses the papers in this volume as examples. Next, the focus is on the validity of disease reports in health surveys, and how collaboration with clinical colleagues can improve the quality and acceptance of disease data based on self-reports. A current project on osteoarthritis that will determine how well respondent reports of chronic joint symptoms match clinical diagnosis and x-ray diagnosis is described.

An Epidemiological Format

Validity studies can be conveniently framed in terms of epidemiological notions of sensitivity, specificity, accuracy, and predictive value. The paper concentrates on sensitivity and specificity since they state amounts of difference between the criterion and test.[1] Figure 1 portrays the analytic framework.

Sensitivity and specificity both measure the quality of the test against the criterion. Sensitivity is the percent of persons with disease (or event) X who are positively identified by the test. Specificity is the percent of people free of that disease who score negative on the test. Thus, both refer to correct classification of persons with or without the disease (or event). High sensitivity means the test does a very good job identifying the ill, and high specificity means it does a very good job of identifying the well. The two measures are intrinsically bound together in a given data set; increasing sensitivity reduces specificity and vice versa. Misclassified cases are of two kinds: the truly ill who nevertheless score well on the test (false negative, FN), and the truly well who happen to score ill (false positive, FP) (Fletcher & associates, 1988; Hulley & Cummings, 1988; Kelsey & associates, 1986; Sackett & associates, 1985).

Lois M. Verbrugge is with the Institute of Gerontology, The University of Michigan, Ann Arbor

The author thanks rheumatology colleagues (especially David Felson, Timothy Laing, and Frederick Wolfe) and biostatistics colleagues (Morton Brown, Lincoln Moses, and Lisa Weissfeld) for their help in the design and analysis structure for the osteoarthritis project described at the end of the paper, and in general discussions surrounding the project that have helped her see the possible, beautiful bridges between clinical and social science research. Preparation of the paper was facilitated by a Special Emphasis Research Career Award (KO1 AG00394) from the National Institute on Aging.

[1]Accuracy is less informative; it states the proportion of correct (true positive plus true negative) tests. Predictive value is the probability that a positive-test person actually has the disease (or event), or a negative-test person does not have it.

Figure 1. An epidemiological format

CRITERION	TEST: YES	TEST: NO
YES	true positive (TP)	false negative (FN) (β)
NO	false positive (FP) (α)	true negative (TN)

$$\text{Sensitivity: } \frac{\text{true positives}}{\text{all who have the disease/event: Criterion = Yes}} \qquad \frac{TP}{TP + FN}$$

$$\text{Specificity: } \frac{\text{true negatives}}{\text{all who don't have disease/event: Criterion = No}} \qquad \frac{TN}{FP + TN}$$

Note: α and β refer to H_0: does not have disease.

Restating a prior point: In scientific research conducted for knowledge accumulation we learn something worthwhile even if the sensitivity and specificity values are low. By contract, when one hopes to use the test as a substitute for the criterion in estimating prevalence rates or evaluating patients, then high values are requisite. Tests yielding anything less than that are judged worthless and of no interest.

The basic data structure to be discussed here is the comparison of reports elicited in health surveys with official records (medical, administrative) of the same phenomena. The health survey reports are the test, and the official records the criterion. Four of the papers in this volume have this format: Calore, Perloff and Morris, Smith and McElwee, and Brambilla and associates. Another paper by Reiser and Eaton is quite different. It studies the internal structure of items used in official definitions of two psychological problems. It provides a detailed exploration of the test rather than a comparison of test with criterion.

In health survey research the unit of analysis is typically a person or an event. Person is appropriate when something happens just once, such as diagnosis (the official onset) of a chronic disease. Event is appropriate when it can happen more than once, such as hospitalization episodes or ambulatory care visits. Most points made in this paper pertain to both people and events, though my examples emphasize one or the other.

Issues in Design and Evaluation of Validity Studies

The following issues should be considered, in the sequence stated, when contemplating a validity study in health survey research (Figure 2):

1. What are the questions you want to answer? What exactly are the questions your data will and will not answer?
2. Who wants to know besides you? Who will be interested in your results and consider them worthwhile?
3. How is one measure designated as criterion and the other as test? Do you staunchly trust the criterion more than the test, or are they really similar in merit—both true, but different?
4. How are the criterion and test defined? Is the criterion a simple or complex measure? If complex, is it based on uniform rules or professional judgment? Is the test derived from a single question, an index, or a multivariate rendering of respondent answers?
5. Who are the reporters for the criterion and test? Are errors the result of conscious effort to misreport, or entirely benign—because people and administrative systems may not care about what you care about or may not organize the world the way you do?
6. (a) Is there a good conceptual match between the criterion and test? Do they really aim to measure the same thing? (b) Can subjects be easily matched in the two data streams containing the criterion and test scores?
7. Ultimately, you want to see a cross-tabulation of yes and no scores for the criterion and test. Will you explore various rules for the empirical match, or is your mind fixed from the start about what constitutes a match? If you opt for exploration, what are the lowest thresholds you will accept for a score of Yes on the criterion and test?
8. How much does validity matter? Are the errors consequential for decision making about social or individual matters? If so, is one error worse than the other: Is "letting the guilty go" (false negative) or "jailing the innocent" (false positive) more bothersome?
9. Will the strength of the match between criterion and test change much if you use another sample? If several settings interest you, do you anticipate the results to be stable across them?

Now consider the issues in detail.

The Questions

Common approaches in validity studies on health surveys are to (a) check whether people who report a disease or event (test = yes) are telling the truth, or (b) if people with an official record of the disease or event

Figure 2. Issues in design and evaluation of validity studies

1. The question(s)
2. The audiences
3. Designating the criterion
4. Defining the criterion and the test
5. Reporters
6. Concept and subject matches for the criterion and test
7. Empirical match of the criterion and test
8. Importance of errors
9. Sample bias

(criterion = yes) report it when asked. Both are incomplete designs, yielding counts for just two cells of Figure 1. One type of error is measured, either false positive or false negative, but not both. The Smith and McElwee paper is an example of the second situation, where false negatives are identified. The results of an incomplete-design study can be very interesting, especially if researchers work to uncover systematic factors that influence the one type of error measured. But the data will not reveal goodness-of-match between the criterion and test.

Only with a full design, in which criterion and test scores are available for everyone, can you evaluate the quality of the test instrument (Marquis, 1978, 1984). Typically, the barrier to this comes from the side of official records, since the clientele must be scanned to, first, identify the sample cases and, second, check for recorded events among the sample cases.[2] When the sample is small or the event rare relative to the whole administrative data set, reluctance about participating in a validity study rises steeply and so do the financial costs. Papers by Calore and by Brambilla and associates have complete designs. Although both have sample sizes of around 2,000 (small relative to the official records) and uncommon events, the incentives for officials were strong enough to overcome any reluctance. The scientists wanted to know how well the health survey performed; but the administrators wanted to know how well the record system performed!

The papers in this session pose these questions: Do survey respondents accurately report the number and dates of their inpatient and outpatient visits (Calore)? Do women report diagnosed malignancies in a health interview (Brambilla & associates)? Do people remember their children's poison events—those that were reported to a poison center (Smith & McElwee)? Do people know the kind of place they receive medical care (its organizational niche) (Perloff & Morris)? How often do survey items actually occur in the combinations required for the Diagnostic Interview Schedule diagnoses of depression and anxiety (Reiser and Eaton)? Readers are invited to place these within the framework of Figure 1.

The Audiences

It is important to consider at the beginning where results of a validity study will go. Besides one's scientific peers, there are professional audiences for the results of validity studies: Administrators of health records systems are interested in how well they capture target events; clinicians are concerned with patient diagnosis; health statisticians want accurate rates of disease prevalence and health services use; and Federal agencies wonder how to collect good information at reasonable monetary cost. If data are to speak to such audiences, scientists must listen to how these groups phrase the questions. Does the study design answer their questions as well as yours?

If one assumes that both data streams can be fallible, then a full design will show not only how well people report health events, but also reveal the record system's coverage. Brambilla and associates provide an example of mutual interest between social science and the cancer registry; both groups learned that they are doing a good job of measurement. Calore also shows a collaborative interest, but the results were less conclusive for either science or recordkeeping (due to issues of matching, discussed below). Similarly, scrutiny of clinical and epidemiological data can show not only whether people report their diagnosed health problems, but also how clinicians make decisions based on patient complaints.

Sometimes, it is not possible to interest relevant professionals in a validity study if they hold steadfastly to the criterion as infallible and reject any alternative approach to measuring it. Neither the scientific results nor the possibility of finding an inexpensive substitute for the criterion will interest them at all. In short, collaboration requires curiosity on both sides.

Designating the Criterion

In validity studies, two instruments (sometimes more) are compared. Typically, one is chosen as the "gold standard", the truest and best way to measure the disease or event of interest. When there is a choice between profession-based and survey-based instruments, the former are always chosen for the criterion. This hegemony is seldom questioned. It inserts a powerful bias in health survey research—the assumption that professional people and procedures can measure a target datum better than the people who experienced it.

Yet health as experienced, perceived, and remembered by individuals is what guides their future behavior and attitudes. (Here, health is a global term for acute and chronic conditions, health services use, and self-care.) It is veridical for individuals. From this perspective, the data produced in self-reports are truth, too.[3]

Then why are validity studies conducted? Shedding the armor of bias, it is because we wonder if certain professional information and contacts can be measured adequately by asking people about them. Are professional truths accessible through personal truths? Stated this way, the question acknowledges that both types of data are "true but different", not that the professional data alone are true.

The power of the medical mind and model in this century has fostered widespread doubt about self-report data. The perspective that the criterion is to be trusted, and respondent reports distrusted, is pervasive among professionals and present even among scientists. In fact, scientists sometimes end up more medically minded than research physicians. Thoughtful physicians know that their gold standards have no intrinsic, enduring eminence and that they will change as medical knowledge

[2]There are variations on this format, depending on whether the administrative system contains records for all eligible persons or just those with target events.

[3]To make this point, the important fact that self-reports are a public truth proffered by individuals to an interviewer or unseen scientist has been dropped. Private truth may be somewhat different. Its details can be included, excluded, or reconfigured when reported to others—spouse, friend, interviewer.

advances. Criterions' truth is a social one, buttressed by medical consensus and convention.

In my own work, I have dropped the term "validity" and use "concurrence" instead. Thus, research that compares clinical and self-report data tells us about the degree of match between decent and interesting sources of data, both collected with skill and attention to quality. Errors can be measured from either perspective—with respect to the professional data as baseline, or the self-report data as baseline. This approach and language are respectful to both types of data and do not assert dominance of one over the other. They also construct a sturdy bridge between medical and scientific colleagues; no one is considered second class.

Defining the Criterion and the Test

1. Is the criterion a plain, fancy, or slippery character? Some criteria are based on a single datum such as a lab test or clinic visit; others are multiplex, based on a multiple-item index or summary judgment. In the multiplex situation there may be a clear rule about how the relevant data items are combined to score yes or, by contrast, allowance for diverse combinations among items to score yes. An example of the latter: Clinical diagnosis of many physical and psychological conditions involves scrutiny of numerous indicators, and physicians vary in their mental algorithms for assembling information; thus, both diverse manifestations and diverse scoring rules lead to yes. This makes the criterion a slippery character, but still considered valid in professional realms. Concern about nonstandard algorithms has led some specialties to develop diagnostic criteria; that is, specified combinations of patient data for case definition (test = yes).

Ideally, the criterion should be measured independently of the test; that is, by different observers and involving different component items. This is a key problem in development of diagnostic criteria, where the criterion measure (diagnosis) is always derived from some of the test items (symptoms, signs, lab tests) themselves. Intrinsic circularity invades empirical assessments of these criteria; the circularity can be diminished but not completely avoided. (Because of it, some clinicians have shifted their terminology to "classification criteria" rather than "diagnostic criteria", a subtle but important difference.)

Is the criterion measurable at all? Social scientists may be willing to accept an unmeasurable criterion (latent trait, or unmeasured concept), but clinicians and administrators certainly are not. For them, the criterion must be present or absent, and not more mystical. When social scientists want professional attention for their validity studies, they must always compare measurable entities; that is an unshakable requirement. From this perspective, Reiser and Eaton's paper is not about validity, but something more introspective to the Diagnostic Interview Schedule.

2. The test variable can also be plain or fancy. Ideally, the test is simpler to conduct, less detailed, or more standardized than the criterion. Some tests are just a single survey question (such as number of hospitalizations in past year). Others involve multiple items; some algorithm must then be found to combine and summarize responses, to form the basis for a final yes or no on the test. Elementary algorithms add up the responses in some manner. Muiltivariate techniques summarize responses with more elegance: Equations that yield the best fit between test items and the criterion are estimated and then used to compute predicted values for individuals. Though seldom used in validity studies to date, multivariate techniques have great advantages because they maximize use of detailed test information.

Reporters

Both the criterion and the test are reported by someone—a health records system or respondent.

Concerns about quality of reporting are prominent in health records systems. Clinicians and administrators are concerned about knowledgeable and timely entry of information in their records: People who fill out the official forms can be quite distant from details of the target event; a prime example is death certificate completions. Paper flow rates in an agency may affect how dates are entered for target events. Radiologists reading x-rays tire and become less painstaking (an issue of intraobserver variability). Physicians are pressured to decrease time spent with patients, thereby learning less during the contact. Different professional observers of the same phenomenon vary in their judgments; this underlies the intense concern about interobserver reliability in clinical research. (In my own experience, it exceeds survey researchers' concern about interviewer effects.)

With respect to survey respondents, knowledge and salience are key features in response. (1) People may simply not know the answer to a question because they are not aware of how the world is formally organized; Perloff and Morris note this in their paper in this volume. The more complex the medical system becomes, the less conceptual information can be expected from respondents. The responsibility to summarize it in theoretical or conceptual terms is ours, not theirs. The lesson in the Perloff and Morris paper is that people should be asked about health events in the way they think, not the way we think. To do good survey research, scientists are obliged to figure out how things are organized; but we cannot ever expect that to be true of respondents. (2) With regard to salience, if the target datum was not especially important to a person, it fades quickly from memory and does not resurface during an interview. What seems important to us may be just too much to ask. The effort to match hospital stays, length, and dates in Calore's paper offers an instructive example. One can expect good recollection of being hospitalized in a time period, but length of stay and exact dates are far less important to the person. Our scientific curiosity can overtake good sense and fairness to respondents. Sometimes, we will ask anyway because the information is critically important to policy or patient care, but that decision must recognize the tough cognitive task posed for respondents.

In short, recorders fill out forms in the context of their whole job and the behavior of a medical record system, and respondents answer questions in the context of life's

hubbub. Contexts compromise people's ability to "get it down right" by scientific standards. Unintentional deviations from the truth as we would like to see it are probably due more to organizational and personal complexity than to conscious decisions to misreport. (The data of Brambilla and associates on a highly sensitive topic, cancer, support this.)

Concept and Subject Matches for the Criterion and Test

1. Do the criterion and test really measure the same concept? In the Calore paper in this volume, respondent reports of medical care visits are compared with billing system records. Billing systems have motivations quite unrelated to patient care, and more so now than medical records systems did several decades ago. There is no one-to-one relationship between visits and bills. To compare them, that relationship has to be constructed; a translation scheme has to be devised that reconfigures bills into visits, or partitions visits to bills. The tables in Calore's paper show how difficult that task is.

The issue of conceptual match can also be stated as a problem of different units of analysis in the two data streams. When there is conceptual distance between the designated criterion and test, one must ask if a validity study is apt. At some point, the distance is so great that one must admit that two very different things are being measured, rather than mount an intensive effort to render them comparable.

2. Subject identification can also differ in the official and survey systems. In this situation, a linkage scheme is required to bring together the relevant records for a respondent. This problem easily arises in our validity studies, since our criterions are typically generated by established medical or health services systems, and our survey is a special record housed in a scientific setting.

When conceptual and subject distances both occur (as Calore faced), the strategy for aligning records and then reconfiguring events is complex and time consuming. The results have to be greatly needed by someone important to justify the effort.

Empirical Match of the Criterion and Test

When the criterion and test are intrinsically dichotomous, so that yes and no are easily scored on both, displaying the empirical results is very simple (Figure 1). But when either has ordinal or continuous gradations, cutpoints are needed to arrive at yes and no. None of the session papers faced this problem, but it is common for medical and health data. For example, if we want to compare x-ray and exam evidence of osteoarthritis of hands, the x-rays (criterion) are graded 0 to 4 and the joint count score (test) can be 0 to 32. Clinical conventions often govern the cutpoints. But they can also be varied during data analysis in an exploratory manner. This is very informative, since sensitivity and specificity levels change as cutpoints do, and one can then empirically derive the points that maximize sensitivity and specificity. An excellent way to portray and evaluate such experimentation is by receiver operator characteristic (ROC) curves.

Consider a challenging multivariate situation: The criterion is an x-ray score (Y, 0 to 4), and the test is a prediction equation based on numerous signs and symptoms (X_i). Predicted values for individuals are generated, either in the form of a predicted score (0 to 4) or a probability if the criterion has been condensed (0 to 1). These $\hat{Y}$ constitute the test scores. Various cutpoints can be tried, and individuals thereby placed in yes or no status for the test; the consequent levels of sensitivity and specificity can be derived. (The cutpoint for the x-ray grade can also be varied.) The ROC curves help the investigator choose the very best cutpoints. Having a clinical collaborator is very helpful in such an endeavor since the final cutpoints have to be clinically reasonable, as well as statistically attractive.

In the general approach described so far, an individual ends up with two dichotomous variables, one for the criterion and one for the test. Whether the scores match or not is readily determined by a cross-tabulation (Figure 1).

That approach cannot be followed when there is significant conceptual distance between the criterion and the test. Then the match between them is consciously constructed and defined; this is called a record linkage scheme. Rules for the TP (yes/yes) cell are made, and the data streams scanned for alignment. Nonmatched events fall into the FP and FN cells. The TN cell is the residual: usually a combination of absent information (no official record exists) with answer of no (for the survey). Typically, the matching rule has stages, with strict match requirements at the start, then relaxed rules, until all acceptable matches are made. Calore's paper in this volume provides an example of this laborious procedure.

Whether the two-variable or record-linkage approach is used, four subgroups are eventually estimated. Then, a decision must be made about any special checks of the FP and FN cases. This question arises more often in record linkage studies, where the two data streams are supposed to both capture the target event. Reports found in one stream but not the other suggest coverage problems in the official system (FP) or conscious and unconscious reporting problems among respondents (FN) or both. Important understanding about record-keeping and response can be gleaned by looking closely at the FP and FN cases, either (a) investigating each case when the sample size is small or (b) finding statistical associations with relevant sociodemographic predictors when the sample is large. Two papers in this volume do detective work on mismatches (Perloff and Morris, Brambilla and associates), and one does statistical work on them (Smith and McElwee).[4]

[4]In clinical situations, once a test is approved and incorporated into practice, only the test value is known for an individual, and not the criterion value. Double-checking the Yes and No statuses (the bottom marginal of FIGURE 1) can be immensely important. Although the test's sensitivity and specificity levels may be known from prior research (so you have a good idea of the number of false cases in each column), you do not know exactly which people have false readings. When the disease is rare, or an invasive procedure is performed on persons with it, then rechecking the positive-test people is critical. When the disease is difficult to detect and life threatening if missed, or if it is extremely common, then rechecking the negative-test people is important.

Importance of Errors

Most validity checks yield some false positives and false negatives. Sensitivity is not 100.0, nor is specificity. The importance of mismatches or errors depends on what kind of decision will be made based on the results. The fundamental question is, can the test substitute for the criterion? Consider specific situations of interest to social and health scientists: In population-based studies, can you replace a physician exam (expensive) by a symptom questionnaire (less expensive)? Can prevalence rates be derived from a multivariate equation based on symptoms and signs?[5] Can identification of cases be standardized in community or clinic samples? Consider some situations of interest to professionals: Does an abbreviated exam yield similar diagnostic information as an extensive one? Will treating false positives hurt them (see footnote 4)? Will missing the proper diagnosis for false negatives shorten their lives?

If scientific knowledge and research are the goals, we take the data as is. After choosing the best test and evaluating its sensitivity and specificity properties, we can then study (a) other measures of association (such as Kappa) and accuracy, (b) risk factors for subgroup assignment, or (c) outcomes such as disability and death among the subgroups. But if interventions on individuals are to be made based on test values, then minimizing the errors is extremely important. Moderate levels of false positives and false negatives may provide grist for the academic mill, but they are completely unacceptable to clinicians. In their view, such tests are totally worthless, not only clinically but scientifically as well.[6]

Human judgment necessarily enters when deciding which kind of error is more important. This is best conveyed by using the language of guilt and innocence: Do you want to be sure to find the guilty (criterion = yes) and settle for scooping up some innocents as well? Or do you want to protect the innocent (criterion = no) at all costs and thereby let some of the guilty escape? In the first example, you abhor FN errors; in the second, FP. You cannot trap all guilty and spare all innocent people at the same time; so which do you care about most? Deciding this brings in considerations besides statistics.

All in all, scientists tend to be more accepting than other professionals about the data from validity studies. The stakes are simply not as high. But professional complaints should be keenly heard and spur scientific efforts to improve the test and reduce the unacceptable error. Listening closely or, better still, engaging in genuine collaborative work will ultimately yield instruments that have both excellent statistical properties and clinical and epidemiological utility.

[5]Signs are clinical findings from physician examination of a patient.

[6]Although disease prevalence does not affect sensitivity and specificity, it can nevertheless be important in decision making about the test's quality. When disease prevalence is high, the number (but not proportion!) of false negatives rises. When disease prevalence is low, the number of false positives rises. There may be a threshold of how many such people an agency is willing to accept; if the test procedure yields too many despite high sensitivity or specificity, it will be rejected.

Sample Bias

If the test is judged to be very good after full review, it may be advocated as a substitute for the criterion in all or specified circumstances. The criterion does not thereby fall from grace; instead, an excellent substitute is available when needed.

Consideration of settings where it will be used now becomes imperative. A test that performs admirably for differential diagnosis in a clinic sample may behave differently in a population-based sample. Good guesses about this are difficult but possible to make. Generally, one can expect lower sensitivity and higher specificity when moving from a clinic to community setting.[7] Why? Consider a test developed in a rheumatology clinic that distinguishes between patients with osteoarthritis (OA) and those without. (1) The latter have other rheumatic diseases (for example, rheumatoid arthritis, scleroderma, lupus) or regional pain syndromes; some of them will be positive on the test anyway because they hurt in some of the same ways OA patients do. When the test is brought to a community setting, where those other rheumatic diseases are infrequent, the chances of false positives diminish. (2) A clinic sample has few OA people with mild symptoms; you have to hurt enough to seek medical help. When a test developed in a clinic setting is placed in the community setting, it will not do a good job of capturing the numerous people with mild or early osteoarthritis; thus, many false negatives occur. In sum, the consequence of increasing false negatives and decreasing false positives is lower sensitivity and higher specificity. Remarkably, this is probably just the direction of changes wanted: If the data are longitudinal, time will help draw in the OA cases (from FN to TP), and you have contaminated the OA sample with few people who really are free of the disease (FP).[8]

In short, the severity of the target disease and the prevalence of related diseases in a sample are key factors that affect the levels of sensitivity and specificity in various settings.

Summary

Doing a validity study is worth the effort if:

1. You have a full design (can estimate false positives and false negatives).
2. There is a professional or policy audience already interested in the matter and ready to collaborate with you.
3. Both the criterion and the test are reputable and interesting.
4. The scores for criterion come from standardized or strongly consensual (clinical judgment) procedures, and from simpler standardized ones for the test.
5. The criterion has good intraobserver and interobserver reliability, and the test items are suitable to respondents' lives and ways of thinking.

[7]A fine example is in the Kulka paper in this volume.

[8]I am grateful to David Felson, a rheumatology colleague at Boston University, for helping me think through the issue of sample bias.

6. The data streams conceptualize the target datum the same way and have the same procedures for subject identification.
7. The rules for positive scores (yes) on each dimension or for a match between them are grounded in clinical or administrative experience, good sense, and interest in maximizing sensitivity and specificity. Compromises on these features will reduce the scientific and professional yield of the validity study.

Once the data are in, decisions about next steps are made: Should the scientific analyses be elaborated? Is the test good enough to replace the criterion in clinical practice, health statistics, or health research? If not, should it be improved or abandoned altogether? Here, the importance and frequency of FP and FN cases (8), and the chance of altered relationships between criterion and test in other settings (9), become central concerns.

Disease Reports in Health Surveys

Validity studies in the past several decades have focused on checking reports of health services use. That is certainly related to the simultaneous growth of public payments for acute and long-term care. But there is equally great need for validity studies of chronic diseases (both physical and mental) and impairments. Medical colleagues scoff at prevalence rates derived from respondent reports in health surveys such as the National Health Interview Survey, and no amount of description about the careful interviewing and coding procedures convinces them.

We have been going about disease rates our way for decades. Perhaps we should start doing it their way. We must bring the perspective of clinical diagnosis into health surveys that rely on self-reports. In return, we offer our medical colleagues the techniques of epidemiology and health survey research for clinical research and thinking. (In fact, clinical researchers have turned in our direction more than vice versa. The term clinical epidemiology is now well-established and reflects several decades of superlative thinking and mentoring in clinical research on patient's well-being and functioning.) The bridge between health survey research and diagnostic knowledge will be made via epidemiology, which provides a language and perspective attractive to both social scientists and clinicians.

How exactly can diagnostic knowledge be brought into (a) estimation of prevalence rates or (b) multivariate research of relationships among risk factors, disease status (caseness), and disease consequences?

Consider how we have approached the matter so far. In health surveys, answers to a single question such as "In the past year, did you have arthritis of any kind?" suffice to identify a case. Validity studies have been conducted, comparing such answers to diagnoses abstracted from respondents' medical records (Balamuth, 1965; Cobb & associates, 1956; Cox & Cohen, 1985; Cox & Iachan, 1987; Daughety, 1979; Densen & associates, 1960; Elinson & Trussell, 1957; Jabine, 1987; Krueger, 1957; Madow, 1973; Scott & associates, 1981; Trussell & associates, 1956). Various rates of mismatch are computed, often including sensitivity and specificity (though not often by those names). These are typically in the range of 30 to 70 percent—certainly intolerably high from a clinical standpoint.

Instead of this, can we find some multivariate rule for identifying cases of a disease? Clinical experience offers knowledge of the symptoms and signs most commonly manifested in the disease. Sometimes a single symptom or sign suffices; but more often particular combinations of them are the clue. Thus, diagnosis proceeds by a multivariate assessment. It is this procedure, statistically modeled into an average clinical behavior equation, that can serve as the foundation for better rates. Equations that use the smallest number of items, and which also yield acceptably high levels of sensitivity and specificity when compared to the criterion (diagnosis), would be sought. The test items could include both questions to respondents (about symptoms, prior diagnosis, even strong risk factors) and evaluations by interviewers (clinical signs).

The point is to bring diagnostic expertise into health surveys, not by requiring the presence of physicians and clinical paraphernalia, but by using a translation device (an equation) that aligns symptoms, signs, and medical history with diagnosis. The result would be both rates and case identification that are medically apt and secured at low cost.

As part of the effort to improve disease reporting, health survey researchers should also pursue the standard strategy of finding good direct questions on disease presence. We should study the utility of probes about prior diagnoses and medical history and probes to clarify colloquial words for chronic conditions. This entails less collaboration with medical colleagues; it should accompany the type of work described above, not supplant it.

An Illustration

To illustrate, a current research project on the concurrence of radiographic, clinical, and symptom evidence of osteoarthritis is described. There are two standard criteria for osteoarthritis in rheumatology: x-ray evidence of changes in cartilage and underlying bone, and clinical diagnosis. X-rays measure pathology but have no direct relevance for patient care since many people with positive x-rays are asymptomatic. Clinical diagnosis is based on a review of medical history, signs (joint exam), symptoms, often x-rays (normally used only to support a presumptive diagnosis), and even disability.

From clinical experience, rheumatologists have noted the low concurrence between signs and symptoms and x-rays. But the empirical literature on the issue is small (especially for population-based samples) (Acheson & associates, 1970; Allander, 1973; Cobb & Rosenbaum, 1956; Cobb & associates, 1957; Cobb, 1971; Davis, 1981; Davis & associates, 1988; Felson & associates, 1987; Gresham & Rathey, 1975; Lawrence, 1977; Lawrence & associates, 1966; Valkenburg, 1981). Both dissatisfaction with x-rays as the gold standard and the desire to standardize diagnostic procedures for selecting patients into

clinical trials have led rheumatologists to develop diagnostic criteria. These are short lists of items that signal clinical diagnosis of particular rheumatic diseases. The requirement for caseness (test = yes) is sometimes "n or more items present", sometimes certain combinations of items present (Altman & associates, 1983, 1986, 1987; Schumacher, 1988).

A study is being conducted at the University of Michigan that gathers five types of evidence of osteoarthritis for hand and knee from sampled individuals: x-rays, clinical diagnosis, joint exam, chronic joint symptoms, and (for hands only) regular photograph. Prediction equations will be estimated with clinical diagnosis (or x-ray evaluation) as the dependent variable, and arrays of signs, symptoms, and photograph readings as the independent variables. The main question is, how well can clinical diagnosis of osteoarthritis (or radiographic grade of osteoarthritis) be predicted from a data array of symptom reports, joint exam, and photograph? Various models will be estimated, and cutpoints for caseness (test = yes) varied. Then ROC curves will be used to detect, visually or statistically, the equations that maximize sensitivity and specificity.

Two samples will be studied separately: 100 patients from a rheumatology outpatient clinic, and 100 from a geriatric medicine clinic. The latter is as close to a community sample as possible, given the constraint that this project must obtain the medical data in established medical settings. Differences in the prediction equations for the two samples will be evaluated. The best equations developed for each will be tried on the other's data, and measures of sensitivity and specificity derived and then compared. We shall end up with important clues about how sample bias affects the relationships between criterion and test.

Ultimately, the project information will have utility in three domains: scientific, clinical, and health survey research methods. With respect to the latter, can diagnostic status be closely predicted by commonplace instruments: a health survey, a photograph, and an interviewer-administered joint exam?[9] This question has prompted both laughter and applause among rheumatology colleagues. Both are important to hear. The laughter probably comes from conviction that no prediction equation can ever substitute for clinical judgment; thus, an unwillingness to study the issue empirically. The applause reflects the intense interest among rheumatologists in epidemiological research on osteoarthritis (and rheumatoid arthritis). They rely on us to do the quantitative work that maps signs and symptoms to diagnoses. They are genuinely curious to see if a technically simpler strategy than represented by the National Health and Nutrition Examination Survey, which includes physician examination and x-rays, can yield fine data. With such fine colleagues ready, what on earth are we waiting for?

[9]In the project, the joint exam will be performed by a physician or nurse. If it proves valuable in the prediction model, it will be developed for use by trained lay interviewers.

References

Acheson, R. M., Chan, Y.-K., & Clemett, A.R. (1970). New Haven Survey of Joint Diseases XII: Distribution and symptoms of osteoarthritis in the hands with reference to handedness. *Annals of the Rheumatic Diseases,* 29, 275-286.

Allander, E. (1973). Conflict between epidemiological and clinical diagnosis of rheumatoid arthritis in a population sample. *Scandinavian Journal of Rheumatology,* 2, 109-112.

Altman, R. D., Asch, E., Bloch D., & associates (1986). Development of criteria for the classification and reporting of osteoarthritis: Classification of osteoarthritis of the knee. *Arthritis and Rheumatism,* 29, 1039-1049.

Altman, R. D., Bloch, D. A., Bole, G. G. Jr., & associates (1987). Development of clinical criteria for osteoarthritis. *Journal of Rheumatology,* 14 (Suppl.), 3-6.

Altman, R. D., Meenan, R. F., Hochberg, M. C. & associates (1983). An approach to developing criteria for the clinical diagnosis and classification of osteoarthritis. *Journal of Rheumatology,* 10, 180-183.

Balamuth, E. (1965). Health interview responses compared with medical records. *Vital and Health Statistics* (Series 2, No. 7). Washington, DC: Public Health Service.

Cobb, S. (1971). *The frequency of the rheumatic diseases.* Cambridge, MA: Harvard University Press.

Cobb, S., Merchant, W. R., & Rubin, T. (1957). The relation of symptoms to osteoarthritis. *Journal of Chronic Diseases,* 5, 197-204.

Cobb, S. & Rosenbaum, J. (1956). A comparison of specific symptom data obtained by nonmedical interviewers and by physicians. *Journal of Chronic Diseases,* 4, 245-252.

Cobb, S., Thompson, D. J., Rosenbaum, J., & associates (1956). On the measurement of prevalence of arthritis and rheumatism from interview data. *Journal of Chronic Diseases,* 3, 134-139.

Cox, B. G. and Cohen, S. B. (1985). A comparison of household and provider reports of medical conditions. In *Methodological issues for health care surveys* (pp. 150-189). New York: Marcel Dekker, Inc.

Cox, B. G. and Iachan, R. (1987). A comparison of household and provider reports of medical conditions. *Journal of the American Statistical Association,* 82, 1013-1018.

Daughety, V. S. (1979). Illness conditions. In R. Andersen, J. Kasper, M.R. Frankel, & associates. *Total survey error* (pp. 52-74). San Francisco, CA: Jossey-Bass.

Davis, M. A. (1981). Sex differences in reporting osteoarthritic symptoms: A sociomedical approach. *Journal of Health and Social Behavior,* 22, 298-310.

Davis, M. A., Ettinger, W. H., Neuhaus, J.M. & associate (1988). Sex differences in osteoarthritis of the knee: The role of obesity. *American Journal of Epidemiology,* 127, 1019-1030.

Densen, P. M., Balamuth, E., & Deardorff, N. R. (1960). Medical care plans as a source of morbidity data. *Milbank Memorial Fund Quarterly,* 38, 48-101.

Elinson, J. & Trussell, R. E. (1957). Some factors relating to degree of correspondence for diagnostic information as obtained by household interviews and clinical examinations. *American Journal of Public Health,* 47, 311-321.

Felson, D. T., Naimark, A., Anderson, J., & associates (1987). The prevalence of knee osteoarthritis in the elderly: The Framingham Osteoarthritis Study. *Arthritis and Rheumatism,* 30, 914-918.

Fletcher, R. H., Fletcher, S. W., & Wagner, E.H. (1988). *Clinical epidemiology: The essentials* (Chapter 3). Baltimore, MD: Williams & Wilkins.

Gresham, G. E. and Rathey, U. K. (1975). Osteoarthritis in knees of aged persons: Relationship between roentgenographic and clinical manifestations. *Journal of the American Medical Association,* 233, 168-170.

Hulley, S. B. and Cummings, S. T. (1988). *Designing clinical research: An epidemiologic approach* (Chapter 9). Baltimore, MD: Williams & Wilkins.

Jabine, T. B. (1987). Reporting chronic conditions in the National Health Interview Survey. *Vital and Health Statistics* (Series 2, No. 105. DHHS Publ. No. (PHS) 87-1379). Hyattsville, MD: National Center for Health Statistics.

Kelsey, J. L., Thompson, W. D. & Evans, A. S. (1986). *Methods in observational epidemiology* (Chapter 11). New York: Oxford University Press.

Krueger, D.E. (1957). Measurement of prevalence of chronic disease by household interviews and clinical evaluations. *American Journal of Public Health,* 47, 953-960.

Lawrence, J. S. (1977). *Rheumatism in populations.* London: William Heinemann Medical Books.

Lawrence, J. S., Bremner, J. M. & Bier, F. (1966). Osteoarthrosis: Prevalence in the population and relationship between symptoms and x-ray changes. *Annals of the Rheumatic Diseases,* 25, 1-24.

Madow, W. G. (1973). Net differences in interview data on chronic conditions and information derived from medical records. *Vital and Health Statistics,* (Series 2, No. 57. DHEW Publ. No. (HSM) 73-1331). Rockville, MD: National Center for Health Statistics.

Marquis, K. H. (1978). Inferring health interview response bias from imperfect record checks. *Proceedings of the American Statistical Association (Survey Research Methods Section)* (pp. 265-270). Washington, DC: American Statistical Association.

Marquis, K. H. (1984). Record checks for sample surveys. In T. B. Jabine, M. L. Straf, J. M. Tanur, and R. Tourangeau (Eds.). *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 130-147). Washington, DC: National Academy Press.

Sackett, D. L., Haynes, R. B. & Tugwell, P. (1985). *Clinical epidemiology* (Chapter 4). Boston, MA: Little, Brown and Co.

Schumacher, H. R. Jr. (Ed.). (1988). *Primer on the Rheumatic Diseases* (9th ed.). Atlanta, GA: The Arthritis Foundation.

Scott, B., Brook, R. H., Lohr, K. N. & associate (1981). *Conceptualization and measurement of physiologic health for adults. Vol. 10: Joint Disorders.* (Rand/R-2262/10-HHS). Santa Monica, CA: The Rand Corporation.

Trussell, R. E., Elinson, J., & Levin, M. L. (1956). Comparisons of various methods of estimating the prevalence of chronic disease in a community. The Hunterdon County Study. *American Journal of Public Health,* 46, 173-182.

Valkenburg, H. A. (1981). Clinical versus radiological osteoarthrosis in the general population. In J.G. Peyron (Ed.). *Epidemiology of osteoarthritis,* (pp. 53-58). Paris: Geigy.

DISCUSSION PAPER

Validity of Reporting in Surveys

Nancy A. Mathiowetz

Introduction

All these papers addressed the question of validity, that is, the extent to which the responses to survey questions measure the phenomena of interest. To the extent that survey data contain errors and are therefore invalid, estimates based on the data will be at best, problematic, and at worst, misleading. Nonresponse bias, sampling errors, measurement error, and interviewer bias are among the factors that contribute to errors in, and therefore the invalidity of, survey data.

The problem of errors in surveys has been of interest to both methodologists and users of survey data for the past 35 years. Beginning with the 1945 Census of Agriculture (Mauldin & Marks, 1950), researchers have enumerated, measured, and attempted to reduce the sources of error. This research has focused on the effects of questionnaire topic and wording (Bradburn & Sudman, 1979; Schuman & Presser, 1981), bias related to interviewer behavior (Cannell & associates, 1977b), mode of questionnaire administration (Groves & Kahn, 1979), characteristics of the interviewer (Schuman & Converse, 1971), and characteristics of the respondents (Cannell & associates, 1977a).

One general model of survey error has been proposed by Andersen and co-workers (1979) based on the work of Kish (1965). The model classifies survey error as either variable error or bias and then further classifies the error or bias according to the source, sampling or nonsampling. The major classification distinction, between variable error and bias, separates random errors from systematic errors. The latter type of error, bias, is considered more detrimental to the quality of survey data than random error, since systematic errors affect any sample with the same constant error. Variable errors are assumed to be random with a mean of zero across replicate samples.

Nancy A. Mathiowetz is with the National Center for Health Services Research and Health Care Technology Assessment, Rockville, Maryland.

Nonsampling bias is believed to be the largest contributor to total survey error. In part because nonsampling errors are difficult or impossible to measure, they are often ignored by the survey analyst or methodologist. "Nonobservation" bias, referring to the bias associated with nonresponse, is often mentioned in survey literature, but its effects are usually not well understood. For example, most users of survey data require high response rates for their studies, believing that the distribution of answers from nonrespondents might differ from those who cooperate with the survey. However, few studies have had the opportunity to question nonrespondents to determine the validity of the notion that higher response rates imply less biased data.

The sources of observation bias are numerous and are associated with the data collection process and the postsurvey data processing. Errors in data collection include factors related to the questionnaire, the interviewer, or the respondent, whereas postsurvey data processing bias may occur during the coding, keying, or analysis of the data. The focus here will be on those errors related to the questionnaire, the respondents, or the interviewers, a group of errors referred to as response errors.

Errors in Surveys: Respondents, Questionnaires, and Interviewers

The sources of response errors can be separated into three groups: errors related to the task, the respondent, and the interviewer. Sudman and Bradburn (1974) state that task variables are the "most important sources of response effects." Task variables include such factors as the nature of the question (subject, social desirability, salience), the administration of the question (open or closed question, position of the question within the survey instrument, question wording), and the interviewing situation (mode of administration, length of the interview). Task variables also include factors related to the retrieval of accurate information, such as the length of

the reference period, the availability of records, and the use of aided recall.

Although Sudman and Bradburn (1974) classify errors into these three groups, the classes are not discrete. For example, a question is not in and of itself socially desirable. A question may focus on topics considered sensitive, but it is the interaction of the question with the particular respondent and/or social climate at the time the question is asked that determines whether responses to a question have socially desirable or undesirable characteristics.

Several studies have evaluated the effects of questionnaire task on response error. The effects of length of recall period, salience and social desirability on reports of hospitalizations, visits to physicians, and reporting of health conditions were detailed by Cannell and colleagues (1977a). They found that errors of omission increased as the length of time between the interview and the event of interest was extended. Events judged to be more important to the respondent were better reported than less salient events and that socially undesirable reports, in this case conditions, were subject to higher rates of response error than conditions not subject to social evaluation. Similar results were reported by Bradburn and Sudman (1979). They found that the degree of distortion between survey reports and validation information was positively related to the level of social threat. The direction of response effects was also related to social desirability, with respondents overreporting rates of socially desirable behavior (voter registration) and underreporting rates of undesirable behavior (drunken driving).

Question length, questionnaire content, and administration have also been shown to be task variables related to level of response errors. For example, question length does not seem to affect the distribution of yes or no responses to behavior questions, but does affect the counts of threatening behavior (Bradburn & Sudman, 1979). Several studies have shown little or no difference in response distributions for face-to-face and telephone interviews (Thornberry, 1987). However, Groves & Kahn (1979) report that length of responses to open-ended questions is affected by mode of administration, with shorter responses being elicited from telephone respondents. Analysis of questionnaire content has indicated that responses to both attitude and factual items can be manipulated by the placement of the question in the survey instrument (Schuman & Presser, 1981; Turner & Krauss, 1978).

Previous research has focused on two aspects of respondent characteristics. Demographic characteristics have been shown to be related to differential error rates in various surveys, although the findings across studies are not consistent. Using data from the 1949 Denver Community Survey, Cahalan (1968) found that younger respondents were more inaccurate than other respondents, especially with respect to voting behavior. Cannell and others (1977a) report that responses from younger respondents were subject to lower rates of errors of omissions in reports of hospitalizations and visits to physicians than older respondents. However, in the same report, these authors state that older respondents were more accurate in reporting chronic conditions than younger respondents.

The relationship between education and response effects has tended to indicate that higher rates of response effects are associated with lower levels of education. A correlation between higher education and lower levels of response error has been found in studies of reporting voting behavior (Presser, 1984), visits to physicians (Cannell, 1977a), and consumer purchases (Sudman & Bradburn, 1974).

The respondent's perceptions of the interview and motivation to complete the task constitute the second class of respondent characteristics related to response effects. Bradburn and Sudman (1979) found that respondents' level of anxiety concerning questions affected their reporting levels. Motivated respondents are believed to work harder at the interviewing task and may therefore complete a more thorough memory search before responding to a question. Second, if a respondent is motivated to perform the task, he or she will be more likely to report socially undesirable behaviors.

Behavior, perceptions and expectations, and demographic characteristics are all means by which interviewers affect response errors. Several researchers have measured the contribution of interviewers to measures of variance (Freeman & Butler, 1976; Groves & Magilavy, 1986). Cannell and Lawson (1971) have documented the extent to which interviewers' behavior varies during the course of an interview. They showed that interviewers tended to change question wording, failed to probe for adequate responses, and provided directed feedback to the respondents. The relationship between interviewer behavior and response effects, however, is not well understood. Attempts to link interviewer behavior and measures of interviewer variance have not proved to be successful (Groves & Magilavy, 1986).

Interviewer's perceptions and expectations have been linked to response distributions. For example, Sudman and associates (1977) found that interviewers who perceived the respondent task as difficult were more likely to have a lower percentage of respondents reporting certain activities than interviewers who did not see the task as difficult. In the same study, the authors reported that interviewers who expected respondents to underreport got lower levels of reporting than other interviewers.

Demographic characteristics of the interviewer are also related to response distributions. Race of interviewer has been correlated to responses not only for question items related to race but to other questions as well (Schuman & Converse, 1971). Interviewer gender has been correlated to responses (Groves & Fultz, 1985). Other interviewer attributes such as voice characteristics and intonation have been shown to affect respondents in telephone surveys (Oksenberg & associates, 1986).

Discussion

As is true with previous research, the papers presented in this volume force us to question the validity of data from surveys. Of particular interest are the level of

omissions and the reporting of false positives. One popular approach to assessing the validity of responses has been to compare survey data to supplementary information; for health care surveys, usually from records from doctors' offices, hospitals, or insurance claims. This is the approach taken by four of the authors. All found, to varying degrees, differences between the survey data and the particular validation data.

Miller and Groves (1985) question the use of matching criteria as a basis for assessing error in social surveys. They found, using varying criteria and both human and computer matching, that the proportion of events matched in comparing reports of victimization with police records was between 0.14 and 1.0. They argue that the reporting of a single match rate, as is often done in validation studies, suggests a level of confidence in the estimate that is impossible to determine. Too often, they feel, researchers have examined the bottom line match rate of these validation studies and interpreted the results as the extent to which survey data depart from truth.

The question concerning the quality of the validation data has not only to do with the completeness of a particular record, but also with the need to obtain information for the set of all possible events that may have been subject to forgetting and misreporting. The various approaches to validation studies have been outlined by Marquis (1978); in his work, Marquis outlines the three basic kinds of record check designs: prospective, retrospective, and full designs. Prospective designs refer to cases in which positive responses from the respondent, for example, stating that the individual had been hospitalized sometime during the reference period, are then validated against records. Retrospective designs sample the events from records and then ask the respondents to report on the attribute of interest. Full designs refer to record check procedures where all reports of individuals, whether positive or negative, are verified against record data following the survey. All positive and negative reports, from both the survey and the record data, are compared to the opposite source of information. Marquis' (1978) study points out that only a complete design which includes validation information for the set of all possible events (the full design) will provide unbiased estimates of the levels of omission and inclusion errors in the survey data.

Finally, it is always useful to look at the possible sources of error in the validation data. Record data are often accepted as more valid since they are not subject to some of the sources of error known to be present in survey data. We hope that in using record data we have eliminated recall error, due either to the respondent's inability to recall an event or desire to protect self-image; interviewer error; error related to miscommunication of the definition of the event we are trying to measure; and error related to the interaction of the questionnaire, the interviewer, and the respondent. However, record data are not without fault and are subject to error from numerous sources.

The Calore paper compares claims from Medicare to reports of hospitalizations in the National Medical Expenditure Survey. The findings parallel other work that has been done in this area, namely, the research of Cannell and associates (1977a) and Cannell and Fowler (1963). The majority of hospitalization events could be matched, a fact that may be attributable to characteristics of hospitalizations:

- Hospitalizations are relatively unique events as compared to doctor visits; this often results in better reporting on the part of the respondent. As noted by Calore, 56.5 percent of the hospitalizations reported in NMES exactly matched the admission or discharge date reported in MADRS.
- The uniqueness of the events permits more flexibility in the matching rules than for 'series' events. The rules used by Calore, allowing two events to match if either admission or discharge dates were in the same month for the two records or within thirty days, resulted in an additional 24.4 percent of the NMES events matching to MADRS.

However, the paper raises questions concerning the quality of validation data and the appropriateness of attempting the matches. With respect to the quality of the data, the MADRS claims include what are labeled "physician claim without hospital claim." How is it possible to have a physician charge when there is no corresponding hospital claim? Does this suggest that the MADRS data are incomplete with respect to hospitalizations? Or are these overreports?

The Perloff and Morris paper, validating the reports of usual source of care, forces us to examine a measure that has been used in health surveys for the past 30 years. Although the sample, based on low income black women in the Chicago area, could certainly not be regarded as representative of the population, it represents a group of interest to health researchers. Perloff and Morris point out, referencing Sudman and Bradburn, that you should not ask respondents questions they cannot answer. The response categories provided in the question, private physician's office, hospital outpatient clinic, health department clinic, emergency room of a hospital, other clinic or some other place, are ill-defined and confusing. Where would you classify the so-called "doc in the box" that has grown in popularity? The authors direct the research community to assess the question and definitions provided to the respondents and urge that more detail be given so as to assist the respondent.

There is one criticism of the paper, which is once again related to the quality of the gold standard; the authors state they "used various secondary sources to independently determine the type of usual place named by each respondent," further citing the American Hospital Association Guide and published information from the Chicago Department of Health. Who did the classification? One person? Is there any way to determine the accuracy of the verification data?

The Smith and McElwee paper discusses a study in which records of reported poisoning of children were sampled and an adult member of the family was interviewed concerning child poisonings. The findings support other work that has been done on recall of events—mild poisonings, those that we would tend to label less salient, were less well recalled than severe poisonings.

However, in contrast to most of the literature, the length of time between the date of the poisoning and the date of the interview was not significant in predicting rates of omissions or errors concerning details about the event when other factors describing the nature of the poisoning were controlled. This parallels research by Mathiowetz (1988), which indicated that it was not the length of time but rather the occurrence of intervening events that resulted in poorer recall.

The fact that each of these poisonings involved a telephone call to a poison center biases the findings to some unmeasurable extent. That each family placed a call to the poison center would, according to cognitive research, reinforce the occurrence of the poisoning.

The paper by Brambilla and colleagues examining reports of cancer incidence, represents a prospective survey study with a full-design record check that permits the researcher to determine both the level of false positives as well as the false negatives. These findings show small numbers of both false positives and false negatives. Of the 13 false positives, 1 was credited to interviewer error (no discussion of how this determination was made); 8 were found to be nonmalignant growths; and 4 were found to be correct reports. The latter point indicates that a small proportion of incident cases was missing from the cancer registry. However, the presence of false negatives in the validation data also should suggest to the researchers that there may be false positives.

The Reiser and Eaton paper, which examined the validity of two questions in the Diagnostic Interview Schedule, points out the problem of using "vague" terms in a survey. Schaeffer (1988), in a paper on gender differences in the use of vague quantifiers, cited a scene from the movie "Annie Hall,"

> On one side Alvie Singer talks to his psychiatrist; on the other side Annie Hall talks to hers. Alvie's therapist asks him "How often do you sleep together?" and Alvie replies "Hardly ever, maybe three times a week." Annie's therapist asks her, "Do you have sex often?" Annie replies, "Constantly. I'd say three times a week."

The vagueness of responding to frequency questions is also applicable to vague terms such as felt sad, blue, depressed, or whether persons classified themselves as nervous. The authors suggest that the small proportion of individuals who respond positively to the two questions of interest, as compared to the specific symptoms indicative of dysphoria and anxiety, is related to the difficulty of drawing inferences about oneself or the undesirability of admitting to the disease rather than just the symptoms. My reaction, both from reading the paper and from working with Diagnostic Interview Surveys, is that individuals do not respond affirmatively to these questions because they are vague.

Conclusions

Survey methods research has several goals with respect to response error or the validity of reports. The first is to measure the extent to which errors exist in survey data. The second is to determine, if possible, the source or sources of the error. The third is to develop means by which to reduce the error, and the fourth is to provide the users of survey data with the means with which to analyze data in the presence of these errors.

All of the papers attempt to measure the extent of error in their respective survey data. To the extent we accept the validation data as "truth" we see once again that survey data can be both incomplete and inaccurate as well as relatively well reported. Work completed in the 1960s and 1970s pointed out how vulnerable survey data are to the actors in the process, including the respondent, the interviewer, and the task. Not only do we need to continue to measure error in survey data, we need to challenge ourselves with the following questions:

- Are the sources of validation data and our means for assessing error, specifically matching, adequate?
- How can we best reduce the error or analyze data in the presence of error? The measurement of error is an academic exercise if we do not use the information to improve our designs or analyses.

The question is the most important aspect of the survey process to reexamine. We ask respondents to do difficult if not impossible tasks. Whether it is reporting on events that have happened over the last 3 months or classifying an event that happened yesterday, we often provide the respondents with little information and even less time to formulate responses. Shaping health policy is one of the many applications of survey data; errors in the data undermine our ability to rely on the recommendations we make. Perhaps as we assess the validity of survey data we need to question what it is reasonable to expect from the process.

References

Andersen, R., Kasper, J., Frankel, M., & associates (1979). *Total survey error.* San Francisco: Jossey-Bass.

Bradburn, N., Sudman, S., & associates (1979). *Improving interview method and questionnaire design.* San Francisco: Jossey-Bass.

Cahalan, D. Correlates of respondent inaccuracy in the Denver validity survey (1968). *Public Opinion Quarterly,* 32, 607-621.

Cannell, C. F. & Fowler, F.J. (1963). *A study of the reporting of visits to doctors in the National Health Survey.* Ann Arbor, MI: Survey Research Center.

Cannell, C. F. & Lawson, S. A. (1971). "Analysis of Individual questions" In J. B. Lansing (Ed.). *Working papers on survey research in poverty areas.* Ann Arbor, MI: The Survey Research Center.

Cannell, C. F., Marquis, K. H., & Laurent, A. (1977a). *A summary of studies of interviewing methodology.* (Series 2, No. 69 Vital and Health Statistics,) Washington, DC: U.S. Government Printing Office.

Cannell, C. F., Oksenberg, L., & Converse, J. (1977b). *Experiments in interviewing techniques: field experiments in health reporting,* 1971-1977. (Publication No. 78-3204) Washington, DC: Department of Health, Education, and Welfare.

Freeman, J. & Butler, E. W. (1976). Some sources of interviewer variance in surveys. *Public Opinion Quarterly,* 40, 79-91.

Groves, R. M. & Fultz, N. (1985). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research,* 14, 31-52.

Groves, R. M. & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Academic Press.

Groves, R. M. & Magilavy, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly,* 50, 251-266.

Kish, L. (1965). *Survey sampling.* New York: Wiley & Sons.

Marquis, K. H. (1978). *Record check validity of survey responses: A reassessment of bias in reports of hospitalizations.* Santa Monica, CA: Rand Corporation.

Mathiowetz, N. (1988). The applicability of cognitive theory to long-term recall questions in social surveys. Unpublished dissertation. The University of Michigan, Ann Arbor.

Mauldin, W. P. & Marks, E. S. (1950). Problems of responses in enumerative surveys. *American Sociological Review,* 15, 649-657.

Miller, P. V. & Groves, R. M. (1985). Matching survey responses to official records: Validity in victimization reporting. *Public Opinion Quarterly,* 49, 366-380.

Oksenberg, L., Coleman, L., & Cannell, C. F. (1986). Interviewer's voices and refusal rates in surveys. *Public Opinion Quarterly,* 50, 97-111.

Presser, S. (1984). Is inaccuracy on factual survey items item-specific or respondent specific? *Public Opinion Quarterly,* 48, 344-355.

Schaeffer, N. C. (1988). Sex differences in the use of vague quantifiers. Paper presented at the annual meeting of the American Association for Public Opinion Research.

Schuman, H. & Converse, J. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly,* 35, 44-66.

Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys.* New York: Academic Press.

Sudman, S. & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis.* Chicago, IL: Aldine.

Sudman, S., Bradburn, N., Blair, E., & associate (1977). Modest expectations: The effects of interviewer's prior expectations on response. *Sociological Methods and Research,* 6, 177-182.

Thornberry, O. (1987). *An experimental comparison of telephone and personal health interview surveys.* (Series 2, No. 106 Vital and Health Statistics,) Washington, DC: U.S. Government Printing Office.

Turner, C. F. & Krauss, E. (1978). Fallible indicators of the subjective state of the nation. *American Psychologist,* 33, 456-470.

SESSION SUMMARY

Validity of Reporting in Surveys

Deborah M. Winn, Recorder, and Daniel C. Walden, Chair

The discussion centered on two questions: First, what are the major methodological problems of validity of reporting in surveys and how do they affect data quality? Second, what procedures or techniques have been found effective in addressing such problems?

Four of the five papers in this session compared household respondent reports of health events with data from administrative records. Conference attendees agreed that the match between record and survey data is subject both to how well respondents report health events and to the nature, purpose, and quality of the administrative record system. They agreed that the value of making the comparison at all should be carefully scrutinized. Accordingly, the session's concluding discussion focused on the quality of self reports, the quality of administrative record data, and the problems of matching the two.

Smith's paper stimulated a discussion of the selection of respondents in household surveys. Significant societal changes have influenced the reporting of health events, and two types of changes were mentioned. First, family composition and interrelationships have changed over time, and family members may have less knowledge of each other's activities due to physical separation (such as living apart or joint custody situations) or simply due to different lifestyles within a household. For health surveys, the result is that it is often difficult to identify a good questionnaire informant. In the study of poisoning episodes experienced by children, Smith pointed out that family members often were not well informed about a poisoning episode unless they themselves had witnessed or reported the event to the poison control center. The author concluded that information about an acute, potentially serious health problem occurring to a child may not be shared among adult household members, as usually assumed. This indicates the importance of selecting respondents who have access to information that is wanted.

A second societal trend concerns the willingness of respondents to admit to (and report) medical problems once thought to be stigmatizing. Participants mentioned other recent studies with findings similar to Brambilla's in which household respondents have accurately identified themselves as cancer patients (Czaja and others, 1984). These studies have shown that there is less denial of the diagnosis of cancer than had been found in earlier studies which suggests that the reporting of cancer probably is no longer subject to the same problems of underreporting in survey research as other stigmatizing diagnoses, such as HIV infection or AIDS.

Discussion on the quality of administrative records led to the conclusion that survey researchers should not use administrative record data as a "gold standard" or even a "gold-plated standard" in their comparisons with reports from household respondents. Most record systems are designed to serve other primary goals. Record systems also may contain errors. For example, the MADRS system of the Health Care Financing Administration is a billing system which has only secondarily been used for linkage with questionnaire data. The well-known and high quality SEER cancer registry data set was reported to have 5 percent fewer records of cancer patients than are believed to exist by comparing SEER data with data from other sources (Young and others, 1981). One participant cautioned researchers about assuming that physician records contain complete data on the use of Pap smears and noted that household respondents, particularly black women, are believed to overreport the use of the Pap smear procedure. Similar concern was expressed about hospital records. New prospective payment systems have created incentives for hospitals to classify stays into diagnosis related groups that will benefit them financially. Several participants documented the impact of these payment systems on diagnostic codes (Baker and Kronenfeld, in process). In one hospital-based study, the apparent incidence of can-

Deborah M. Winn is with the Division of Health Interview Statistics, National Center for Health Statistics. Daniel C. Walden is with the Division of Intramural Research, National Center for Health Services Research and Health Care Technology Assessment.

Background of the Health Survey Methods Conference Series

Norman W. Weissman

The National Center for Health Services Research and Health Care Technology Assessment (NCHSR) and the National Center for Health Statistics (NCHS) have a tradition of providing support for the improvement of health survey methods and of supporting these conferences. The original intention was to have a conference every two years, but circumstances and resources were such that this was not possible. However, this meeting continues the outstanding tradition generated by the four previous conferences over the past 14 years.

At the first conference at Airlie House in 1975, topics were specified but no papers were prepared in advance. Invited researchers reported their experiences in each of the topic areas.

In 1977 in Williamsburg, topics were specified but in a more structured way. For each topic a speaker was invited to present a position paper summarizing major findings and pointing out areas of unsolved problems and research needs.

The format was more structured for the third conference in Reston in 1979, and for the fourth in Washington, D.C. in 1982. Selection was made from among many submitted papers on specified topics, and a discussant commented on each of the papers. Open discussion followed each paper and in a general discussion later in the session. The papers and summaries of the discussion were incorporated into the conference proceedings.

The format for this fifth conference in 1989 is similar to that of the third and fourth conferences. Papers were selected from well over 100 submitted abstracts. The process was further refined by grouping the papers on related topics and having two discussants provide comments on all five papers together, followed by an open discussion.

The three topic areas include improved methods for (1) designing questions, (2) collecting data from special groups, and (3) improving data quality. They are all timely in the present research and policy context. They should provide important methodological contributions, and the proceedings, like its predecessors, will be a valuable resource for researchers and students.

The proceedings of the previous meetings are greatly valued by the health services research community and its students. When we have made them available for distribution they have been among the most avidly consumed. Survey research will continue to be a fundamental part of health services research and will share in its growth for the foreseeable future. I am delighted by the contribution already being made by survey researchers to the Medical Treatment Effectiveness Initiative.

Norman W. Weissman is director of the Division of Extramural Research, National Center for Health Services Research and Health Care Technology Assessment.

The Role of the National Center for Health Statistics in Meeting the Needs for Health Data

Manning Feinleib

The Changing Climate of Health Statistics

Most readers of these proceedings are probably familiar with the recent report, *The Future of Public Health* (Institute of Medicine, 1988). Among the many recommendations, several related to the need for and role of health statistics. The report recommended that "every public health agency regularly and systematically collect, assemble, analyze, and make available information on the health of the community, including statistics on health status, community health needs, and epidemiologic and other studies of health problems." Among the Federal public health obligations cited in the report is the "support of knowledge development and dissemination through data gathering, research, and information exchange . . . and the provision of technical assistance to states and localities. . . . In addition to conducting research directly, the Federal government should support research by states, localities, universities, and by the private sector."

This emphasis by the Institute of Medicine reflects important social, demographic, political, and technological changes that have impacted on the need and demand for health data and, subsequently, on the role of health statistical agencies. Each of these presents numerous challenges to survey methodologists, on whom we are dependent for leadership, expertise, and guidance as to the best methods and procedures for the collection and analysis of data.

For example, some changes result from alterations in the nature of the impact of disease and disability on our society. The effective implementation of community-level public health measures during the latter half of the 19th century and the first part of the 20th, which aimed at providing clean water, pure food, sanitation, and the elimination of disease vectors, led to the prevention and control of many infectious diseases. The emphasis has now shifted to health promotion and the prevention of chronic diseases, violence, substance abuse, and human immunodeficiency virus (HIV) infection, all of which focus on changing individual behaviors. This requires accurate data in areas previously considered soft or more qualitative, such as knowledge and attitudes and health-related behaviors. This new emphasis also requires the development of measures that are real indicators of outcomes, so that changes in indicators reflect changes in health status and other outcome variables. A number of our current and past indicators have shown changes over time that may well have been artifacts of improved survey methodology, changes in respondents' willingness to report events or behaviors, or changes in utilization and diagnostic patterns rather than real changes in health status or behavior.

These issues are of particular importance to the National Center for Health Statistics (NCHS) because it has been designated the agency with lead responsibility for the priority area: Improve Surveillance and Data Systems for Promoting Health/Preventing Disease: Year 2000 Objectives for the Nation. This gives NCHS the lead role in providing through our data systems the intelligence that is needed to track progress in many of the other 20 priority areas. It is clear that improvements in survey methods and analytic models are needed if NCHS is to meet this mandate. Models need to be developed for projecting trends; assessing the impact of interventions; assessing the impact of meeting individual objectives on broader goals, for example, the impact on life expectancy of reducing selected chronic diseases to projected levels; assessing the consistency between numerical objectives for population subgroups, for example, are total reduction targets consistent with targets set for minority groups?; and understanding competing risks in the assessment of progress toward individual objectives.

Improvements are also needed in methodologies for measuring and presenting information on health. Along with the improvements to the methodology of measuring

Manning Feinleib is Director of the National Center for Health Statistics, Hyattsville, Maryland.

years of potential life lost (YPLLs), new measures such as quality adjusted life years (QALYs), which can take disability into account, need to be further developed. At the same time, it will be necessary to improve statistical aspects of measurement of exposure to risk, adapting new research technologies to national data systems.

There have also been changes in serious threats to health and well being. The impact of certain behaviors like drug use and high-risk sexual behavior have necessitated a reevaluation of data priorities. There is an increasing demand for data on these and other subjects that traditionally have not been discussed or have not been well received by the public owing to their sensitive nature. Because of the urgent need for these data, surveys on HIV and the acquired immunodeficiency syndrome (AIDS) epidemic are appearing routinely with little methodological development or assessment of data quality. New ground is currently being broken with the initial success of the National Household Seroprevalence Survey and the National Survey of Health and Sexual Behavior. Some progress has been made in the last decade in the use of the survey mechanism to collect data in selected sensitive areas. The National Health Interview Survey (NHIS), for example, has recently addressed issues of alcohol consumption, chronic mental illness, and digestive disorders; the National Health and Nutrition Examination Survey (NHANES) has asked questions on sexual experience and on the use of selected drugs; and of course, the National Survey of Family Growth (NSFG) has for many years asked sensitive questions specific to reproduction. It is obvious that research needs to be expanded to evaluate and improve the quality of these data.

Other significant changes impact on health statistical systems. Never before have health status and health care changed at such a rapid rate. Shifts in the demographic composition of the United States population—the graying of America—have resulted in dramatic changes in health care demand and utilization, restructuring of services for cost containment, and a change in the role of the government in promoting healthy behavior.

This has resulted in public pressure for government accountability for its programs and for the evaluation of the effectiveness of these very expensive legislated mandates. In turn, considerable pressure is put on statistical agencies to produce data of unquestionable quality and maximum utility with minimum turnaround time. This situation has been described by Bonnen (1983) as a "growing intimate embrace between statistics and public policy decision making" that "has greatly increased the significance and decision value of the statistics we produce." The stress that this "intimate embrace" places on the Federal statistical system has been compellingly described by Wallman (1988).

> Policymakers have called upon the statistical agencies to produce new types of data, for smaller geographic areas, to use in funds distribution. In many cases, the agencies have neither the technical nor the operational resources to meet these mandates, and are left with the unenviable choice: failing to meet the lawmakers' requests, leaving public programs bereft of a good statistical base for fund allocating (and risking congressional displeasure that could threaten support in future years); or reallocating resources to meet new demands, resulting in neglect of already starved core agency programs.

The National Center for Health Statistics has been and continues to be faced with these unenviable choices resulting from politically mandated research. Two such examples already mentioned are the National Household Seroprevalence Survey and the National Survey of Health and Sexual Behavior. As Wallman has questioned relative to the AIDS epidemic: What will happen if survey pretests indicate the data will be of poor quality, while public pressure to understand the epidemic calls for trustworthy national measures?

As another example, the Office of Management and Budget has recently mandated that NCHS population-based surveys collect detailed data on sources and amounts of personal income and health insurance coverage. These data will serve the legitimate purpose of assessing program participation for policy decisions. However, this creates a serious dilemma for NCHS. For example, with the National Health Interview Survey, the detailed income and health insurance questions could take as much as 30 minutes in each household, leaving little time for other topics. This severely limits NCHS in meeting its legislated mandate to provide the data needed by other health agencies. In addition, income is one of the most sensitive areas to address in population-based surveys and the data are always suspect and subject to high levels of imputation.

Furthermore, there are the many legitimate and serious concerns about the Federal "need to know" versus the individual's right to privacy and confidentiality and protection from undue burden. Clearly, the environment in which surveys are conducted has changed. A serious dilemma is faced in maintaining the scientific integrity of data collection systems while responding to political priorities.

National Center for Health Statistics Research Emphases

These are some of the major changes that have occurred over the past decade and some of the new challenges to statistical data systems in general and survey research in particular. What follows is a brief overview of the efforts at NCHS to address some of these issues.

At NCHS, special emphasis is placed on methodological research. A primary function of NCHS, as stated in its authorizing legislation, is to conduct and support such research. For this reason NCHS has developed the philosophy that methodological research is not simply a means to an end but an end in itself. Methodological research is not merely responsive to the data needs; it is assertive in that basic methods research opens data collection opportunities that may not have been anticipated. The National Center for Health Statistics has learned that no single particular survey method or technology can meet all needs and no single one is best under all conditions. Rather a broad repertoire of methods must be developed which can be selectively applied to meet a wide variety of circumstances.

Over the past few years major methodological accomplishments have helped NCHS produce data sets that are more timely, of higher quality, and more relevant to public health and health policy concerns. At the same time, attempts have been made to contain costs and respondent burden.

Sampling

The National Center for Health Statistics has several achievements in the area of sampling. For example, the health care provider surveys were redesigned and integrated into the National Health Care Survey (NHCS). The National Health Care Survey includes both traditional settings of health care and alternative settings such as hospices, home health agencies, free-standing surgical centers, and hospital outpatient clinics and emergency rooms. The four individual surveys that comprise National Health Care Survey are the Hospital Discharge Survey (HDS), the National Ambulatory Care Survey (NACS), the National Nursing Home Survey (NNHS), and the Master Facilities Inventory (MFI).

In addition, the National Health Care Survey is part of the integrated survey design; that is, the samples for these and the population-based surveys at NCHS are being drawn from the same primary sampling units (PSUs) as the sample of the largest survey, the National Health Interview Survey. This process makes the design of the combined sample survey more utilitarian and cost effective than designing each survey independently. With the integrated design, each of the surveys will cover some of the same primary sampling units—and in the case of the population-based surveys, some of the same people—making cross-analyses at the microlevel of health issues possible.

Potentially, there are six major areas improved by the integrated design:

1. Subnational statistics can be produced for smaller geographic areas.
2. Similarly, the capability to produce statistics on population subdomains such as racial and ethnic groups is enhanced.
3. Primary sampling units will overlap in both population and provider-based surveys so that statistics on health care service areas can be produced.
4. Longitudinal and follow back studies can be conducted to monitor the effects of lifestyle on health since individuals can be targeted for tracking over time based on random selection or certain specific characteristics.
5. Analytic reports can be produced that combine statistics from several reporting systems that are collected or maintained by NCHS such as the National Health Care Survey and the National Death Index.
6. Because the composition of the household will be known, individuals—not just addresses—can be targeted.

The National Health Interview Survey is also being redesigned to enhance small area estimates and estimates of other demographic subdomains. Currently, most small area estimates are model based. If a state has the resources to supplement an NCHS survey and wishes to collaborate with us, direct estimation is possible for the state and selected localities. State supplementation enables a survey to be of a larger scale and more tailored to state data needs. One example of this supplementation is occurring now. The State of California needed additional information on health insurance coverage of its residents for proposed legislation. The National Center for Health Statistics assisted in the development of a brief set of questions that will be administered in households in California by the Census Bureau interviewers following the standard NHIS health insurance questions.

A recent experiment using dual-frame sampling resulted in some interesting and unexpected findings. In a joint effort between the Bureau of the Census and NCHS, the possibility of basing the National Health Interview Survey on a dual-frame sample design using the current area sample supplemented with a random digit dialing (RDD) sample was investigated. To explore this possibility, the Census Bureau conducted a national RDD survey using a slightly modified version of the National Health Interview Survey. On the basis of the results, the dual-frame sample approach to the National Health Interview Survey was rejected for three reasons:

1. The RDD costs were higher than anticipated, so the dual-frame system would not have been as cost effective as expected.
2. Nonresponse rates were unacceptably high—response rates were only 80 percent compared to the usual NHIS rate of 95 percent.
3. For the integrated survey design to be feasible, it is imperative to have good address information and for the sample of households to be fairly heavily clustered. With the RDD portion of the sample, this would not be the case.

One use of the dual-frame approach that has proven feasible is for states to supplement the National Health Interview Survey with an RDD survey within their states. Most states have too few NHIS primary sampling units to produce estimates from the National Health Interview Survey alone. The National Center for Health Statistics is working on the development of an optimum dual-frame design for these states based on the National Health Interview Survey and a supplemental RDD telephone survey.

Questionnaire Design Research Laboratory

One of the most progressive and successful endeavors was the establishment in 1986 of the National Laboratory for Collaborative Research in Cognition and Survey Measurement which is jointly funded by NCHS and the National Science Foundation. The mission of the National Laboratory is to promote and advance interdisciplinary research on cognitive aspects of survey methodology among Federal agencies, universities, and research centers. The work has two facets: the Collaborative Research Program and the Questionnaire Design Research Laboratory (QDRL). The Collaborative Research Program funds research through small contracts—mostly with universities—to conduct basic research. The Questionnaire Design Research Laboratory carries out the in-house applied research, that is, it as-

sists in the development and testing of questionnaires that NCHS fields.

The Questionnaire Design Research Laboratory has mainly been used to test Supplements to the National Health Interview Survey and has proven effective in designing questionnaires for a wide range of topics and respondent groups. Current activities also include designing an experimental protocol that presents hypothetical survey situations to individual laboratory subjects, and focuses on the motivations and response strategies that these subjects use in deciding how to respond to sensitive questions in each of the presented situations. Other projects include tests of cued recall, the use of landmarks and time lines, and estimating social pressure bias.

Because the demand for the QDRL's services is becoming greater than can be met, NCHS is requesting additional funding to expand the Laboratory's services to provide design assistance to other components of the Centers for Disease Control (CDC) and Public Health Service (PHS) agencies.

Automation of Data Collection, Access, and Dissemination

Technology has improved survey methods most in the areas of data collection and data processing. The benefits of technology can be seen in the increased timeliness of survey products, the increased scope of analytic output, improved access to data, and improved quality of data.

So that NCHS remains responsive and timely, it has major plans for the use of emerging technology. The overall approach to automation is based on converting existing manual methods for collection and preparation of data in electronic form. In the near future, data will be collected from individuals or institutions in electronic form, for example, through the use of computer-assisted interviewing; edited rapidly and processed through upgraded computer facilities; accessed for analysis and report writing by NCHS analysts through automated software and networked desk-top work stations; and electronically transmitted from work stations to print or electronic release to the public. Similarly, the research community should be able to access NCHS data by compact disk-read only memory (CD-ROM), microcomputer public use files, or other modes that will facilitate their analysis. This assumes that rapid progress will continue in the automation of data collection, processing, and access mechanisms.

Although funding for technology implementation has been constrained, NCHS has made vast strides in computer-assisted interviewing. During the late 1970s and early 1980s, NCHS conducted and sponsored research on computer-assisted telephone interviewing (CATI). Most of the targeted population surveys are conducted using the CATI method. In the past 3 years, NCHS has devoted its limited resources to research and development in computer-assisted personal interviewing (CAPI). Beginning in late 1988, the AIDS Knowledge and Attitudes portion of the National Health Interview Survey was successfully administered by Bureau of the Census interviewers using laptop computers in sample households. This has facilitated fast turnaround and the publication of data on a monthly basis and made available very timely data for evaluation of the impact of the mass mailing of the brochure "Understanding AIDS."

Efforts are underway to develop more generalized CAPI software, including an authoring system so that the more complex surveys like the National Health Interview Survey, which requires extensive rostering, can be fully accommodated. The goal is to implement fully computer-assisted personal interviewing in the near future in conducting the National Health Interview Survey, the Hospital Discharge Survey, and the National Health and Nutrition Examination Survey.

As Nichols (1989) has aptly stated, data collection technologies must be viewed as dynamic rather than static or rarely changing, and "government data collection agencies need to prepare themselves for continued technological change in their data collection operations similar to the more familiar periodic upgrades of their data processing systems". In the near future, NCHS will be exploring the potential of such developing technologies as the use of voice technology and portable, handwritten character recognition (HCR) devices.

Automation will also have a positive impact on data availability, analysis, and dissemination. The National Center is developing for use by its staff on-line data retrieval systems and automated means of quality control and analysis. Data will be accessed for analysis and report writing by NCHS staff through automated software and networked desk-top work stations. This automation will result in data tapes being cleaned and released sooner than ever. This will also release our data users from dependence on mainframe computers.

Besides CAPI and CATI development, NCHS research has led to the development of automated coding systems for mortality data derived from state vital registration systems. These systems, Mortality Medical Indexing Classification Retrieval (MICAR) and Automated Classification of Medical Entities (ACME), are critical to the implementation of the upcoming revision of the International Classification of Diseases, scheduled for 1993. Beyond systems already developed, NCHS plans to improve its capability to develop automated systems for coding occupation and industry from death certifications, and to fund technologies that will allow states to automate more fully the process of obtaining mortality data from hospitals and other sources, as well as electronically transmitting data to NCHS.

New methods for improving access to data are being developed. Consideration is being given to including on floppy disks tables for NCHS hard copy reports, utilizing compact disk-read only memory (CD-ROM), and establishing an on-line access system, such as WONDER. Compact disk-read only memory is a remarkable new data dissemination medium being applied to NCHS data. The storage capacity of a single disk, equivalent to 1,800 floppy disks, has tremendous implications for archiving NCHS datasets and improving accessibility for the microcomputer user. Data files, documentation, tutorials, and programs could be put on a single compact disk for use by in-house and other analysts. Similarly, cross-tabulations could be put on a compact disk to be

accessed by users with limited statistical expertise. Development of a CD-ROM product is underway at NCHS; testing of the first version will be completed this calendar year.

All of this will be made easier with the development of the Automated System for Survey Information and Statistical Tools (ASSIST). As the name implies, this is a two-part system for data retrieval and specialized statistical tools. It is a user friendly tool for both the novice and the professional. It allows the user to choose a dataset from an options screen—for instance, the option selected may be the National Health Interview Survey. The user could then choose the subject of interest, say coronary obstructive pulmonary disease, and the next screen would display the years for which this information is available. There is then a series of screens by which the user may select the precise variables of interest and the statistical measures that are appropriate. The system would then run the program and provide the tables. The user does not have to write any JCL, SAS, PL-1, or FORTRAN routines.

Variance Estimation

Many NCHS surveys use a multistage cluster probability design which poses unique analytic challenges. One of the most complicated problems is in the area of variance estimation. The National Center for Health Statistics has shifted to the more sophisticated Taylor linearization for the computation of variances. To assist the user in applying this complex technique, NCHS has awarded a contract for the development of a software package to make this process more user friendly.

Analysis

The biggest problem encountered in analysis is that many of the analytic methods used today have not been adapted for complex surveys. Analysis of data based on complex samples presents a distinct challenge above the simple random sample assumptions widely used today.

Another analytic demand is the production of statistics on subnational areas. Subnational data are increasingly needed to implement and monitor health programs. Obviously the resources, both financial and statistical, are not always available. The National Center for Health Statistics has been using a procedure known as synthetic estimation to meet the demand for small area statistics. This procedure obtains small area estimates of characteristics by combining national estimates of the characteristics specific to population-subgroups within estimates of the proportional distribution of the local population.

All these methodological developments are aimed at the cross-sectional surveys conducted by NCHS. An epidemiologic approach has been introduced to many ongoing surveys by developing longitudinal or followback components. In addition to repeated contacts with selected respondents over time, survey data can be matched to death records through the National Death Index and to administrative records such as Medicare claims files. Three NCHS surveys currently have follow-up components—the National Health Interview Survey, the National Survey of Family Growth, and the National Maternal and Infant Health Survey. The National Center for Health Statistics is building a longitudinal component into the current wave of the National Health and Nutrition Examination Survey. Undoubtedly, as these longitudinal surveys mature, new methodological issues will require further research endeavors.

Conclusion

There is constant pressure for NCHS to address a plethora of health issues which is placing an enormous burden on resources. Thus, there is a need to develop innovative approaches to meet these demands. Alternatives are needed to the stratified, multistage, cluster surveys of representative samples of the United States. One approach might be a grant system which would fund special approaches to the collection of data. This approach has some important advantages in that it gives researchers with the most knowledge of a specialized area—whether it be a minority group or a geographic area—an opportunity to devise a feasible strategy for measurement of health status. Also, Request for Application and other means could be used to encourage work on specific subgroups that need attention at any given time.

References

Bonnen, J. T. (1983). Federal statistical coordination today: A disaster or a disgrace? *The American Statistician,* 37(3), 184.

Institute of Medicine. (1988). *The future of public health.* (Committee for the Study of the Future of Public Health, Division of Health Care Services). Washington, DC: National Academy Press.

Nichols, W. L., III. (1989). *The impact of high technology on data collection* (p. 49). (CATI research report no. GEN-1. Computer Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee). Washington, DC: Bureau of the Census.

Wallman, K. K. (1988). *Losing count: The federal statistical system* (p. 11). (Population trends and public policy, no. 16.) Washington, DC: Population Reference Bureau.

cer increased dramatically when the coverage policy for mammography was changed. In another medical record study cited by a participant, there was a large increase in the frequency of reports of serious diagnoses within 2 months of a change to a DRG-based reimbursement system. Thus, hospital record data may now be less useful in assessing the accuracy of the reports of household respondents or in supplementing their reports with specific diagnoses than had been true before the implementation of this type of payment system.

Conference participants agreed that current standards for good methodological practice require that researchers have a healthy skepticism about record sources when comparing these data with those from survey respondents.

Several participants noted that many Federal statistical reports from national data collection efforts include extensive technical notes on sources of sampling errors but little discussion of nonsampling errors. Participants agreed that it would be helpful to readers if more documentation on sources of nonsampling errors were provided in Federal reports. It was also recommended that the steps taken in survey planning that are believed to affect data quality be documented. These include the use of interaction coding and information on whether cognitive techniques were employed in questionnaire development.

Matching data from questionnaire items with similar data from record sources is a complex process requiring judgment about what constitutes close agreement or a good match. Many participants believed that matching activities should not be limited to exact matching and that relaxing the matching criteria may be useful in many instances.

Well established survey designs exist to make comparisons between questionnaire data and administrative reports. The appropriate strategy is to have questionnaire and administrative data on all persons in the study population. Sometimes investigators obtain administrative records only on those persons with a positive response to particular questionnaire items. This allows the calculation of the predictive value of a positive response; that is, the likelihood that those with a positive response also will have a corresponding positive value from the administrative record source. However, sensitivity and specificity are usually more relevant objectives of these comparisons between questionnaire data and administrative records, and these statistics cannot be obtained under such circumstances. As an example of the correct design, in a comparison of data on the presence of ulcers from questionnaire data and medical record data, medical record data should be obtained for both those persons with and those without questionnaire evidence of the presence of ulcers.

Considerations for Further Research

Two major areas for future investigation were discussed:

1. Some participants felt that very little is known about how people think about health-related issues: that is, how they encode, organize, and retrieve information. Thus, studies should be undertaken to understand how respondents think so that questions using that framework can be developed. One participant pointed out that definitional and classification issues have an enormous impact on the survey researcher's counts of health events. Many survey researchers generate questions from a classification scheme unknown by respondents who are then asked to classify phenomena into this scheme. For example, questionnaire items on medical conditions are often structured on the International Classification of Diseases and related documents, which have a particular approach to the classification of disease and which were developed for statistical tabulations of death or clinical events. The International Classification of Diseases may have little relationship to the concepts of disease classification used by the general public. It was noted that this definitional framework must then be communicated effectively to respondents to elicit good data.

2. Survey researchers also must be more sensitive to and have a fuller understanding of the limits of a respondent's ability to answer questions. For example, one participant pointed out that health survey instruments continue to include questionnaire items on physician visits occurring over a 1-year period when it is known that the ability of persons to recall these data accurately is poor. Another participant noted that in Perloff's study, information on the business/management status of health care facilities might have been more accurately and appropriately obtained from the medical establishments, as patients should not reasonably be expected to have direct knowledge of this information.

References

Baker, S., and Kronenfeld, J.J. (in process). DRG jump from prospective payment for Medicaid in South Carolina.

Czaja, R. & associates. (1984). Locating patients with rare diseases using network sampling: Frequency and quality of reporting. In *Health survey research methods* (pp. 311-324). Proceedings of the Fourth Conference on Health Survey Research Methods (DHHS Publication No. (PHS) 84-3346). Rockville, MD: Public Health Service.

Young, John and others. (1981, June). *Cancer incidence and mortality in the U. S.* Bethesda, MD: National Cancer Institute.

SESSION 3

Collecting Data from Samples of Older Adults and Nursing Home Populations

Introduction by Floyd J. Fowler, Jr.

Collecting data from and about older Americans is a particularly important focus of health surveys because they are frequent users of medical care services and have many needs for health support that affect their well being. In addition, collecting data from and about older people poses special methodological problems. Two sampling-related issues are particularly important. First, only about one in five households has a person age 65 or older. Hence, to draw a sample of older people in households requires screening a sample of households five times as large as the actual desired sample. On the other hand, the alternative is usually to use imperfect lists which may leave out important segments of the older population.

A second problem for survey research results from the fact that a significant portion of the older population is not in housing units; some are in long-term care nursing homes. Moreover, those in nursing homes are distinctively in need of health services. Survey methodology is primarily geared to sampling people by sampling housing units, and the problems posed by omitting or trying to include nursing home residents are challenging.

Surveying the elderly also poses special challenges with respect to nonresponse. It has been found that older people often decline to be interviewed in telephone surveys that use random digit dialing techniques. In addition, illness and mental deterioration may mean that surveys of older people are distinctively likely to encounter respondents who are unable to answer some or all of the questions posed. Leaving out data for these respondents is a problem because they are different from the rest of the population in important ways, including their need for and use of medical care.

All feature papers in this session address some aspect of the problem of how to collect adequate and complete data about samples of older people. These papers attempt to provide a better methodological perspective on both sampling the elderly and addressing various types of nonresponse in order to reduce survey error.

FEATURE PAPER

Sampling Strategies for Surveys of Older Adults

Dorothy W. Kingery

Introduction

The last two decades have been a period of increasing productivity for researchers in the psychology of aging. Between 1968 and 1979 the number of published studies in the field increased by 270 percent (Poon & Welford, 1980). An article examining research between 1975 and 1982 suggests that the increase in production has been accompanied by improvements in research quality (Hoyer & associates, 1984). Authors of that research summary note that although the field is becoming more sophisticated, methodologically there continue to be some problems, especially in the area of sampling. One criticism of the published research is that sampling methods vary widely and often are not discussed in enough detail to allow the reader to critique the method employed. An associated criticism is the lack of information on what sampling problems were encountered and how they were resolved.

Construction of a good sampling frame is a major practical, as well as methodological, problem (Cochran, 1977). As older adults become more popular as a target group for research, attempts to solve the problems involved in sampling and collecting data from this age group become even more important. Despite an increase in the proportion of our population that is over 65, older adults presently account for only 11.3 percent of the total U. S. population (1980 U. S. Census). Thus they are usually difficult and expensive to reach using methods of sampling common to general population telephone and mail surveys. As a result, studies often use samples of convenience or attempt to recruit specific types of volunteers (Bolla-Wilson & Bleecker, 1989; Holahan & Holahan, 1987; Windle & Sinnott, 1985).

In an effort to adapt probability sample designs to studies of older adults the Survey Research Center (SRC) at the University of Georgia conducted studies using two different types of sampling strategies. These studies were designed to recruit participants using procedures that would be efficient with respect to time and budget and produce a sample representative of the population. Methodological data from those studies were used to examine the efficiency of each sampling strategy, the comparability of samples to the population and problems associated with each technique. The first strategy used Random Digit Dial (RDD) procedures. The second collected a targeted age sample based on listed data. Sampling methodology, problems associated with each method and the implications of using each strategy are summarized below.

Classic Random Digit Dial General Population Survey

If research results have to be projectable to the universe then a probability sample design must be used (Kerlinger, 1973). For studies using telephone interviewing, RDD samples are the design of choice because they provide all (one) telephone households an equal probability of being included in the sample. However, this option is not well suited to samples that are difficult to target. Thus, it is a relatively inefficient mechanism for producing samples of subgroups such as older adults.

One example will demonstrate both the coverage and the inefficiency of RDD sampling for older individuals. Survey Research Center uses RDD samples to collect data for semiannual omnibus surveys of adult residents of the State of Georgia. Sample size ranges from 500 to 550 with an average interview length of 20 minutes. Data from these surveys demonstrate that it is time consuming to screen a general population sample to acquire a

Dorothy Kingery is with the Survey Research Center, University of Georgia, Athens.

The author would like to thank Dr. Leonard Poon, Director, Gerontology Center, and Dr. Matt Perri, College of Pharmacy, University of Georgia, for permission to use data from research supported by Grant No. LCR-21R01MH43435-01 from the National Institute of Mental Health and Grant RFAARPFDN from the Andrus Foundation.

target population of older respondents. Approximately 500 hours of calling yields 70 to 80 respondents age 66 and over. However, these omnibus surveys consistently pick up a representative number of adults in the over 65 age group. According to 1980 U. S. Census figures, 13.5 percent of Georgia's adult population is over the age of 65, 1.7 percent are 80 plus. Over the last 4 years an average of 14.4 percent of omnibus poll respondents has been over 65, 1.9 percent have been 80 plus. Thus, the efficiency of screening general population samples to target older Georgians is low, but the samples that result are representative in terms of the proportion of elderly found in the larger population.

Although survey efficiency is important in selecting a sampling method, if necessary, an RDD sample, even of difficult-to-reach groups, is a reasonable choice. This was the case with two surveys conducted in the past year. Surveys sampled adults 50 and older in two contiguous counties in Georgia. The counties consist of rural and urban areas and have a combined population of approximately 87,000. The study used a pretest-posttest design to determine the effectiveness of media advertising over knowledge and use of generic prescription medications. In addition to the pretest and posttest surveys, a comparison group of randomly selected adults 18 and older from the same geographic area was sampled during the time of the posttest. For all three surveys, telephone numbers were randomly generated using two-stage RDD procedures as developed by Waksberg (1978).

Methodological data from the three surveys were used to compare data collection efficiency and response rates for the two age groups. Figure 1 shows that of the telephone numbers called, 16 percent (pretest) and 14 percent (posttest) yielded an eligible respondent in the 50 plus sample versus 23 percent for the general adult sample. Thus, it took almost twice as long to contact an eligible respondent in the 50 plus sample as for the sample using all adults 18 and older. Response rates were lower for the two older samples, 66 percent and 70 percent versus 81 percent for the 18 plus sample (Table 1 and Figure 2). Using the Troldahl-Carter respondent selection method (1964), males and females were appropriately represented in both samples.

Table 1. RDD general population survey households contacted

	County samples		
	Age 50 and over		Age 18 and over
	October 1988 pretest	December 1988 posttest	December 1988 comparison group
Eligible household	16%	14%	23%
Response rate	66%	70%	81%
Sample size	312	392	225
Numbers used	3,057	3,993	1,222

Figure 1. Household contact RDD method two-county sample

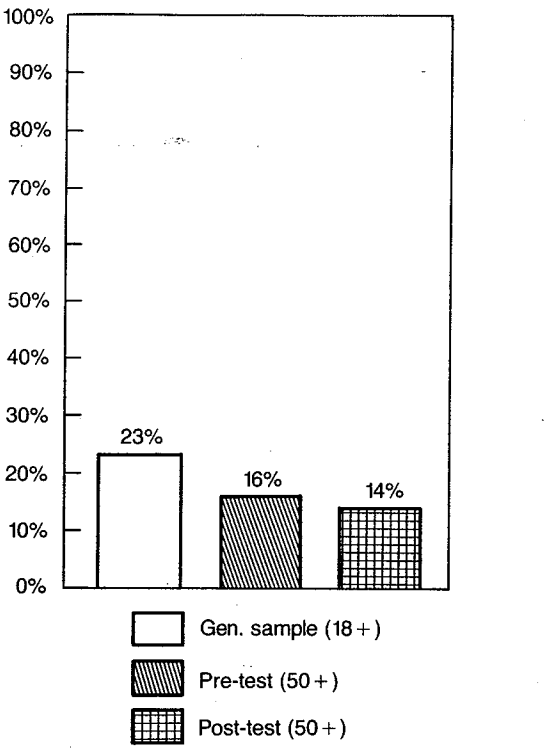
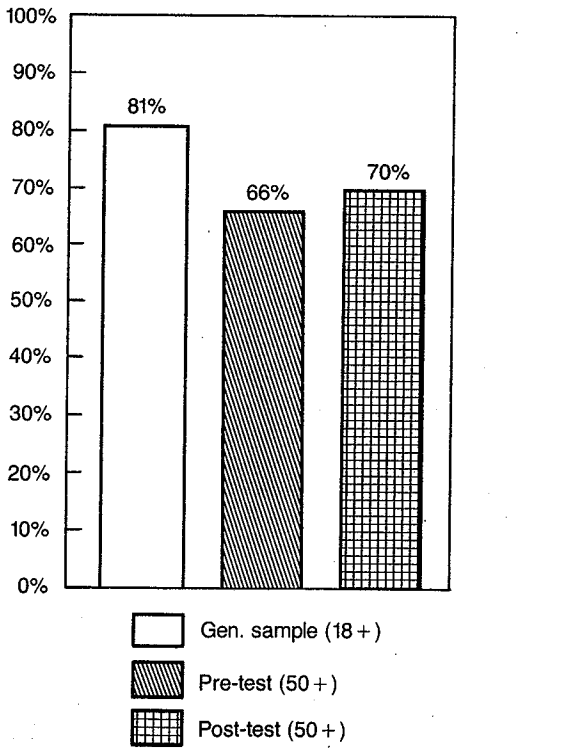


Figure 2. Response rate RDD method two-county sample



Targeted Age Sample Based on Listed Data

A targeted age sample is one option for increasing the efficiency of RDD samples. These samples use as broad a base as possible while targeting age groups to increase efficiency. Targeted samples first select specific areas of a city or county from which a random sample of telephone numbers is then generated. The increase in efficiency is gained at the expense of a decrease in statistical projectability because not all households in the universe have an opportunity to be selected. This type of RDD sample does result in known probability of selection thus relative "weights" can be computed (Scheaffer & associates, 1979).

Efficiency and dispersion of target RDD samples will, of course, vary with the geographic area covered by a study. In the metropolitan Atlanta area, 8.8 percent of the population is age 65 or older. To increase efficiency in sampling this age group to 12 percent decreases coverage from 225 telephone exchanges to 68. An increase in efficiency to 15 percent decreases coverage to 38 exchanges. In addition to the decrease in dispersion of subjects, this method does not allow us to target specific age groups in the over 65 category. Thus, it was not an appropriate method to use for a study investigating physical and mental health and health practices of persons age 60 to 69 and 80 to 89. Rather, a targeted age sample based on county lists of registered voters was used to recruit 20 participants in each age group. The list sample increases efficiency and has the advantage of allowing a random selection of eligible respondents from a known population. As with other types of targeted samples, there is a loss in statistical projectability because not all eligible respondents have the opportunity to be selected (Singh & Chaudhary, 1986). Probabilities vary according to the type of list used.

Voter registration lists offer broad coverage while allowing a researcher to examine potential demographic biases evident in the resulting sample. Although counties vary in the amount of information they are willing to make available to researchers, in all cases available lists contain name, mailing address, and birth date for currently registered voters.

The health study used cluster samples drawn from the following geographic areas of Georgia: metropolitan Atlanta, two metropolitan statistical areas (MSA) that are smaller than Atlanta (Augusta and Athens) and one large rural county in the southern part of the state. The Board of Elections in each county chosen as a part of the sample furnished a list of current voter registrants age 60 to 69 and 80 to 89. Data presented are from the Athens metropolitan area, the first sample collected. Names were randomly selected from the voter list for the two age groups of interest. Each respondent selected was telephoned and asked to participate in the study by being interviewed in their homes on two to three occasions for a total of approximately 5 hours. Participants were paid $50 and given a free physical examination. To the extent possible, demographic data were collected on respondents who refused to participate in the study.

A comparison of demographic characteristics of registered voters and the county's adult population demonstrates a similar composition for race and sex (Table 2 and Figure 3). Note, however, that there are 3.5 percent fewer blacks on the voter registration list than in the county's population. By age, the voter list contains a larger percent of persons age 60 to 69 and 80 to 89 than does the county population. The large number in the 60 to 69 age group may result in part from the high educational level of county residents. The state's flagship university, located in the county sampled, has a large number of older, tenured faculty, and the community has a large number of university retirees. The positive association between level of education and voting behavior may account for this particular finding. The disproportionately large percent of registered voters in the

Figure 3. Voter list and population figures

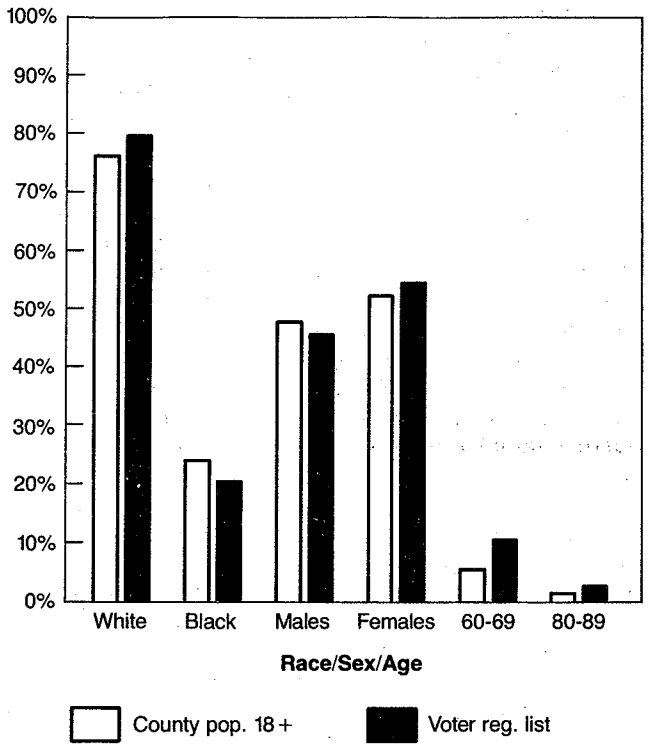


Table 2. County population and voter registration list demographics

	Percent of county population age 18 and over	Percent of voter registration list
Race		
White	76.1	79.6
Black	23.9	20.4
Sex		
Males	47.7	45.5
Females	52.3	54.5
Age		
60-69	5.7	10.7
80-89	1.6	2.9

60 to 69 age group was not expected. A question on voter registration on SRCs 1988 statewide omnibus poll shows that a similar proportion of respondents in the 55 to 64 age group (84.8 percent) and in the 65 and older group (83.2 percent) report being registered to vote. Despite the inconsistency in actual percentages, these findings indicate that neither State nor county voter registration lists are biased against older adults.

How the list sample reported on here compares to the specific population of interest is not known. The study's sample frame was individuals who are cognitively and physically intact and living at least semi-independently. The data are useful in examining reasons for nonparticipation and for dropout between selection as a respondent and the actual interview. Of major interest, also, are data on survey efficiency. In 36 hours of telephone calling, interviewers recruited 37 participants in the 60 to 69 age group, 43 in the older group; 7 black participants were recruited.

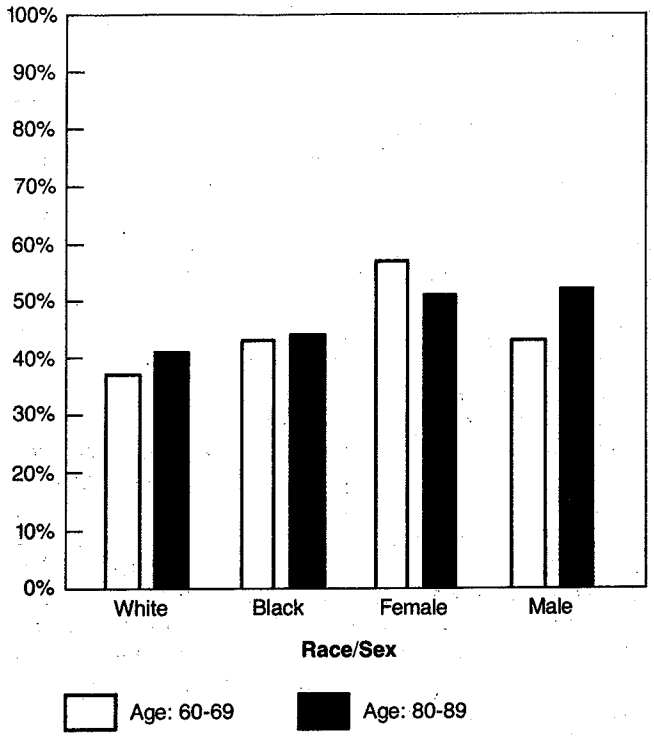
Table 3 suggests one bias attached to the use of a voter registration list coupled with an initial telephone contact. This is evident in the large number of individuals in the 80 to 89 age group for whom we were unable to locate a telephone number. It is reasonable to assume that some of these people live in a household with a telephone listed in another name or in a nursing or retirement home and do not have their own telephone. Also, some potential older respondents will be deceased. Of interest, too, is the sharp increase in the percent of potential respondents in the older group who are unable to participate because of illness.

Table 3. Voter registration list survey efficiency

	Age groups	
	Age 60-69	Age 80-89
Reasons for nonparticipation		
No phone number	38	104
Nonworking number	1	6
Incorrect age	1	8
Out of town	0	1
Ill or in hospital	6	26
Refused	41	45
Refusal rate	37%	32%
Unresolved		
Callbacks	38	55
Will participate (recruited)		
White female	15	23
White male	14	15
Black female	5	4
Black male	3	1
	37	43
Total	162	288

Refused by Age/Sex/Race Percent of category who refused	Age 60-69	Age 80-89
White	37%	41%
Black	43%	44%
Females	57%	51%
Males	43%	52%

Figure 4. Refusal rates by age, sex, and race



Overall, the response rate for eligible respondents was similar for the two age groups. By category there are some differences. In both age groups, slightly more blacks than whites refused to participate in the study. The difference was greater in the younger sample (Figure 4). The figures also demonstrate that a much larger percent of women than men age 60 to 69 refused to participate.

A large number of the participants recruited dropped out before their initial interview. As a result it was necessary to recruit 80 participants to fill 40 slots. Our data show that the second-stage participation rate was uneven. Interviewers were unable to reach some respondents who had agreed to participate. Two respondents are still willing to interview but cannot schedule an appointment. The percent of each category, age by sex by race, who have actually participated is shown on Figure 5. There was a high participation rate for black women and a low rate for black men. The low percent of black males interviewed results in part from the small number recruited.

Implications

A classic RDD method is appropriate for collecting a sample of older adults when a telephone interview is used for data collection, as in our pretest and posttest surveys. Also, if the survey targets a group containing the "younger old," RDD methods are quite acceptable. This method takes longer to contact eligible respondents but fulfills a major requirement of sampling theory—that the sampling frame is comprehensive (Fowler,

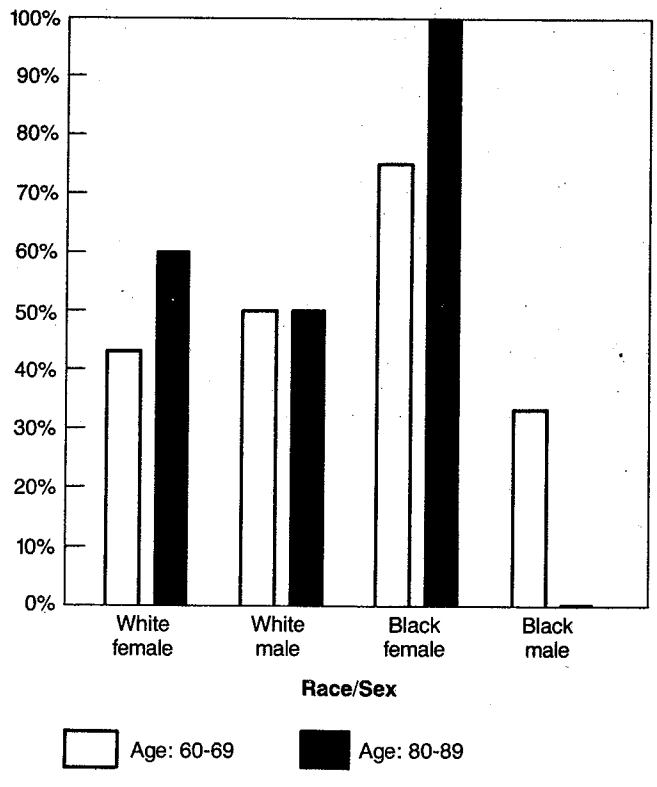
1988). A lower overall response rate for the over 50 sample than for the general population sample may signal a potential bias, in that respondents who agreed to be interviewed may be different from the population.

The voter registration list provided a relatively efficient method for reaching respondents. However, although telephone screening is effective, the attrition rate from the time a respondent agrees to participate and the time of the interview is a major concern. More attention needs to be given to retaining respondents originally recruited. The strategy at the Survey Research Center has been revised so that respondents are given an appointment time when they initially agree to participate, and elapsed time has been decreased so it will be no more than 2 to 3 days. A packet of information plus an incentive is being prepared to be mailed to respondents immediately after their recruitment.

Although data on education have not been processed, there is concern with the potential for the voter registration list to be biased against less educated respondents and blacks. Neither the list nor the sample originally recruited demonstrated an obvious bias against blacks, but the list did contain 3.5 percent fewer blacks than the population. This plus a higher dropout rate before the interview resulted in a sample skewed toward whites. In the 60 to 69 age group, lag time before the interview was conducted led to a loss of some of the more active respondents who had travel or other plans. In the older group, illness of either the selected respondent or a spouse was a major problem.

A target RDD sample or a probability sample of census tracts stratified on the basis of age are reasonable alternatives to consider, especially for large, metropolitan areas. They cluster respondents, thereby making contact easier for interviewers. The latter suggestion would eliminate the drop-out rate between contact and completed interview, but is more expensive than other methods discussed. Also, use of a targeted age RDD-sample will not always result in the dramatic decrease in coverage demonstrated by the figures for the metropolitan Atlanta area. Researchers need to examine potential coverage for the areas of interest to them.

Figure 5. Recruited and participated voter list



Overall, the effort to use probability designs to sample older adults was successful. Although the RDD surveys performed well, recruitment of study participants using the voter registration list has obvious problems. The new procedures adopted are expected to eliminate some of the second-stage dropout. Data presently being collected in a rural Georgia county for another list-based sample should furnish additional information about voter registration list effectiveness and about potential differences in using the list sample in rural and in urban areas.

References

Baron, A., & Menich, S. R. (1985). Age-related effects of temporal contingencies on response speed and memory. *Journal of Gerontology,* 40, 60-70.

Bolla-Wilson, K., & Bleecker, M. L. (1989). Absence of depression in elderly adults. *Journal of Gerontology,* 44, 53-55.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Fowler, F. J., Jr. (1988). *Survey research methods* (rev. ed.). Beverly Hills, CA: Sage Publications.

Holahan, C. K., & Holahan, C.J. (1987). Self-efficacy, social support and depression in aging: A longitudinal analysis. *Journal of Gerontology,* 42, 65-68.

Hoyer, W. J., Raskind, C. L., & Abrahams, J. P. (1984). Research practice in the psychology of aging: A survey of research published in the *Journal of Gerontology,* 1975-1982. *Journal of Gerontology,* 39, 44-48.

Kerlinger, F. N. (1973). *Foundations of behavioral research.* (2nd ed.). New York: Holt, Rinehart & Winston, Inc.

Poon, L. W., & Welford, A. T. (1980). Prologue; A historical perspective. In L. W. Poon (Ed.). *Aging in the 1980s: Psychological issues.* Washington, DC: American Psychological Association.

Scheaffer, R. L., Mendenhall, W., & Ott, L. 1979. *Elementary survey sampling* (2d ed.). Boston, MA: Duxbury Press.

Singh, D., & Chaudhary, F. S. (1986). *Theory and analysis of sample survey designs.* New York: John Wiley & Sons.

Troldahl, V. C., & Carter, R. R., Jr. (1964). Random selecting of respondents within households in telephone surveys. *Journal of Marketing Research,* 1, 71-76.

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

Windle, M., & Sinnott, J.D. (1985). A psychometric study of the Bem Adult Sample. *Journal of Gerontology,* 40, 336-343.

FEATURE PAPER

Collecting Health Data From and About Older People: The Longitudinal Study of Aging

Mary Grace Kovar

Survey Sample and Procedures

The National Center for Health Statistics (NCHS) has long been known for conducting large cross-sectional health surveys. Analysts knew that longitudinal studies were needed to study cause and effect and to learn about change in individuals, but until very recently the Center had never conducted one. The study of older people described in this paper is the first survey by the Center designed to be longitudinal.

The design of the Longitudinal Study of Aging, whose purpose is to measure transitions in functional ability and living arrangements, is shown in Figure 1. It takes advantage of the National Health Interview Survey (NHIS), the large continuing survey of the National Center for Health Statistics that has been used since 1957 to collect information about the health of people living outside of institutions in the United States. The National Health Interview Survey is designed to have a basic questionnaire and supplements. In 1984 there were two supplements. The Health Insurance Supplement was similar to earlier supplements; The Supplement on Aging (SOA) was new.

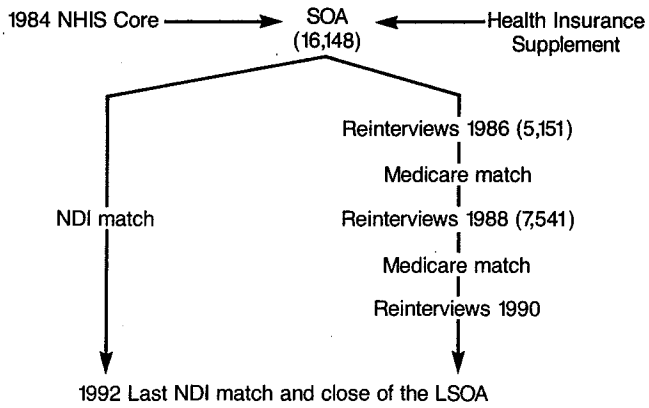
The Supplement on Aging had two major purposes. First, it was designed to obtain information about people age 55 and older and to obtain as much of that information as possible from self-respondents. Second, it was designed to be the baseline survey for a longitudinal study (Fitti & Kovar, 1987).

In preparation for a longitudinal study, participants in the Supplement on Aging (or their proxies) were asked for information, such as a social security or railroad retirement number, that was needed to link with other files. They were also asked for the name, address, and telephone number of someone who did not live in the same household and who would be likely to know where they were in case we wanted to contact them again. Of the 16,148 participants in the Supplement on Aging, almost 16,000 provided linkage information and virtually all provided information on a contact person.

Because reinterviews were to be conducted by telephone and some people doubted that older people would respond to a telephone survey, a small feasibility study was conducted. The results showed that response rates for such a study would be good provided certain procedures were followed (Kovar & Fitti, 1985).

Funds were provided by the National Institute on Aging to follow a sample of the oldest old, defined for this purpose as people who had had their 70th birthday before they were interviewed in 1984, through telephone interviews every 2 years. Because funds were limited for the 1986 interview, a sample was selected from the participants in the Supplement on Aging. All persons age 80 and over, all black persons age 70 to 79, one half of the remaining sample age 70 to 79, and all other persons age 70 to 79 who lived in the same household as one of the above were selected for the 1986 interview. The 1986

Figure 1. The Longitudinal Study of Aging



SOURCE: NCHS, Longitudinal Study of Aging, 1986

Mary Grace Kovar is with the National Center for Health Statistics, Hyattsville, Maryland

sample included all of those who had been 80 and over, and 56 percent of those 70 to 79 years. For the 1988 interview, all participants age 70 and over were included.

All interviewing in both 1986 and 1988 was conducted using Computer Assisted Telephone Interviewing (CATI) by U.S. Bureau of the Census interviewers at the Census facility in Hagerstown, Maryland. Before interviewing began, all potential respondents were sent a letter, addressed to them by name, from the Director of NCHS. The letter told them that participation was voluntary, gave them the authorizing legislation and a telephone number to call if they had questions, and told them the areas that would be covered in the interview. Interviewing began in August each time and continued through September in 1986 and through October in 1988. Persons who had not given a telephone number and those who could not be contacted by telephone were sent a mail questionnaire with one followup.

Because we had originally intended only to ascertain the fact of death, and decided to obtain additional information about decedents after the study was designed, that information was collected after the 1986 Computer Assisted Telephone Interviewing had been completed. The same telephone interviewers and a paper and pencil questionnaire were used.

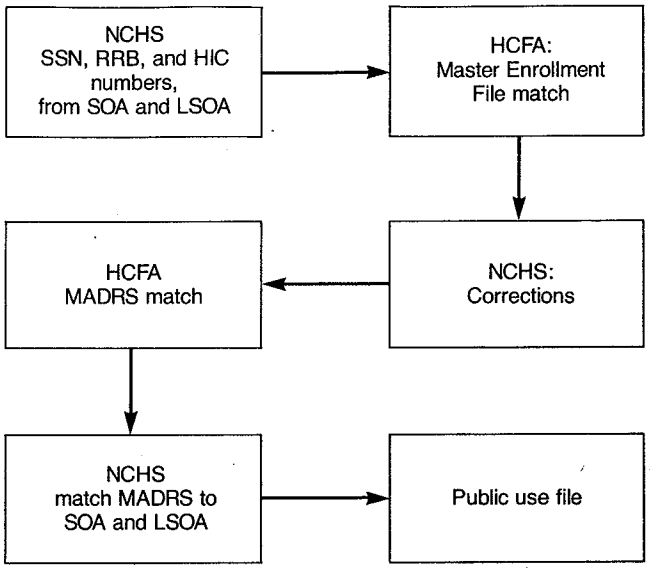
Participants' records are matched with the National Death Index (NDI) each year. If the participant is known to be dead from the NDI match, that information was on the CATI system. Otherwise, the interviewer has to ask. The date of death from the NDI match is on the public-use file. The states have now given permission to match with the multiple cause-of-death computerized records. That information will be added to files as soon as the procedures are approved.

Participants' records are also matched against Medicare records maintained by the Health Care Financing Administration (HCFA). To preserve confidentiality, only the Social Security or Railroad Retirement number is sent to HCFA to be matched against the Master Enrollment file (Figure 2). Name, address, sex, date of birth, and the Health Insurance Claim number is returned for the matches. That information is matched against the Longitudinal Study of Aging file and all records with a discrepancy are reviewed. In almost all cases the difference was in the spelling of a name or in an address. In those cases, the HCFA information is accepted as correct and it is used to correct the Longitudinal Study of Aging files for future interviews and matches. In rare cases the match is obviously the wrong person. In those cases, the number is deleted from the Longitudinal Study of Aging files. When the only difference is in the date of birth or in sex, the information is retained on the matched file.

After the corrections have been made, the file of Social Security, Railroad Retirement, and Health Insurance Claims numbers is sent back to HCFA to be matched against the Medicare Part A and Part B claims files. Selected information from those files is added to the public-use Longitudinal Study of Aging file.

Response rates to the Supplement on Aging were high. The household response rate for the National Health Interview Survey was 95 percent. Information for

Figure 2. LSOA Medicare match



SOURCE: NCHS, Longitudinal Study of Aging

the Supplement on Aging was obtained for 96 percent of the people age 55 and older in the interviewed households (97 percent of the people age 65 and older), giving an effective response rate of 92 percent.

The self-response rate for the Supplement on Aging was 92 percent. People ages 65 to 74 were somewhat more likely than those 55 to 64 or 75 to 84 to be self-respondents (94 vis 92 and 91 percent) and much more likely than those 85 and older (73 percent).

The response rate in 1986 was 92 percent, including 12 percent who were reported to be deceased (Table 1). The response rate for the followup on the 604 reported decedents was 91 percent. Tracing conducted after the 1986 interviews had been completed enabled us to locate about half of the people not located and add corrected information for the 1988 interviews.

Poststratification weights were calculated for the Supplement on Aging using the same age-race-sex categories as used in the National Health Interview Survey so that estimates from the Supplement on Aging for those

Table 1. Response to the 1986 LSOA reinterview

	Supplement on Aging	
	Number	Percent
Total	5,151	100.0
Status known	4,734	91.9
Alive	4,130	80.2
Live alone	1,597	31.0
Live with others	2,323	45.1
Live in institution	192	3.7
Not reported	18	0.3
Deceased	604	11.7
Status unknown	417	8.1

NOTE: Some with status unknown will be located through NDI matches
SOURCE: NCHS, Longitudinal Study of Aging, 1986

groups would be identical to those from the basic survey. New weights were calculated for the 1986 sample using the same poststratification cells. The weights for the 1988 sample were the same as those for the SOA sample.

What Has Been Learned

It is possible to obtain high response rates when collecting information about older people by telephone. The advance letter addressed to the potential respondent by name with the indication of the questions to be asked helps keep the rates high. Using Public Health Service letterhead for the advance letter and having Census Bureau interviewers may also help.

A survey of the oldest-old imposes a psychological drain on the interviewers. Many older people are in good health and are coping well. However, some older people are not, and the interviewer has to try to obtain information from people who are in bad health or without resources, or from their relatives who are trying to care for them. Older respondents cannot be rushed; thus, interviews take longer than with younger people. Every interviewer who worked on this survey broke down at least once after completing such an interview. At the same time, those interviewers feel very possessive about this survey and want to be the ones to contact "their people" again.

Evidence is accumulating that the oldest-old do not respond well to the structured questionnaires used on the National Health Interview Survey. They have a way of telling about their lives and, in response to repeated questions, will tell the same story again. This was observed while listening to the CATI interviews, and some work in the NCHS Cognitive Research Laboratory confirms it (Jobe & Mingay, unpublished).

The best time to interview older people by telephone is before noon on weekdays. When the Longitudinal Study of Aging began, interviewing started at 9:00 a.m. local time, but it was soon learned that older people preferred being called early and interviewing could begin at 8:00 a.m. Sunday is a bad time to call because that is family visiting time, and evenings are bad because older people tend to be tired. These findings are in contrast with those for the general adult population (Weeks & associates, 1987).

A household respondent will protect or prevent an older person from being self-respondent. The NHIS respondent rules call for all adults who are at home to join in the interview and respond for themselves. Those rules clearly are not followed for older people. The self-response rate for the Supplement on Aging, which was usually conducted immediately after the basic National Health Interview Survey, was 92 percent in contrast with 82 percent answering some questions and 76 percent answering all questions for themselves on the basic National Health Interview Survey (Fitti & Kovar, 1987).

Moreover, women protect men from being questioned, especially over the telephone. In spousal households, men age 70 and over were 1.5 times as likely as women to have a proxy respondent in 1984 when the interview was in the household; they were 2.1 times as likely as women to have a proxy respondent in 1986 when the interview was over the telephone (26.6 vis 12.8 percent).

Table 2. Characteristics of persons in the sample who were self-respondents or had proxy responses

	Total	Self	Proxy
Total 65 plus	100.0	100.0	100.0
	Percent distribution		
Age			
65-74 years	61.7	63.3	45.1
75-84 years	31.1	31.0	32.9
85 years and older	7.2	5.8	22.0
Number of ADL with difficulty			
None	77.3	79.8	50.7
One	9.3	9.1	11.3
Two	4.7	4.3	8.4
Three	2.9	2.6	5.5
Four or more	5.9	4.2	24.0
Number of ADL with help			
None	90.3	93.1	61.5
One	3.9	3.4	8.9
Two	1.9	1.4	6.8
Three or more	3.9	2.1	22.7

SOURCE: NCHS, Supplement on Aging, 1984

Among older people, self-respondents are not necessarily better than proxy respondents, when better is defined as more. The amount of disability reported by proxy respondents to the Supplement on Aging for people age 65 and older was higher than that reported by self-respondents (Table 2).

Although some people have argued that caretakers tend to overestimate the level of disability and that all participants in studies of aging should be self-respondents, restricting a study to self-respondents seriously biases estimates. For example, if information had been accepted only from self-respondents, death rates would be underestimated (Table 3). Further, the number of people with difficulty in Activities of Daily Living would have been badly underestimated, estimates that are being used by the Congress to design long-term-care legislation (Table 4). That does not mean that chrono-

Table 3. Status in 1986 of persons age 70 and over in 1984 by respondent status in 1984

Respondent status in 1984	Total number	Status in 1986			
		Total	Living	Dead	Not located
		Percent distribution			
Total	5,151	100.0	80.2	11.7	8.1
Self-respondents	4,578	100.0	82.4	9.5	8.0
Proxy in household	411	100.0	62.8	29.2	8.0
Proxy not in household	83	100.0	59.0	28.9	12.0
Not recorded	79	100.0	62.0	29.1	8.9

SOURCE: NCHS, Longitudinal Study of Aging, 1986

Table 4. Population estimates based on all persons and on self-respondents only

	Total	Self-respondents
Total 65 plus	26,433	26,433
Age		
65-74 years	16,288	16,708
75-84 years	8,249	8,208
85 years and older	1,897	1,517
Number of ADL with difficulty		
None	20,434	21,108
One	2,440	2,393
Two	1,243	1,141
Three or more	2,315	1,791
Number of ADL with help		
None	23,884	24,608
One	1,028	908
Two or more	1,521	918

SOURCE: NCHS, Supplement on Aging, 1984

logical age should be used as the sole criteria for deciding on self-response or proxy response. Information from an older self- respondent may be more accurate than that from a proxy. A case in point is Medicare coverage, which we can evaluate from the link with Medicare data files.

People who were reported to have Medicare coverage were more likely than those who did not to provide a Medicare number and more likely to match with Medicare files (Table 5). However, 18 percent of the records of those who were reported to have coverage could not be matched, and the records of 44 percent of those who did not have coverage were matched.

Self-respondents were more likely than proxy respondents to report coverage, to give a number if covered, and to match if a number was given (Table 6). They were also more likely to give a number and to match if they said that they were not covered by Medicare. Salient use may be a factor in whether proxy respondents can report Medicare coverage. Some of the difference in the match rate for self-respondents and proxy respondents may be due to proxies failing to report a number correctly if the index person has never had hospitalization paid for by Medicare (Table 7).

For 6 percent of the people with interviews in both 1984 and 1986, Medicare coverage reported in 1984 differed from what was reported in 1986 (Table 8). A self-respondent in both 1984 and 1986 was more likely to report the same Medicare coverage than a self-respondent proxy or a proxy respondent. People for whom different reports of coverage were given were less likely to have a number for the match and, if a number was given, were less likely to match.

Summary

Using the National Health Interview Survey as the base for a longitudinal study has several advantages. The costs of screening are avoided and, if the study is designed as this one was, all of the baseline data are collected through the household interview. However, there is a disadvantage if the population of interest is relatively small. Because the NHIS sample has almost equal probability of selection, there will be few participants who are in such populations. In this case, there were fewer people age 80 and over and far fewer age 85 and older than we would have wished.

Table 5. Comparison of reported Medicare coverage in 1984 with match to Master Enrollment file

Match status 1984-1987	Total	Coverage in 1984	
		Yes	No
	Number in sample		
Total	5,151	4,983	
	Percent distribution		
Total	100.0	100.0	100.0
No number given	10.8	10.3	23.8
Number given	89.2	89.7	76.2
No match	18.2	17.8	32.1
Match	71.0	71.9	

NOTE: No includes 8 people with coverage not reported.
SOURCE: NCHS, Longitudinal Study of Aging, 1986

The match rates with the Health Care Financing Administration (HCFA) files were disappointing. Of the 4,600 persons in the 1986 sample for whom we had a number, 20 percent (939) did not match the HCFA Master Enrollment File. We asked for the numbers again in 1988 for everyone for whom a number had been given in an earlier interview but there was no match. People conducting a one-time study do not have that luxury and should exercise extreme care to make certain that the number is reported and recorded correctly.

Production schedules should be modified to take into account the longer time needed to interview very old people. They should also be modified to allow time for interviewers to recover from the stress of interviewing people who are sick and in need of help.

Table 6. Comparison of Medicare coverage reported in 1984 and match with 1984-1987 Master Enrollment file

Coverage reported in 1984 interview	Total	Respondent in 1984	
		Self	Proxy
	Number in sample		
	5,072	4,578	494
	Percent of sample		
With Medicare coverage	96.8	97.1	94.5
If covered, gave number	89.8	91.2	76.7
If number given, matched	80.1	81.1	69.8
Without Medicare coverage	3.2	2.9	5.5
If not covered, gave number	77.8	80.7	63.0
If number given, matched	57.1	58.7	47.1

NOTE: Excludes people with either respondent or coverage not recorded
SOURCE: NCHS, Longitudinal Study of Aging, 1986

Table 7. Match status of people who were reported as covered by Medicare in 1984

Match status 1984-1987	Total	Respondent in 1984	
		Self	Proxy
	Number in sample		
Total	4,910	4,443	467
	Percent distribution		
Total	100.0	100.0	100.0
No number given	10.2	8.8	23.3
Number, no match	17.8	17.3	23.1
Match	72.0	73.9	53.5
No hospital use	33.1	34.7	17.8
Hospital use	38.9	39.2	35.8

NOTE: Excludes people with respondent not recorded
SOURCE: NCHS, Longitudinal Study of Aging, 1986

Table 8. Medicare coverage and respondent as reported in 1984 and in 1986

Coverage in 1984 and 1986	Total	Respondent in 1984 and 1986		
		Self/ Self	Self/ Proxy	Proxy/ Proxy
	Number in sample			
	4,113	3,001	860	252
	Percent distribution			
Total	100.0	100.0	100.0	100.0
No change	93.4	95.4	91.3	90.0
Yes/yes	93.1	94.5	89.9	88.1
No/no	1.2	1.0	1.4	2.8
Change	5.7	4.6	8.7	9.1
Yes/no	4.0	3.1	6.3	7.5
No/yes	1.7	1.5	2.4	1.6

NOTE: Excludes people with respondent or coverage not recorded and people with no 1986 interview
SOURCE: NCHS, Longitudinal Study of Aging, 1986

The high rates of successful contacts on weekday mornings suggest that a survey organization could conduct interviews with the general population and with the oldest population simultaneously because the best time to interview one differs from the best times for the other.

Data from the Longitudinal Study of Aging suggest that selecting the respondent for older people deserves more attention than it has been given. Many older people know more about their affairs than a proxy does. However, if people are extremely ill or have lost much of their cognitive functioning, a proxy can provide information better than the index person. In any case, *all* older people must be included in any study of aging, or estimates of disability and death will be biased.

References

Fitti, J. E., & Kovar, M. G. (1987). The supplement on aging to the National Health Interview Survey. *Vital and Health Statistics* (Series 1, No. 21. DHHS Pub. No. (PHS) 87-1323). Washington, DC: U. S. Government Printing Office.

Kovar, M. G., & Fitti, J. E. (1985). A linked follow-up study of older people. *Proceedings of the Survey Research Section of the American Statistical Association*. Alexandria, VA: American Statistical Association.

Weeks, M. F., Kulka, R. A., & Pierson, S. A. (1987). Optimal call scheduling for a telephone survey. *Public Opinion Quarterly,* 4, 540-549.

FEATURE PAPER

The Effects of Nonresponse and Attrition on Samples of Elderly People

Cynthia Thomas

Introduction

Nonresponse presents a potential threat to validity in any survey. Participation in a long-term study can be more problematic for older than for younger people because poor health may make it difficult to provide information on a continuing basis. In panel surveys, attrition as well as nonresponse complicate the interpretation of the data. Attrition is potentially a more serious problem in panel studies of elderly people because they are more likely than younger people to experience a serious, disabling illness or to die.

This paper examines the impact of nonresponse and attrition in a panel of 1,855 elderly people in the Bronx who were first interviewed in 1984, and every 6 months thereafter for 3.5 years, for a study of health, health care, and aging. Seventy-three percent of the sample, selected from a list of Medicare enrollees, agreed to participate and completed a baseline or wave 1 interview. On average, 2 percent of remaining respondents withdrew from the study on any given wave or, more rarely, could not be found, and 3 to 4 percent of the original sample died between any two waves. The response rate and attrition patterns are considered to be quite respectable for a study in a large metropolitan area such as New York City among a sample with a mean age of 75.

Response rates in surveys have been found to decline with age (Herzog & Rodgers, 1988), and nonrespondents generally are in poorer health, have less education, and are in other ways more disadvantaged than respondents. Little is known, however, about the impact of nonrespondents' missing data on analytic results, although it has been suggested that the absence of such data can decrease the relationship between risk factors and a disease outcome or distort the relationship between sociopsychological factors and death (Criqui & associates, 1979; Riegel & associates, 1967).

Similarly, elderly panel-study dropouts have been found to be sicker early in the study, both mentally and physically, and to have other traits generally more like those of the disadvantaged members of the sample than of the continuing participants. This difference is due mainly to characteristics of those who die rather than of voluntary dropouts, however (Cooney & associates, 1988; Norris, 1985). Whether dropouts from panel studies distort study findings has drawn mixed conclusions (Streib, 1966; Schaie & associates, 1973; Markides & associates, 1982; Goudy, 1985; Rusin & Siegler, 1985; Norris, 1987). Differences between deceased dropouts and study participants may not matter in some analyses because each cross-sectional wave should still be representative of the cross section of the population from which it was drawn.

If the objective is to measure change, the absence of higher proportions of sicker or disadvantaged members of the population may present some difficulties, as evidenced by the tendency of mean values on health variables to remain unchanged or to change less than one would expect in samples of elderly people. In our sample, for example, smokers are healthier than nonsmokers, possibly because the unhealthy smokers within the study area died before the sample was selected.

Whether those who die or are too ill to continue in a study present the same kind of problem for interpreting the data as voluntary dropouts, or indeed whether they present any problem at all, remains a controversial issue (Markides, 1986; Markides, 1987; Norris & Goudy, 1986).

Methods

After the health and health care study had been underway for approximately 18 months, half of the original 630 nonrespondents were recontacted. Information was

Cynthia Thomas is with the Division of Health Services Organization and Policy, Department of Epidemiology and Social Medicine, Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, New York.

This work was supported by Grant No. PO1 AGO 3424 from the National Institute on Aging.

Table 1. Characteristics of nonrespondents and respondents on three study waves (means/percents)

Characteristics	Nonrespondents	Respondents		
		Wave 1	Wave 4	Wave 7
Sociodemographic				
Age	75.79[a]	75.36[a]	76.54[b]	77.08[c]
Male (%)	22.59[a]	33.66	31.80	31.15
Lives alone (%)	62.67[a]	42.88	46.99	48.17
Nonwhite (%)	3.67	5.01	6.05	5.90
Income (%)				
Under $5,000	35.00[a]	18.20	16.32	11.72
Over $15,000	25.83	23.04	23.86	21.95
Medicaid	15.23	14.61	12.35	11.77
Health				
Self reported (%)				
Fair/poor	49.72[a]	41.42[b]	37.27[b, c]	33.93[c]
Difficulties				
Getting around	20.93[a]	13.07[b]	13.13[b]	17.51[a, b]
Stairs	32.95[a]	19.71	19.24	19.57
Step stool	29.17[a]	10.42	—	—
Depression	0.71[a]	0.60	0.59	0.63
Observed (%)				
Frail	22.76[a]	13.64	14.28	9.04
Ill	12.23[a]	8.08[a, b]	10.58[a]	5.30[b]
Unhappy	23.24[a]	14.32[b]	14.09[b]	7.49[c]
Confused	7.64	7.61	6.60	2.98[a]
Source of care (%)				
None	9.20	7.40	8.95	9.84
Hospital	12.27	13.89	12.34	11.61
Group	28.83	25.45	24.24	22.38
Physician	47.24	50.61	51.97	53.74
Hospitalizations				
Number	.31.	.30	.31	.23
Mean total days	21.37	23.11	19.07	14.03
Average length of stay	18.78[a]	17.54[b]	12.34[a, b]	10.27[b]
Total	271	1855	1426	1201

—Data not available

[a-c]Means or percents with the same superscript (or no superscript) are not significantly different at $p < 0.05$ by t-test.

NOTE: Data for respondents and dropouts are weighted to adjust for unequal probability sampling in this and subsequent tables. Data are unweighted for nonrespondents. All totals are unweighted. The same differences in mean values shown in Table 1 between nonrespondents and respondents remain significant at < 0.05 using unweighted data for respondents.

obtained from about 86 percent of them (271 people) in the form of a completed interview or information about their death. From the living, information was obtained on age, sex, living arrangements, health and functional status, income, primary source of medical care, and any hospitalization in the preceding 12 months. In the analysis described below, nonrespondents are compared with respondents according to basic characteristics, and dropouts with nondropouts. Whether mean values on certain characteristics would have changed if nonrespondents had agreed to respond and, similarly, if sample characteristics would be different if dropouts had continued in the study were examined. Then the impact of nonresponse and attrition on results of multivariate analyses was explored.

Results

Table 1 shows characteristics of the nonrespondents compared with study respondents at the first, fourth, and seventh survey waves. Most of the nonrespondent interviews were completed during survey wave four. The nonrespondents—those who refused initially to take part in the study but later were willing to answer a series of questions—are considerably different in a number of respects from their would-be counterparts at wave four and, indeed, at waves one and seven as well, since the mean values of characteristics of study respondents are similar throughout the data collection period. Most noticeably, their health was poorer, whether self-reported or observed by interviewers, including their ability to get

around inside the house, climb stairs, or (available in contrast to wave 1 only) use a step stool. They were more often women, more likely to live alone, and to report annual incomes under $5,000. Although in poorer health and with higher rates of other unfavorable circumstances, they were not any older than study participants.

Two sets of characteristics reported in this table are key analytic measures in the health and aging study: source of medical care and hospitalizations. Although nonrespondents and respondents differ according to certain health and socioeconomic characteristics, distributions across the various sources of medical care are remarkably comparable, with 12 to 13 percent of nonrespondents and respondents alike reporting a hospital as the primary source of care, and 47 to 54 percent reporting a private physician as the care source. Nonrespondents' rates of hospitalization and mean total days in the hospital also are comparable to those of respondents at waves one, four, and seven. In general, then, study respondents are similar among themselves wave to wave, but differ considerably from the nonrespondents with respect to health and socioeconomic status. Nonrespondents and respondents are remarkably alike, however, according to key analytic measures of hospitalization rates and source of medical care.

It might be expected that wave four interview respondents would be a very different group if only nonrespondents had been enrolled. They would be expected to be sicker and more disadvantaged than they are. Indeed, one might suspect that the literature on the elderly residing in the community, which tends to find a surprisingly high proportion of spry, active people, could be biased if nonrespondents generally are sicker than respondents. Within the reported sample, the nonrespondents (each one weighted by 2 to represent the correct ratio of nonrespondents to respondents) were combined with wave four respondents, and mean values on characteristics were examined, in contrast to means for wave four respondents alone. These values are shown in Table 2. As can be seen, it makes little difference whether the nonrespondents are included or not. There are no apparent differences between the two groups on any characteristic except the proportion living alone. Apparently, the differences between respondents and nonrespondents were small enough and the response rates were high enough that mean values are unchanged. Still, it has not yet been shown that the nonrespondent group is unimportant. Disadvantaged participants may differ in some important way from disadvantaged nonparticipants, and such a difference many affect the relationships among variables, if not mean values of characteristics themselves. This point will be discussed later.

Not only must nonresponse be contended with, but attrition as well. Whether nonrespondents pose a threat to survey results, dropouts may create biases. Table 3 compares respondents with two categories of dropouts: those who died or entered nursing homes and those who chose not to continue in the study. Measures are taken either from wave 1 data for unchanging characteristics such as gender and age or, for characteristics such as health status, at the previous survey wave when all three groups answered questions within the same timeframe. Information from adjoining waves of dropouts has been collapsed to provide adequate sample sizes for analysis. Table 3 shows that dropouts, particularly those who died, tend to be different from those who continue in the survey. The most striking differences are the extremely poor health, greater likelihood of being male, and greater age of those who died in contrast both to those who remained in the study and those who refused to continue. It is not surprising, of course, to find high rates of poor health among a group of people who will die within 6 months. The dropouts who left the study voluntarily are more often similar to those who continued. The voluntary dropouts are close to the same age, roughly similar in level of income, and on balance, seem to be just about as healthy as ongoing participants, as others have reported in the literature.

As for nonrespondents, whether mean values of sample characteristics would change if dropouts had not dropped out was determined by examining sample characteristics on a given survey wave, first with and then without those who died or left the study before the next wave. Table 4 shows that dropouts—essentially those

Table 2. Characteristics at wave 4: Respondents versus respondents plus nonrespondents

	Respondents at wave 4	Respondents at wave 4 plus nonrespondents
Sociodemographic		
Age	76.54	76.37
Male (%)	31.80	29.70
Lives alone (%)	46.99[a]	50.85
Nonwhite (%)	6.05	5.50
Income (%)		
Under $5,000	16.32	18.23
Over $15,000	23.86	24.26
Medicaid	12.35	12.81
Health		
Self reported (%)		
Fair/Poor	37.27	39.36
Difficulties		
Getting around	13.13	14.33
Stairs	19.19	21.42
Depression	0.59	0.61
Observed (%)		
Frail	14.27	15.52
Ill	10.58	10.77
Unhappy	14.09	15.61
Confused	6.60	6.72
Source of care (%)		
None	8.95	8.95
Hospital	12.34	12.28
Group	24.24	25.13
Physician	51.97	51.15
Hospitalizations		
Number	.15	16.99
Mean total days	15.76	18.14
Average length of stay	12.34	13.21
Total	1426	1988

[a] $p < 0.05$ (t-test)

Table 3. Characteristics of respondents and dropouts: Early and late (means/percents)

	Wave 2 respondents at wave 1	Wave 3 respondents at wave 2	Wave 2 and 3 dropouts: Died/ NH	Wave 2 and 3 dropouts: Refused/ Lost	Wave 6 respondents at wave 5	Wave 7 respondents at wave 6	Wave 6 and 7 dropouts: Died/ NH	Wave 6 and 7 dropouts: Refused or Lost
Sociodemographic								
Age	75.23	75.63	79.81[a]	74.77	76.69	77.08	81.97[a]	76.77
Male (%)	33.57	33.10	41.21	33.06	31.24[a, b]	31.17[a, b]	41.01[a]	19.23[b]
Lives alone (%)	43.77	46.50	46.06	30.99[a]	47.08	47.97	45.32	65.38[a]
Nonwhite (%)	5.40[a]	5.59[a]	3.03[a, b]	0.01[b]	5.98	5.91	2.16	13.46[a]
Income (%)								
Under $5,000	18.01	17.88	26.12[a]	14.57	15.68	15.09	26.88	16.28
Over $15,000	23.48	23.33	15.67	26.49	25.25	25.67	12.90	23.26
Medicaid	14.58	13.96	29.63[a]	9.62	12.16	11.67	20.14	17.31
Health								
Self-reported								
Fair or poor (%)	40.68[a]	40.18[a]	50.00[b]	45.96[a, b]	35.91[a, b]	35.77[a, b]	42.97[b]	26.09[a]
Difficulties (%)								
Getting around	12.32	11.47	38.27[a]	9.13	16.14	14.88	43.17[a]	23.08
Stairs	19.35[a]	18.85[a]	42.24[b]	11.62[c]	20.82	19.57	42.45[b]	28.85
Step stool	9.93	9.59	26.85[a]	6.49	—	—	—	—
Depression	0.60	0.61	0.63	0.55	0.04[a]	0.63[a]	0.95[b]	1.04[b]
Observed (%)								
Frail	12.92	12.84	39.60[a]	13.76	12.35	10.77	31.00[a]	10.81
Ill	7.40	11.00	25.33[a]	7.59	7.64	5.41	16.67[a]	4.76
Unhappy	14.03	14.93	25.34[a]	17.04	10.79[b]	6.22[b]	25.22[a]	20.45[a]
Confused	6.86	6.38	23.49[a]	7.59	4.32[b]	3.18[b]	16.67[a]	13.64[a]
Source of care (%)								
None	7.30	7.13	8.54	8.68	9.21	9.10	5.04	17.65[a]
Hospital	14.32[a, b]	13.86[a, b]	18.90[a]	10.74[b]	12.30	12.07	13.67	15.69
Group	25.29	25.83	21.34	25.21	23.97	24.14	24.46	21.57
Physician	50.40	50.43	49.39	82.89	52.05	52.31	53.96	41.18
Hospitalizations								
Number	.30[a]	.46[b]	.64[c]	.24[a]	.26	.23	.81[c]	.48
Mean total days	22.47	24.04	28.67	28.05	15.09	13.83[a]	24.95[b]	44.88[c]
Average length of stay	16.92	16.16	22.59	21.23	10.25[a]	10.30	16.14	24.74[a]
Total (per person hospitalized)	1,675	1,555	111	174	1,269	1,201	131	45

—Data not available

[a-c] Means or percents with the same superscript (or no superscript) are not significantly different at $p < 0.05$ by t-test

who died, since they and not voluntary dropouts had different characteristics from continuing participants—have virtually no impact on any of the mean values of sample characteristics. Although nonrespondents and dropouts who died are very different from ongoing study respondents, their presence (or absence) has little impact on the characteristics of the sample. However, this result does not preclude the possibility that these groups, in their absence, have the power to distort the analytic results of the study.

Next, the possible impact of nonrespondents and dropouts on multivariate analysis was examined. Dropouts and nonparticipants who are ill or have low incomes or some other unique characteristic may be different in such a way from those who are just as ill or poor, but who enroll in the study and remain as participants, as to change the relationship between illness or income measures and some other variable in a multivariate relationship. One key variable of interest in the health and aging study is the number of hospitalizations among respondents, and the factors that either lead to or predict a hospitalization.

Does attrition have an impact on the predictors of whether someone will be hospitalized? First, using logistic regression, results were compared at baseline for two sample groups: the entire sample, and those who responded to all seven waves of survey questions, the continuing participants. This comparison, in effect, takes account of the impact of attrition on the results.

The p values for each independent variable, indicating significance above or below the 0.05 level, are presented in Table 5. In both groups, poorer health, including self-assessed health and presence of difficulties, lead to the likelihood of a hospital admission. Within the entire sample, older people are more likely to be hospitalized than younger people and men more often than women. Among the continuing participants, however, being a man or being older is not associated with hospitalizations. One possible explanation is that men or older people who died were more likely to have been hospitalized, so that those who accounted for the relationship between either age or gender and hospitalizations dropped out of the study. Indeed, when the coefficient for gender and age in the two dropout groups is exam-

ined (Table 5, columns 3 and 4), gender significantly predicts hospitalizations for those who died but not for those who voluntarily left the study, and age predicts hospitalizations only for those who dropped out voluntarily. Consequently, it appears that the difference in the significance of gender and age between the entire sample and those who remain at the end of the study is due to attrition of certain hospitalization-prone men and older people.

The impact of the nonrespondent group on predictors of hospitalization can be assessed similarly. Columns 5 to 7 in Table 5 show the p values of logistic regressions for nonrespondents, for respondents at wave 4, and for respondents and nonrespondents combined. For five of the eight variables, age, gender, difficulties in daily living, fair-to-poor self-reported health, and presence of depression, the addition of nonrespondents to the wave 4 data does not cross the significance level of 0.05. Being nonwhite is no longer a significant variable among the combined wave 4 and nonrespondent group, and income and Medicaid status both have taken on significance. Given that small proportions of nonwhites were nonparticipants, one could argue that nonparticipating nonwhites may be atypical and that probably the true role of race is best ascertained in the wave 4 data without their inclusion. The true importance of Medicaid status and income are more difficult to fathom, since they only emerge as significant variables when the nonrespondents are combined with the respondents, and not within either group alone. The presence of a relatively high proportion of low income nonrespondents may be affecting the coefficient for the combined sample. It could be argued that failure to convince low income people to participate was a handicap to the study. Depending on whether nonrespondents are included with the wave 4 data, variables associated with race and income may (or may not) be associated with whether one is hospitalized.

Discussion

Nonrespondents in this study of health and aging in the Bronx are different in several respects from respon-

Table 4. Respondents' characteristics at selected intervals, with and without dropouts.

Characteristics	Wave 1 means		Wave 4 means		Wave 6 means	
	Without wave 2 dropouts	With wave 2 dropouts	Without wave 5 dropouts	With wave 5 dropouts	Without wave 7 dropouts	With wave 7 dropouts
Sociodemographic						
Age	75.23	75.36	76.39	76.54	77.08	77.32
Male (%)	33.57	33.66	31.64	31.80	31.17	31.35
Lives alone (%)	43.77	42.88	47.01	46.99	47.97	48.18
Nonwhite (%)	5.40	5.01	5.84	6.05	5.91	5.90
Income (%)						
Under $5,000	18.01	18.20	16.36	16.32	15.09	15.68
Over $15,000	23.48	23.04	24.10	23.86	25.67	25.25
Medicaid	14.58	14.60	12.02	12.35	11.67	12.16
Health						
Self reported						
Fair or Poor (%)	40.68	41.42	36.09	37.27	35.77	35.97
Difficulties						
Getting around	12.32	13.07	12.14	13.13	14.88	16.00
Stairs	19.35	19.71	17.93	19.19	19.57	20.69
Step stool	9.93	10.42	0.70	1.05	0.72	0.87
Depression	.60	.60	.57	.59	.63	.64
Observed						
Frail	12.92	13.64	12.88	14.27	10.77	11.02
Ill	7.40	8.08	9.40	10.58	5.41	5.63
Unhappy	14.03	14.32	13.16	14.09	6.22	6.82
Confused	6.86	7.61	5.98	6.06	3.18	3.91
Source of care (%)						
None	7.30	7.39	9.02	8.95	9.10	9.23
Hospital	14.32	13.89	12.30	12.34	12.07	12.35
Group	25.29	25.45	24.14	24.24	24.14	24.14
Physician	50.40	50.61	52.80	51.97	52.31	51.84
Hospitalizations						
Number	.30	.30	.14	.15	.12	.13
Days	22.47	23.11	15.44	15.76	12.08	14.99
Average length of stay	16.92	17.54	11.77	12.34	10.30	11.19
Total	1,676	1,855	1,336	1,426	1,201	1,269

Table 5. Factors associated with hospitalizations among nonrespondents, dropouts, and respondents: *p* values from logistic regressions

Characteristics	Wave 1 data: Entire sample (subsample)	Respondents to wave 7	Dropouts: Died/NH	Dropouts: Lost	Nonrespondents	Respondents at wave 4	Nonrespondents plus wave 4: subsample
Column number	1	2	3	4	5	6	7
Age	0.01	0.91	(−)0.71	0.01	(−)0.26	0.44	(−)0.67
Gender (male)	.00	.13	.01	.13	.67	.02	.05
Nonwhite	.65	(−).64	.81	.44	(−).36	.03	.13
Medicaid	.53	.65	.25	(−).01	.51	.54	.02
Income-$5,000	(−).30	.81	.10	(−).97	(−).19	(−).15	.00
Difficulties in daily living	.00	.00	.12	.00	.00	.00	.00
Health							
Fair/poor	.00	.00	.00	.04	.00	.00	.00
Depressed	.30	.30	.15	(−).36	.35	.34	.61
Total[a]	1,201	1,201	362	292	173[b]	1,426	1,426
R^2 [c]	.07	.04	.07	.12	.18	.05	.07

[a]Subsamples were selected for analyses in columns 1 and 7 so that totals would be equivalent for appropriate comparison groups
[b]Information missing on hospitalizations for 98 of 271 people; other missing values imputed to the sample mean
[c]Pseudo R^2, calculated according to Judge and associates, 1980, p. 601, equation 14.4.20

dents and among dropouts, those who have died are different from continuing study participants. When nonrespondents and dropouts are either merged with or removed from appropriate groups of respondents, their presence (or absence) does not in any way alter mean values on a series of characteristics used as independent and dependent variables in the study.

This lack of impact is based on three factors: the rates of nonresponse and attrition, both of which were relatively low for a sample of this age group in the particular study site; the magnitude of the difference in mean values on each characteristic between pairs of the separate groups; and the absolute numbers of dropouts, nonrespondents, and respondents. Using data from this study and similar ones a formula could be constructed specifying, given alternative values for these four variables, when nonresponse and attrition would change marginal values.

It is more difficult to interpret multivariate analyses than cross-sectional mean values of characteristics, however. Clearly, the difference in results with and without dropouts is due to different characteristics of those who die and, indeed, a comparison of the results first with dropouts and then without them adds information to an analysis rather than confusing matters. The dynamics of the multivariate relationships between nonrespondents and respondents is more complex, however. Nonrespondents among the elderly may be very different from respondents, even within a given category such as low income. Remedies for this potential problem, from the design of the survey, to procedures for calling back nonrespondents, to later analyses, may require even greater attention than they have received in the past.

References

Cooney T. M., Schaie, K. W., & Willis, S. L. (1988). The relationship between prior functioning on cognitive and personality dimensions and subject attrition in longitudinal research. *Journal of Gerontology: Psychological Sciences,* 43(1), Pl2-Pl7.

Criqui M. H., Austin, M., & Barrett-Connor, E. (1979). The effect of non-response on risk ratios in a cardiovascular disease study. *Journal of Chronic Diseases,* 32, 633-638.

Goudy, W. J. (1985). Sample attrition and multivariate analysis in the retirement history study. *Journal of Gerontology,* 40(3), 358-367.

Herzog, A. R., & Rodgers, W. L. (1988). Age and response rates to interview sample surveys. *Journal of Gerontology: Social Sciences,* 3(6), S200-205.

Judge, G. G., Griffiths, W.E., Hill, R.C., & associate (1980). *The theory and practice of econometrics* (p.601). New York: Wiley & Sons.

Markides, K. S. (1986). Letter to the Editor. *Journal of Gerontology,* 41(6), 806.

Markides, K. S. (1987). Characteristics of dropouts and prediction of mortality in a longitudinal study of older Mexican-Americans and Anglos. In R. A. Ward & S. S. Tobin (Eds.). *Health in aging: Sociological issues and policy directions* (pp. 86-97). New York: Springer.

Markides, K. S., Dickson, H. D., & Pappas, E. (1982). Characteristics of dropouts in longitudinal research on aging: A study of Mexican Americans and Anglos. *Experimental Aging Research,* 8(3), 163-167.

Norris, F. H. (1985). Characteristics of older nonrespondents over five waves of a panel study. *Journal of Gerontology,* 40(5), 627-636.

Norris, F. H. (1987). Effects of attrition on relationships between variables in surveys of older adults. *Journal of Gerontology,* 2(6), 597-605.

Norris, F. H., & Goudy W. J. (1986). Letter to the Editor. *Journal of Gerontology,* 41(6), 806-807.

Riegel, K. F., Riegel, R. M., & Meyer, G. (1967). A study of the dropout rates in longitudinal research on aging and the prediction of death. *Journal of Personality and Social Psychology,* 9(3), 342-348.

Rusin, M. J., & Siegler, I. C. (1985). Personality, dropout, and death. In E. Palmore, E. Busse, G. L. Maddox, & associates (Eds.). *Normal aging III* (pp. 242-246). Durham, NC: Duke University Press.

Schaie, K. W., Labouvie, G. V., Barrett, T. J., & associates (1973). Selective attrition effects in a fourteen-year study of adult intelligence. *Journal of Gerontology,* 28(3), 328-334.

Streib, G. F. (1966). Participants and drop-outs in a longitudinal study. *Journal of Gerontology,* 21, 200-209.

FEATURE PAPER

Nonresponse to Survey Questions by Elderly in Nursing Homes

Judith Garrard, Carol Skay, Edward R. Ratner, Robert L. Kane, and Hung-Ching W. Chan

Introduction

Item nonresponse exists when answers to some, but not all, questions in a health survey are missing. Different approaches (Sudman & Bradburn, 1974; Fowler, 1984; Rubin, 1976; Little & Rubin, 1987) exist for handling this problem; however, statistical solutions require an assumption that the data are missing at random (Rubin, 1976; Little & Rubin, 1987). Often this assumption is not evaluated, and a growing body of health survey research suggests that there is nonresponse bias in some types of items (Siemiatycki & Campbell, 1984). Age has been found to be one source of such bias. The age range in most studies has been quite broad, however, and there has been a tendency to collapse respondents 65 years and older into a single category. Although decline in biological and cognitive function has been reported, there is evidence that major differences exist among the young-old (65 to 74), old (75 to 84), and oldest-old (85 or older) (Suzman & Riley, 1985). With the growth in the elderly population, there is increased interest in health surveys with individuals 65 years and older. Research on nonrandom nonresponse in health surveys with elderly has been recent and limited to only a few studies (Herzog & associates, 1983; Brock & associates, 1986; Wallace, 1987; Colsher & Wallace, 1989).

The purpose of this paper is to examine item nonresponse in health surveys with elderly in a particular segment of the population, those living in nursing homes. Although it has been estimated that only 5 percent of the elderly population are in a nursing home on any one day, approximately 25 to 40 percent are expected to spend one or more days in such a facility after their 65th birthday (Vicente & associates, 1979). The cognitive status of many nursing home residents precludes any possibility of an interview. Others, however, suffer only from functional or physical impairments (Tennstedt & McKinlay, 1987) but are as cognitively intact as their able-bodied counterparts in the community and are therefore capable of being a respondent in an interview. Recent Federal regulations for certifying nursing homes require that selected nursing home residents be interviewed as part of a standard assessment of quality of care (Federal Register, 1988). Nonetheless, there are factors that must be taken into account in designing survey research with this population, such as the choice of which groups to sample, careful consideration of subject eligibility criteria, and methods for screening potential respondents.

A systematic examination of variables that impact quality of data from elderly across a variety of settings and variables that are unique to particular settings has not been reported in the literature. For this reason, this paper first describes briefly a matrix for organizing such a study and specifies where in that matrix the focus of this presentation belongs. Next, results of a recently completed study on item nonresponse by elderly in nursing homes will be summarized, and finally some specific suggestions for health survey research with elderly in nursing homes and recommendations for further research on item nonresponse will be provided.

A Matrix for Examining Item Nonresponse in Research with Elderly

Although health survey research in the social sciences is generally limited to the interview method, two additional data collection procedures are included here: the self-administered questionnaire and the record review.

Judith Garrard, Robert L. Kane, and Hung-Ching W. Chan are with the Division of Health Services Research, School of Public Health, University of Minnesota, Minneapolis. Edward R. Ratner is with the St. Paul-Ramsey Medical Center, St. Paul, Minnesota. Carol Skay is with the Professional Assessment Services Division of National Computer Systems, Minneapolis.

This paper was prepared under the auspices of a Special Emphasis Research Career Award in Behavioral Geriatrics to Judith Garrard from the National Institute on Aging, NIH Grant No. 1K01-AG 000434-01. Funding for the Mountain States Geriatric Nurse Practitioner Project was provided by grants from the Robert Wood Johnson Foundation to the University of Minnesota School of Public Health, from the Health Care Financing Administration to The RAND Corporation, and from the W.K. Kellogg Foundation to the Mountain States Health Corporation; Robert L. Kane, Principal Investigator.

Figure 1. Matrix for the study of item nonresponse bias in different settings with data collection procedures

	Interview							Questionnaire			Record review
	Face-to-face interview				Telephone interview			Self-administered			
	Subject		Proxy		Subject	Proxy		Subject	Proxy		
	Self-Report	Demo/ Observ	Relative	Health professional		Relative	Health professional		Relative	Health professional	
Elderly in the community											
1. Study design variables											
2. Subject variables	A										
3. Environmental variables											
Elderly in the hospital											
1. Study design variables											
2. Subject variables											
3. Environmental variables											
Elderly in the hospice											
1. Study design variables											
2. Subject variables											
3. Environmental variables											
Elderly in nursing homes											
1. Study design variables	B	B									B
2. Subject variables	B	B									B
3. Environmental variables											

A = Studies by Brock & associates, 1986; Wallace, 1987; Herzog & associates, 1983; Colsher & Wallace, 1989
B = Data presented in this paper

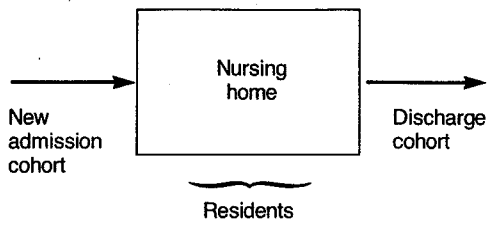
As shown in Figure 1, these three types of data collection procedures are listed across the top of the matrix. An interview can be face-to-face or telephone. Within each, the respondent can be the subject or a proxy, and a proxy can be a relative or friend, a health professional, or other. In a face-to-face interview, information can be obtained by asking the individual to respond directly to an item, or by asking him or her to demonstrate a capability, or by having a trained interviewer observe the subject's condition or behavior during the course of the interview.

The settings in which elderly are studied could determine the type of information that could be collected and the opportunity (or lack of it) for validation. Characteristics of individuals living in these settings might also have implications for item nonresponse. The four settings on the left of the matrix describe elderly in the community, the hospital, the hospice, and the nursing home. Within each setting, variables that might impact item nonresponse could be classified by at least one of three categories: study design, subject, and environmental or institutional. Examples of study design variables might include variations in the type of subject group sampled (such as whether the sample represents a new admissions cohort or elderly currently living in the home or a discharge cohort) or variables related to interviewer (or record abstracter) characteristics, such as training or quality control in the field. Subject variables are those that are often included in papers on missing data, such as respondent demographics, health status, or mental status. With nursing homes and hospitals, environmental and institutional variables can be considered as having an impact on item nonresponse, for example, rural or urban location, number of beds, type of ownership. Whether there are analogous types of variables that might be classified as environmental in studying nonresponse by elderly in the community is an issue left unresolved. Using this matrix, previous research studies in the literature on nonresponse by elderly in the community are shown as A in Figure 1 in the face-to-face column. The study described in this paper is shown as B in the face-to-face interview and the record review columns.

Choosing a Sample of Nursing Home Elderly

It would be virtually impossible to conduct a health survey with a random sample of all residents in a nursing home because not all are equally capable of providing self-report information. Not only must this fact be taken into account in formulating the research question, it is also a consideration in developing a sample design. Even with subjects who are alert and ostensibly capable of an interview, there are some additional decisions to be made about the sample. One of the major points to be emphasized here is the importance of distinguishing among at least three groups of individuals in a nursing home (Figure 2) in deciding who to sample and how: a new admissions cohort (individuals who enter a nursing

Figure 2. Groups of elderly in nursing homes



home during a specified period of time), residents (who are living in the nursing home, regardless of when they entered or subsequently left), and a discharge cohort (those who were discharged from the nursing home for any reason: death, returned to the community, hospitalized, transferred to another facility). In sampling a new admissions or discharge cohort, a period of time has to be specified because there are so few actual admissions or discharges on any one day. To draw a sample of residents, a time, that is, a particular day, can be used to "freeze" the sampling frame.

Space does not permit a discussion of the differences among these three groups; however, the reader is referred to the literature on differences in short-stay and long-stay residents (Liu & Manton, 1984; Keeler & associates, 1979). Suffice to say, these groups have clearly distinguishable characteristics: compared to all residents in the facility, a new admissions cohort tends to be younger, with a greater percentage of males, proportionally more with a diagnosis of hip fracture and stroke and less with dementia, and a greater probability over the short term of being discharged to the community or dying. The discharge cohort consists of those who are leaving. If the intent is to interview those being discharged to the community, then this subset of a discharge cohort represents those who have survived the problems that brought about their admission to the facility in the first place. This group will include a very small proportion of residents who have resided in the home for several months or more but the majority will have been newly admitted within the last 1 to 6 months. The residents living in the home are likely to be the most chronically ill and cognitively impaired. The research question should determine which group is the most appropriate to survey. The characteristics of the type of group chosen could have implications for item nonresponse and should be considered separately in the development of a screening protocol for sample selection and decisions about which kinds of items to include in the interview protocol. The extent to which these three groups differ in item nonresponse has not been reported in the literature.

Study of Nonresponse Bias in Nursing Home Elderly

The purpose of this study was twofold: (1) To determine the level of item nonresponse in interviews with elderly in nursing homes and missing data in records and (2) to examine whether items were missing at random or were biased by differences between resident groups (new admissions versus residents), and within each of those two groups, by sex, age, self-reported health, or mental status. The hypotheses were the following: Item data from elderly in nursing homes would not be missing at random. Based on gerontological research, item nonresponse is expected to be biased by resident group (with residents having more missing data than new admissions). From previous research (Colsher & Wallace, 1989) on nonresponse by elderly in the community, differences were expected by sex (males having more missing data), age group (progressively more missing data from young old to oldest old), self-reported health (greater nonresponse with poor health), and mental status (greater amounts of missing data with increasing cognitive impairment).

Method

Respondents. The data set used was from a previous interview study on quality of care of residents in nursing homes (Kane & associates, 1989). For purposes of this nonresponse study, only the data from the control groups were analyzed. To be eligible for participation in the original study, an individual had to be alert enough to respond to the interviewer's questions, be able to give informed consent, and agree to participate in the interview. Alertness was determined by the interviewer at the baseline interview on the basis of specific behavioral criteria.[1] Within each of 5 nursing homes, 100 subjects were to be chosen: 60 consecutive new admissions and 40 residents selected randomly from a list of all living in that home at the beginning of the study. After eliminating individuals who did not meet the eligibility criteria and those below the age of 65, 246 interviews were available from the new admission cohort and 152 from the resident group. All subjects were 65 years old or older at the time of the interview. Within the new admission cohort, the mean age was 82.73 years, and 68 percent were female; within the resident group, mean age was 85.1 years, and 80 percent were female. In both subject groups, 99 percent or more were Caucasian.

Instrument. A structured 1.5-hour interview was conducted by a registered nurse not affiliated with the nursing home; all nurse-interviewers received a standard 3-day training program conducted by the RAND Corporation. All interviewers were Caucasian and female; each was assigned to a different nursing home. The interview protocol, developed, validated, and used in a previous study (Kane & associates, 1982), consisted of 149 items, of which 68 items were not contingent on previous responses and were designated as required to be administered in instructions to the interviewer. This nonresponse study was limited to these 68 items.

[1]The interviewer was instructed to evaluate the individual's alertness by a greeting, or if that failed by touching the arm, or if that failed by gently pinching the upper trunk. The individual was then rated as alert, semi-alert or comatose. For purposes of this study, only those evaluated as alert were included.

For purposes of this study, an appropriate response was one in which the information was elicited. Incorrect answers on the factual questions and negative responses on the self-report items were included as responses. A nonresponse was one in which the subject stated that he or she did not know or could not respond, gave no response, the item was not administered or was not appropriate (for example, a blind subject could not read the eye chart), the data were not available (in the record), or the response was missing in the data set. Data could be missing for a variety of reasons, depending on the content of the item; therefore, we assigned each of the 68 items to one of five categories based on these reasons. Different coding options existed within each of these categories for the interviewer to explain why a response was not available. In addition to these nonresponse codes available to the interviewer in the field, a code for "blank" was added when no code was given. The five categories were the following:

1. Record item (10 items), data to be gathered from the nursing home record by the interviewer before the interview (for example, subject's sex, birth date, legal status). A nonresponse code for "data missing from record" was available.
2. Interviewer item (30 items), required that the interviewer either supply the information (for example, time of the interview, her code number) or that she make an observation of the respondent's condition or situation during the interview (for example, evidence of incontinence, blood pressure reading, presence of a hearing aid). A nonresponse code of "unable to determine" was available for items that required information about the subject.
3. Factual item (3 items), subject's response could be validated (for example, today's date, name of current president). Nonresponse codes included subject's statement, "I don't know," and for "no response" from subject.
4. Demonstration item (5 items), items that required the respondent to demonstrate his or her capability (for example, use a spoon, read an eye chart). Nonresponse codes were available to indicate that subject "states unable to respond," to indicate "no response," and to code "not administered or not applicable."
5. Self-report item (20 items), items about the respondent's feelings, opinions, or other subjective states in which the answers were not or could not be validated by another source (for example, presence of pain, opinion about food, feeling of happiness, visits by adult children). Nonresponse codes were available for subject to state "I don't know," or to indicate that subject "refused to answer," or to show that subject "did not respond."

Statistical Analysis. This initial study was limited to a univariate analysis using chi-square analysis to determine the possibility of bias at the item level. The study began with an analysis of differences in response frequencies between the new admission and resident cohorts with each of the 68 items. Within each subject group, it went on to examine the potential of nonresponse bias by sex, age group: 65 to 74, 75 to 84, 85 or older, self-reported health: excellent, good, fair, poor, and level of cognitive impairment based on the 10-item short portable Mental Status Questionnaire (MSQ) (Pfeiffer, 1975) using standard scoring levels without adjustment for education or race: 0 to 2 errors = cognitively intact, 3 to 4 errors = mild impairment, 5 to 7 errors = moderate impairment, 8 to 10 errors = severe impairment.

Results and Discussion

The percentage of missing responses across these 68 items ranged from 0 to 10 percent, with the exception of three items, a record item for legal status (16 to 20 percent missing) and two self-report items: whether life was interesting or not (9 to 15 percent missing), and whether residents felt they could keep as many personal possessions in their rooms as they liked (10 to 14 percent missing). Item nonresponse rates within both subject groups are given in Table 1. The mean levels of nonresponse varied by type of item: nonresponse for the nine record items (without legal status) was less than 1 percent and between 1 and 2 percent for the interviewer items. Average nonresponse levels were highest for the three types of items that depended on the respondent's cooperation: 3 to 8 percent of the subjects had missing data on the factual questions, and approximately 7 percent each on the demonstration and self-report items.

Differences in frequency of response and nonresponse between the two groups of nursing home elderly (new admissions and residents) were virtually nonexistent: of the four items in which significant ($p < 0.05$) differences were found, the resident group had a greater percentage of nonresponses. Further analysis showed that within each of these two groups, there appeared to be no response bias by sex, age group, or self-reported health. Only three items were significant by age group, and none was significant by sex or self-reported health.

Nonresponse Bias by Mental Status. There was a major and consistent bias in item nonresponse by mental status score within both the new admission and the resident groups. In both groups the percentage of items with missing data was greatest for those subjects with severe cognitive impairment. Items in which there was a statistically significant difference in distribution of subjects by response or nonresponse and mental status level are shown in Table 2.

These data also demonstrate a clear difference in nonresponse by type of item. Of the 10 record items, only the item on legal status showed a statistically significant difference by mental status score. Further examination of the raw data showed that the absence of legal status for both resident groups was concentrated in one of the nursing homes, although there was a definite bias ($p < 0.0001$): 34 to 38 percent of the severely cognitively impaired had this item missing, compared to 3 to 6 percent of those who were intact.

Of the 30 interviewer items, only three within the new admission data set showed a significant difference in frequency of missing data, report of blood pressure, apical pulse readings, and whether a resident was wear-

Table 1. Percent item nonresponse across all respondents in new admission cohort (N = 246) and long stay group (N = 152)

Type	Item	Description	New admissions % missing	Long stay group % missing
R	5	Most recent admission	0	0
R	6	First admission to facility	0	0
R	7	Birthdate	0	0
R	8	Sex	0	0
R	9	Ethnic background	2.0	0.7
R	10	Marital status	2.4	0
R	11	Legal status	16.3	19.7
R	13	Admitted from	0.4	0
R	15	One or more diagnoses	1.6	1.3
R	16	Current status of patient	0	0
I	1	Patient identification	0	0
I	2	Wave number	0	0
I	3	Trial date	0	0
I	4	Interviewer number	0	0
I	26	Time of interview	0.4	0
I	27	Level of consciousness	0	0
I	55	Evidence of incontinence	1.2	0.7
I	56	Level of mobility	0.4	2.0
I	79	Blood pressure	2.0	0.7
I	80	Apical pulse	2.0	0.8
I	81	Respiration rate	2.8	2.0
I	86	Wear hearing aid	2.0	0.7
I	87	Raise voice to hear	2.0	1.3
I	88	Emaciated	0.8	0
I	91	Edema in legs	2.0	1.3
I	92	Temperature diff in legs	3.3	3.3
I	93	Cyanosis in legs	2.0	1.3
I	94	Mottling in legs	2.0	0
I	95A	Articulation—R hand	2.8	0
I	95B	Articulation—R elbow	2.4	0
I	95C	Articulation—L hand	1.6	0
I	95D	Articulation—L elbow	2.8	0.7
I	95E	Articulation—R hip	4.9	3.9
I	95F	Articulation—R knee	3.7	5.3
I	95G	Articulation—L hip	6.1	2.6
I	95H	Articulation—L knee	5.3	5.3
I	96A	Amputated below shoulder	5.3	3.3
I	96B	Amputated below knee	5.3	2.6
I	96C	Amputated above knee	4.5	3.3
I	97	Tracheostomy	0.8	0.7
D	47	Use spoon	4.5	3.3
D	48	Spread butter on bread	5.7	3.3
D	49	Put on shirt	9.3	7.2
D	84	Read eye chart—5 ft.	6.5	9.9
D	85	Read eye chart—10 ft.	7.3	11.8
F	28	Today's date question	2.4	6.6
F	29	Day of week question	2.4	7.2
F	40	Who is president	5.7	11.2
S	39A	Reading past month	2.8	2.6
S	39B	Sewing past month	2.8	6.6
S	39C	Played games past month	2.8	7.2
S	69	Gone outside home	6.1	7.2
S	70	Stayed out overnight	10.2	8.6
S	71	Children visited	6.9	9.2
S	89	Trouble chewing	5.7	3.3
S	98A	Have any joint pain	6.5	3.3
S	98B	Any chest pain	6.9	3.9
S	98C	Any shortness of breath	5.7	5.3
S	98D	Any dizziness	6.1	5.3
S	98E	Any itching	6.9	4.6
S	98F	Any headaches	7.3	5.6
S	99	Any severe pain	6.5	6.6
S	100	Bothered by nerves	6.1	8.6
S	101	Felt happy past month	6.9	7.9
S	102	Life interesting	8.5	14.5
S	113A	Food good	6.9	7.2
S	113B	Room clean	7.7	5.3
S	113C	Keep possessions in room	14.2	10.5

Type of item: R = record, I = interviewer, D = demonstration, F = factual, S = self report.

ing a hearing aid. Within the resident group, none of the interviewer items had a statistically significant bias by cognitive level.

All five of the demonstration items showed a statistically significant nonresponse bias by mental status level. As shown in Table 2, 14 percent of new admissions who were severely cognitively impaired had missing data on the item requiring a demonstration of the use of a spoon, 18 percent on spreading butter on bread, and 24 percent on putting on a shirt. The possibility existed that the

Table 2. Percent of item nonresponse within each MSQ[a] score level for new admission cohort and long stay group

			New admissions (N = 246) MSQ score level					Long stay group (N = 152) MSQ score level				
Number of subjects			50	80	45	71		60	35	20	37	
Type	Item	Description	Severe	Moderate	Mild	Intact	p^b	Severe	Moderate	Mild	Intact	p^b
R	5	Most recent admission	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	6	First admission to facility	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	7	Birthdate	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	8	Sex	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	9	Ethnic background	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	10	Marital status	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	11	Legal status	34.0	20.0	6.7	5.6	.0001	38.0	11.4	10.0	2.7	0.001
R	13	Admitted from	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	15	One or more diagnoses	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
R	16	Current status of patient	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	1	Patient identification	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	2	Wave number	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	3	Trial date	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	4	Interviewer number	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	26	Time of interview	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	27	Level of consciousness	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	55	Evidence of incontinence	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	56	Level of mobility	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	79	Blood pressure	8.0	1.3	0	0	.01	. . .	. . .	. . .	. . .	. . .
I	80	Apical pulse	8.0	1.3	0	0	.01	. . .	. . .	. . .	. . .	. . .
I	81	Respiration rate	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	86	Wear hearing aid	8.0	1.3	0	0	.01	. . .	. . .	. . .	. . .	. . .
I	87	Raise voice to hear	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	88	Emaciated	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	91	Edema in legs	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	92	Temperature diff in legs	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	93	Cyanosis in legs	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	94	Mottling in legs	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95A	Articulation—R hand	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95B	Articulation—R elbow	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95C	Articulation—L hand	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95D	Articulation—L elbow	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95E	Articulation—R hip	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95F	Articulation—R knee	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95G	Articulation—L hip	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	95H	Articulation—L knee	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	96A	Amputated below shoulder	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	96B	Amputated below knee	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	96C	Amputated above knee	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
I	97	Tracheostomy	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
D	47	Use spoon	14.0	1.3	2.2	2.8	.01	8.3	0	0	0	.05
D	48	Spread butter on bread	18.0	2.5	2.2	2.8	.001	8.3	0	0	0	.05
D	49	Put on shirt	24.0	2.5	0.0	7.0	.001	15.0	2.9	0	2.7	.05
D	84	Read eye chart—5ft.	28.0	2.5	0.0	0.0	.001	23.3	2.9	0	0	.0001
D	85	Read eye chart—10ft.	30.0	2.5	0.0	1.4	.001	28.3	2.9	0	0	.0001
F	28	Today's date question	10.0	0	2.2	0	.01	15.0	2.9	0	0	.01
F	29	Day of week question	10.0	1.3	0	0	.01	18.3	0	0	0	.001
F	40	Who is president	18.0	5.0	2.2	0	.001	25.0	5.7	0	0	.001
S	39A	Reading past month	12.0	1.3	0	0	.001	. . .	. . .	. . .	. . .	. . .
S	39B	Sewing past month	14.0	0.0	0	0	.001	16.7	0	0	0	.001
S	39C	Played games past month	12.0	1.3	0	0	.001	16.7	2.9	0	0	.01
S	69	Gone outside home	26.0	2.5	0	0	.001	16.7	2.9	0	0	.01
S	70	Stayed out overnight	24.0	12.5	4.4	1.4	.001	18.3	0	0	0	.01
S	71	Children visited	24.0	2.5	0	4.2	.001	23.3	0	0	0	.0001
S	89	Trouble chewing	24.0	2.5	0	0.0	.0001	8.3	0	0	0	.05
S	98A	Have any joint pain	22.0	3.8	0	2.8	.0001	. . .	. . .	. . .	. . .	. . .
S	98B	Any chest pain	24.0	2.5	0	4.2	.0001	10.0	0	0	0	.05
S	98C	Any shortness of breath	22.0	2.5	0	1.4	.0001	11.7	2.9	0	0	.05

Table 2. Percent of item nonresponse within each MSQ[a] score level for new admission cohort and long stay group—Continued

			New admissions (N=246) MSQ score level					Long stay group (N=152) MSQ score level				
	Number of subjects		50	80	45	71		60	35	20	37	
Type	Item	Description	Severe	Moderate	Mild	Intact	p^b	Severe	Moderate	Mild	Intact	p^b
S	98D	Any dizziness	22.0	2.5	0	2.8	.0001	13.3	0.0	0	0	.01
S	98E	Any itching	24.0	5.0	0	1.4	.0001	11.7	0.0	0	0	.05
S	98F	Any headaches	26.0	3.8	0	2.8	.0001	13.3	2.9	0	0	.05
S	99	Any severe pain	26.0	3.8	0	2.8	.0001	15.0	2.9	0	0	.01
S	100	Bothered by nerves	24.0	3.8	0	0	.0001	20.0	0	0	2.7	.001
S	101	Felt happy past month	28.0	3.8	0	0	.0001	18.3	0	0	2.7	.01
S	102	Life interesting	34.0	3.8	2.2	0	.0001	31.7	2.9	5.0	2.7	.0001
S	113A	Food good	26.0	5.0	0	0	.0001	16.7	2.9	0	0	.01
S	113B	Room clean	28.0	3.8	4.4	0	.0001	11.7	2.9	0	0	.05
S	113C	Keep possessions in room	38.0	11.3	6.7	5.6	.0001	25.0	2.9	0	0	.0001

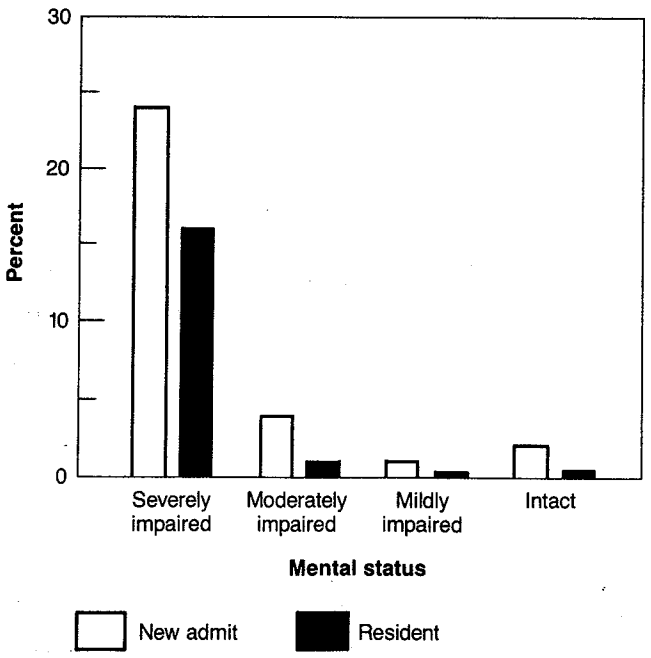
[a]MSQ—Short Portable Mental Status Questionnaire
[b]p value from 2×4 Chi square analysis ($df=3$)
Type of item: R = record, I = interviewer, D = demonstration, F = factual, S = self report

data were missing because the individual was paralyzed or blind, and the codes of "not applicable" indicated that this was the case in approximately one third to one half of these missing responses.

The three factual questions showed a statistically significant difference with the largest amount of data missing among the severely impaired: 10 to 18 percent of the cognitively impaired in the new admission group; 15 to 25 percent in the resident sample.

Finally, all of the 20 self-report items showed a significant difference in level of nonresponse by mental status among the new admission cohort. As shown in Figure 3, an average across these 20 items of 24 percent of the subjects had missing responses among the severely impaired (compared with 3 percent missing among the moderately impaired, and less than 1 percent in each of the other two mental status levels). Within the resident group, 18 of the items showed statistically significant differences, with an average of 17 percent of the subjects missing among the severely impaired, dropping off to 2 percent among the moderately impaired and close to 0 percent in the other two levels.

Figure 3. Mean percent nonresponse by mental status (self-report items)



On further examination it was found that the self-report items were missing because there was about an even split between no response from the subject and the omission of any code. Given the availability in this interview protocol of three other options to show why a valid response was not coded (that is, subject states "I don't know," or "Refused," or "No response" from subject), the existence of a blank might indicate a data quality problem in the field. Alternatively, there are clearly subjects who cannot or will not respond to many of these self-report items, perhaps because of physical conditions or depression or feelings about the negative social appropriateness of their potential answers ("No, my [adult] children never visit me").

Discussion and Recommendations

The average across all items of 0 to 8 percent missing data in this study of nursing home elderly was somewhat higher than that of 0 to 4 percent nonresponse from elderly in the community (Colsher & Wallace, 1989). There was considerable variation in the level of missing data, depending on the source of the information and the behavior required on the part of the interviewer or respondent. Nursing home records had a high degree of completeness, with a range of 0 to 2 percent missing (with the exception of the legal status item). Since these nine items consisted of fundamental information about the individual that was often required on legal or finan-

cial documents, this level of completeness cannot be assumed to be representative of that of clinical notes in the same charts.

Although some of the items in which the interviewer was primarily responsible could be explained by the "not applicable" codes, there was concern about the higher levels of 5 to 6 percent of missing data, which were actually blanks. Perhaps it is more difficult to maintain quality control of interviewer behavior in health surveys in nursing homes. The 3 (out of 30) items that showed a mental status bias depended on the interviewer to perform a task (take a blood pressure) or make an observation. Although this finding may be more a matter of a Type I error rather than actual fact, there was concern that interviewers (even nurse interviewers) may be more reluctant to carry out these tasks with severely cognitively impaired new admissions, who are possibly clinically sicker than the greater numbers of cognitively impaired but clinically more stable residents.

The three types of items that required respondent capability (demonstration, factual, and self-report) all had higher levels of nonresponse, and this is the data set most comparable to health surveys with elderly in the community. Unlike their counterparts in the community, however, nursing home elderly did not show a bias in nonresponse or missing data by sex, age, or self-reported health. Perhaps the need for the kind of care that nursing homes provide transcend response differences within these three variables.

The finding that missing data were biased by mental status score was not surprising, although missing data among a greater proportion of moderately cognitively impaired residents had been expected.

This was a preliminary study, limited to five nursing homes and a modest sample size; therefore, generalization to other nursing homes or samples of elderly residents must be made with caution. From this study, the following conclusions have been drawn:

1. Nonresponse levels. In health surveys of elderly in nursing homes, the level of item nonresponse can be expected to be higher (3 to 8 percent), on the average than that found for elderly in the community (0 to 4 percent), depending on the cognitive level of the respondents and the type of item (observation, demonstration, self-report).
2. Sample selection. Despite the lack of differences between the new admission cohort and the resident group in missing data, we nevertheless maintain a resolve to differentiate between these two groups in designing research with this population. The gerontological literature shows these groups to be different (and also different from a discharge cohort), and those differences are sufficient to warrant consideration in developing a sampling design.
3. Mental status questions. Any study of elderly in nursing homes should include some measure of mental status level, whether it is the Mental Status Questionnaire as used in this study or one of the other standard measures (Kane & Kane, 1981). The variables found to be important in determining nonresponse bias in health surveys with elderly in the community, that is, age, sex, or health status, will probably not be sufficient for examining nonresponse bias for elderly in nursing homes. Furthermore, a reliance on careful screening criteria for alertness probably will not rule out the possibility of missing data due to cognitive impairment. On the other hand, the difference in the amount of missing data between severely impaired and moderately impaired (or better) was dramatic, often by a factor of 10 to 20 percent. This suggests that screening by mental status, but excluding severely cognitively impaired elderly might be acceptable, if the intent is to reduce the amount of missing data. This also suggests that if severely cognitively impaired elderly are included in health surveys in nursing homes, their subgroup means should be considered in the substitution of means for this subset of new admissions and residents.

Finally, research on the topic of item nonresponse with this population of elderly has barely begun. We are interested in exploring differences in missing data and inconsistent data gathered by different data collection procedures (for example, birth date reported in the record, and as asked of the respondent in the form of birth date and then again as age); in the possibility of examining a subset of items such as those relating to a standard set of activities of daily living, by interviews with a proxy (relative versus staff nurse) and self-report and by record review; and in comparing nonresponse rates across multiple waves of interviews with the same subject over a 1-year period. The results of this study suggest that data are not missing at random in a sample of elderly in nursing homes, and future health survey research should examine this possibility rather than make the assumption of random item nonresponse.

References

Brock, D. B., Lemke, J. H., & Woolson, R. F. (1986). Identification of nonrandom item nonresponse in an epidemiologic survey of the elderly. *Proceedings of the Section on Survey Research Methods.* Washington, DC: American Statistical Association.

Colsher, P. L., & Wallace, R. B. (1989). Data quality and age: Health and psycho-behavioral correlates of item nonresponse and inconsistent responses. *Journal of Gerontology: Psychological Sciences,* 44, 45-52.

Federal Register. (1988, June 17). 53(117), pp. 22850-22861. Medicare and Medicaid: Long-term care survey—final rule. Office of the Federal Register, National Archives and Records Administration. Washington, DC: U.S. Government Printing Office.

Fowler, F. J. (1984). *Survey research methods.* Beverly Hills, CA: Sage Publishing.

Herzog, A. R., Rodgers, W. L., & Kulka, R. A. (1983). Interviewing older adults: A comparison of telephone and face-to-face modalities. *Public Opinion Quarterly,* 47, 405-418.

Kane, R. A., & Kane, R. L. (1981). *Assessing the elderly: A practical guide to measurement.* Lexington, MA: Lexington Books.

Kane, R. L., Garrard, J., Buchanan, J., & associates. (1989). Assessing the effectiveness of geriatric nurse practitioners. In M. D. Mezey, J. E. Lynaugh, & M. M. Cartier (Eds.). *Nursing homes and nursing care: Lessons from the teaching nursing homes.* New York: Springer Publishing Company.

Kane, R. L., Riegler, S., Bell, R., & associates. (January, 1982). *Predicting the course of nursing home patients: A progress report* (N-1786-NCHSR). Santa Monica: Rand Corporation.

Keeler, E. B., Kane, R. L., & Solomon, D. H. (1979). Short- and long-term residents of nursing homes. *Medical Care,* 19(3), 363-369.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley & Sons.

Liu, K., & Manton, K. G. (1984). The characteristics and utilization pattern of an admission cohort of nursing home patients (II). *The Gerontologist,* 24, 70-76.

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society,* 23, 433-441.

Rubin, D. B. Inference and missing data. (1976). *Biometrika,* 63, 581-592.

Siemiatycki, J., & Campbell, S. (1984). Nonresponse bias in early versus all responders in mail and telephone surveys. *American Journal of Epidemiology,* 120, 291-301.

Sudman S., Bradburn, N. M. (1974). *Response effects in surveys.* Chicago, IL: Aldine Publishing Company.

Suzman, R., & Riley, N. W. (1985). The oldest old. *Milbank Memorial Fund Quarterly,* 63, 177-451.

Tennstedt S. L., & McKinlay, J. B. (1987, July). Choosing the most appropriate field approach for older populations: the case for mixed-mode surveys (DHHS Pub. No. (PHS) 88-1214). *Proceedings of the 1987 Public Health Conference on Records and Statistics* (pp. 427-432) Washington, DC: U.S. Department of Health and Human Services.

Vicente, L., Wiley, J. A., & Carrington, R. A. (1979). The risk of institutionalization before death. *The Gerontologist,* 19, 361-367.

Wallace, R. B. (1987, July). The relationship of cognitive function, health status and mood to missing data and inconsistent responses in an interview of the elderly (DHHS Pub. No. (PHS) 88-1214). *Proceedings of the 1987 Public Health Conference on Records and Statistics* (pp. 423-426) Washington, DC: U.S. Department of Health and Human Services.

FEATURE PAPER

The Consequences of Accepting Proxy Respondents on Total Survey Error for Elderly Populations

Willard L. Rodgers and A. Regula Herzog

Introduction

In evaluating any survey design or data collection procedure it is important to think in terms of total survey error rather than restricting one's vision to a specific type or source of error. It is entirely possible to find that a particular innovation is effective in decreasing one type of error, but that that beneficial effect is overwhelmed by increases in other types of error. Narrowness of vision may account for a peculiar divergence of practice among those who design and implement surveys with respect to the use of proxy respondents. For those who focus on one source of error—nonresponse—it is taken for granted that proxy respondents should be sought when sampled individuals are unable or unwilling to be interviewed. For those who focus on a second source of error—inaccuracies in reports of respondents—it is generally assumed that respondents tend to provide less accurate answers about others than about themselves. Clearly, a broader perspective is appropriate in assessing the usefulness of seeking data from proxy respondents: one that encompasses at least the two sources of error already mentioned. Not so clear, however, is how to design a study that would allow those two sources of error to be assessed simultaneously, and so permit an assessment of total survey error under different policies with respect to the acceptance of proxy respondents.

In existing surveys that accept proxy respondents, the primary reason for doing so is to achieve acceptably high response rates at relatively low cost. The use of proxy respondents is especially attractive if it is expected that a substantial proportion of the target population is unable to participate because of death (for example, retrospective epidemiologic studies using samples drawn from death registries), age (for example, studies of young children), health problems, or cognitive deficits. Moreover, the accuracy of responses from persons with cognitive deficits may be questionable even if they do participate. For broader population surveys, the acceptance of proxy respondents reduces the need for repeated and expensive callbacks. For persons living in households, the proxy is almost always someone in the same household—preferably the spouse; whereas for those in institutions it is often a caregiver or a grown son or daughter.

Some of the reasons for going to proxy reporters are particularly important with respect to the age group that is growing the fastest and that accounts for a disproportionate amount of health needs and health care expenditures: the oldest old, generally defined as those age 85 and older. For example, concern about response rates among elderly persons surfaced during pilot testing for the Health and Nutrition Examination Survey III: only 57 percent of 46 women age 80 or older were successfully interviewed and only 35 percent of these could be given the physical examination, compared to response rates of 86 percent and 76 percent, respectively, for these two types of data collection from women under 70. The low response rates for the oldest old and the underlying health problems are reflected in the higher proportion of designated persons in this age group for whom proxy interviews are obtained. For example, in the 1984 Supplement on Aging to the National Health Interview Survey of people over 70 years of age, 8.5 percent of the interviews were with proxies, but 26.6 percent for those age 85 and older (Fitti & Kovar, 1987). This increase in reliance on proxies for the oldest old parallels an increase of those having difficulty or receiving help or both with activities of daily living and those experiencing cognitive impairments (Havlik & associates, 1987; Cornoni-Huntley & associates, 1986). Similarly, a high proportion of those who are institutionalized have been reported to be unable to answer survey questions. The fact

Willard L. Rodgers and A. Regula Herzog are with the Institute for Social Research and the Institute of Gerontology, The University of Michigan, Ann Arbor, Michigan.

Research described in this paper was supported by Grant No. R01 AG02038 from the National Institute on Aging. Special thanks are due to Lynn Dielman for her able assistance in all aspects of this project, and in particular for the collection of the validation data.

that over 20 percent of the oldest old are institutionalized (mostly in nursing homes) makes this a major source of concern for this age group.

The reluctance of many survey researchers to accept the use of proxies is based on the assumption that errors due to inaccurate responding are more frequent in proxy than in self-reports. In particular, it is generally assumed that many types of information, especially attitudes and other subjective states, can only be provided by the targeted individuals. As Moore (1988) points out in a recent review, however, evidence on the truth of this assumption is sparse. In many studies it is impossible to separate the effects of proxy reporting from self-selection in the distinction between self-reporters and proxy reporters; and when differences are found, rarely are outside criteria available by which to distinguish which (if either) type of report is the more accurate.

This paper examines differences between respondents and nonrespondents, and differences between proxy reports and self-reports. The design of the study from which the data are taken allows us to compare self-reports and proxy reports for a set of items, albeit only for a rather small number of respondents: those who are married and living in two-respondent households, for whom both proxy reports and self-reports are available on a set of characteristics. Moreover, for a subset of those characteristics information is available from another source, allowing both self-reports and proxy reports to be compared to a criterion.

Methods

The data are from the Study of Michigan Generations, a methodological study funded by the National Institute on Aging. The primary objective of this study is to assess the prevalence and the consequences of several types of nonsampling measurement error, with a focus on the elderly population. The design of the study is described in more detail elsewhere (Rodgers & Herzog, 1987a). In brief, the target population consisted of all adults living in households in the Detroit metropolitan area in 1984. Sampled households were screened to determine whether any member was age 60 or older; the remaining households were selected at a lower rate. Members of the selected households were randomly designated as respondents, and interviews lasting an average of about 90 minutes were obtained with a total of 1,491 respondents, 1,016 of whom were age 60 or older. Interviews were again sought in 1987 with all of the originally sampled persons, and were obtained from 943 respondents, 604 of whom were age 63 or older (that is, had been 60 or older in 1984).

In addition to the survey data, additional information was obtained in both 1984 and 1987 from a variety of sources: administrative data (for example, voting records for several recent elections), census data for the blocks on which sampled households were located, distances to neighborhood facilities measured from maps, and interviewer observations about the households and designated respondents.

A final feature of the design that is central to this paper is that if the randomly designated respondent in a household was married (or living with someone), the spouse was also designated as a second respondent in that household. Moreover, the married respondents were asked a set of questions about their spouses, parallel to questions about themselves, so that if both members of a married couple completed interviews, self-reports and proxy reports are available for these characteristics.

Results

Nonresponse as a Source of Measurement Error

The focus of this paper is on data available from married respondents, because only for these respondents were proxy reported as well as self-reported data requested, thereby permitting comparisons of respondents and nonrespondents and assessments of the quality of the self-reports and proxy reports that would not be possible for nonmarried respondents. A total of 1,496 persons who were married (or quasimarried) and living with their spouses were selected as eligible respondents. That is, given the sample design, 748 couples were selected. By 1987, 97 of those couples were found to be no longer intact due to dissolution of the marriage or the death of one or both spouses. Another 29 of the couples could not be located in 1987, so whether they remained intact could not be determined. The remaining 622 couples were found to remain intact in 1987.

The response rates of these couples in 1984 and 1987 are shown in Figure 1. Each year, in somewhat less than half of the eligible households, both members of the selected couple participated in the study; these will be referred to as "two-respondent households." In another

Figure 1. Response rates by married couples

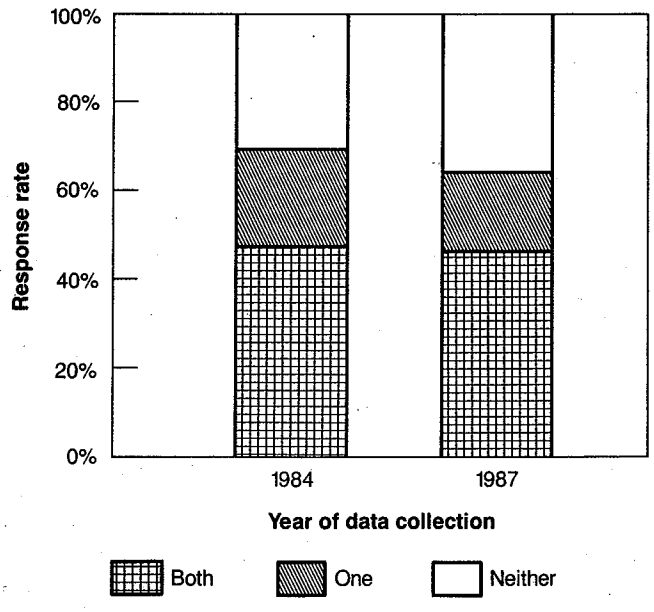
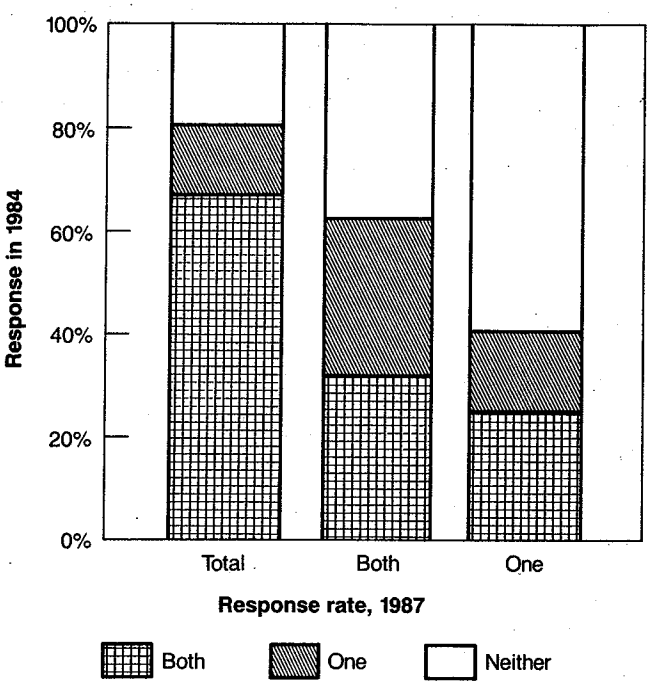


Figure 2. Reinterview response rates



fifth or so of the households just one member of the couple participated ("one-respondent households"), and in the remaining third or so of the households neither person was interviewed ("zero-respondent households"). Stated from the perspective of the individual respondent, in 1984 self-responses were obtained from 58 percent of those selected, proxy responses from another 11 percent, and neither type of response from 31 percent. In 1987, self-responses were obtained from 55 percent of those who remained in intact marriages, proxy responses from another 9 percent, and neither type of response from 36 percent.

Another perspective on the response rates across the two interviews is provided by Figure 2, which shows the 1987 response rate for three groups of couples defined by their responses in 1984. There is some stability in the response pattern: 67 percent of the two-respondent households in 1984 were again two-respondent households in 1987, compared to 32 and 25 percent of the 1984 one-respondent and zero-respondent households, respectively. The stability is by no means perfect, however: On the one hand, cooperation at the first data collection does not assure cooperation at the second, but on the other hand, some of those who refused the initial interview did participate at the second wave.

Differences Between One-respondent and Two-respondent Households. The importance of nonresponse as a source of measurement error is conditional on (1) the frequency of nonresponse, and (2) whether nonrespondents differ from respondents on characteristics that are related to substantive variables of importance to the study. The frequency of nonresponse in this study is clearly high and therefore a potential source of measurement error: more than 40 percent of the designated married individuals did not grant interviews.

To get at the issue of whether nonrespondents differ from respondents, reports were used from married respondents about characteristics of their households and of their spouses, and comparisons were made between one-respondent and two-respondent households: that is, between spouses who were respondents and those who were nonrespondents. It is not obvious, of course, that the nonrespondents for whom proxy reports were obtained adequately represent the total set of married nonrespondents. It seems likely that observed differences would underestimate the differences between respondents and all nonrespondents, since one-respondent households are probably intermediate in most respects between two-respondent and zero-respondent households.

Comparisons of reports about one-respondent and two-respondent households were made for a total of 37 items from the 1984 interview and 34 from the 1987 interview. Each year, statistically significant differences ($p < 0.05$) were found for 10 items, but only 3 items showed significant differences in both years. The variables for which there were statistically significant differences in at least 1 of the 2 years are shown in Table 1.

In both years, one-respondent households had incomes more than $7,000 lower, on the average, than did two-respondent households. The respondents were asked about whether any income was obtained from a total of 16 different sources, and some of these sources were reported by significantly different proportions of the one-respondent and two-respondent households. As is seen in Table 1, for only one of these sources is the difference significant in both years, so caution is advisable. However, in some cases differences that reach significance in one of the years are, while not significant at the 0.05 level, nevertheless in the same direction the other year. One-respondent households are apparently more likely than two-respondent households to receive income from Social Security and welfare, whereas two-respondent households are more likely to receive income from both the husbands' and the wives' jobs and from rent and other investments. Those in one-respondent households were less likely than those in two-respondent households to live in single-family houses, and the houses or apartments in which the one-respondent households lived were smaller, with fewer total rooms and fewer bathrooms. One-respondent households had moved into their current residence an average of 3 years before the two-respondent households.

There is no indication that the nonrespondent biases with respect to these household characteristics are any worse for older than for younger persons. On the contrary, the differences on the variables that were examined are generally smaller, and statistically nonsignificant, for those age 60 and older.

Differences in Reports by Spouses about Respondents and Nonrespondents. In addition to the questions about household and neighborhood characteristics that apply to all members of a household (though often, perhaps, perceived and evaluated differently by each person), married respondents were also asked a set of ques-

Table 1. Household and neighborhood characteristics as reported by married respondents in one-respondent and two-respondent households

Item	1984			1987		
	Two-respondent		One-respondent	Two-respondent		One-respondent
Family income	$33,949	*	$27,119	$41,079	*	$33,513
Percent reporting income from:						
Social Security	18.7		26.3	22.0	*	37.4
SSI	1.3		.5	1.8	*	6.7
Husband's job	79.4	*	67.0	75.5		64.9
Wife's job	59.3	*	47.0	53.3	*	31.1
Rent	8.9		5.5	13.3	*	3.2
Investments	56.3		48.1	67.8	*	54.0
Disability	3.6	*	14.0	9.9		9.6
Welfare	5.9		11.9	2.1	*	13.6
Charity	0.2	*	3.1	0.7		.0
Number of rooms	6.09	*	5.63	6.15		5.97
Number of bathrooms	1.69	*	1.50	1.79		1.71
Percent heating with gas	93.8	*	87.5	—		—
Date moved	1974.2	*	1971.2	1974.8		1971.8
Percent of neighbors with incomes less than $10,000	17.5		20.9	15.9	*	26.4
Percent living in single family home	83.6	*	72.2	88.0	*	74.1
Block: percent with all single family homes	72.6		66.1	79.0	*	64.2

* Difference is statistically significant ($p < 0.05$).

tions specifically about their spouses. These items included a factual item (date of birth), the frequency of certain activities, evaluations of their overall and functional health, and their perceptions of how satisfied the spouse was with three domains of life. Statistically significant ($p < 0.05$) differences were found on 6 of 20 items in 1984, and on 6 of 22 items in 1987. Those variables are displayed in Table 2.

The first entry in Table 2 is the sex of the respondents and nonrespondents, on which one of the largest differences between respondents and nonrespondents is observed: in both years, more than three out of four of the nonrespondents whose spouses did participate were males. The nonrespondents were 4 or 5 years older, on the average, than the respondents.

Married persons were asked to rate the functional health of their spouses in terms of the difficulty the spouse had on four types of activity: seeing things up close, hearing, getting around the house, and remembering things. Nonrespondents were judged by their

Table 2. Characteristics of selected individuals who did and did not respond, as reported by their spouses

Item	1984			1987		
	Respondent		Nonrespondent	Respondent		Nonrespondent
Sex: Percent male**	50.0	*	77.5	50.0	*	78.2
Percent with spouse employed	70.7	*	49.4	66.5	*	33.9
Date born	1940.0	*	1934.7	1939.3		1935.7
Functional health (difficulties, 1–5):						
Hearing	1.41		1.51	1.46	*	1.81
Mobility	1.17	*	1.33	1.21		1.30
Number of religious services in past year	19.21		17.96	20.95	*	13.57
How often depressed (1–5)	2.40	*	2.65	2.43		2.51
Satisfaction with friends (1–7)	5.65	*	5.22	5.70		5.50
Percent who voted in:						
1980	72.3		68.0	77.4	*	57.3
1984	—		—	80.9	*	63.4

* Difference is statistically significant ($p < 0.05$).
** Interviewer report.

spouses to have greater problems with each of these activities in both years, although each year only one difference was statistically significant (mobility in 1984, hearing in 1987).

Respondents were reported by their spouses to attend religious services more regularly than did nonrespondents. Respondents were also reported as depressed less often than were nonrespondents, and more satisfied with their friendships. Finally, respondents voted in national elections more often than did nonrespondents.

These differences in the personal characteristics of respondents and nonrespondents were generally as large, or larger, for older as for younger persons.

Accuracy of Data from Proxy Reporters

The data shown in Tables 1 and 2 indicate that there are differences between respondents and nonrespondents, both at the individual and the household level. These differences are problematic because they call into question the representativeness of the respondents with respect to the population from which the original sample was selected. The use of proxies can reduce the nonresponse rate and thereby reduce concerns about the representativeness of the sample data, but at the same time it raises questions about the accuracy of the data so collected. In this section the accuracy of data collected from proxy respondents is examined.

Statistically significant ($p < 0.05$) differences in the means of the proxy reports and self-reports were found for 8 of the 16 personal characteristics for which proxy reports were obtained in 1984, and for 5 of the 18 asked in 1987. These differences are listed in Table 3, which shows the means for husbands and wives separately since in two cases the proxy effect is significantly different across the sexes. Proxy reports indicate a quarter to a third more visits to physicians in the preceding 12 months than did self-reports. On the other hand, the proxy reports indicate that the respondents visited with friends less often than did the self-reports.

With respect to the functional health of the respondents, the proxy reports indicate that respondents had more difficulty in getting around the house, and less

Table 3. Self-reports and spouse-reports on characteristics of married respondents

Variable	Husband		Wife		Proxy effect	Interaction
	Self-report	Spouse report	Self-report	Spouse report		
Frequency of:						
Doctor visits						
1984	1.50	1.68	2.21	2.50	0.24*	0.06
1987	1.61	2.01	1.96	2.20	.32*	−.08
Visit friends						
1984	4.77	4.35	4.62	4.55	−.24*	.17
1987	4.78	4.49	4.63	4.55	−.19	.10
Functional health (difficulties, 1–5):						
Hearing						
1984	1.51	1.59	1.30	1.25	.01	−.07
1987	1.40	1.64	1.27	1.23	.10	−.13*
Mobility						
1984	1.10	1.15	1.11	1.20	.07*	.02
1987	1.10	1.17	1.14	1.22	.07	.01
Memory						
1984	1.67	1.72	1.83	1.49	−.15*	−.20*
1987	1.85	1.88	1.72	1.58	−.06	−.08
Satisfaction with(1–7):						
House						
1984	5.73	5.46	5.50	5.22	−.27*	−.00
1987	5.69	5.52	5.51	5.16	−.26*	−.10
Life						
1984	5.92	5.41	5.87	5.57	−.41*	.11
1987	5.45	5.29	5.73	5.42	−.24*	−.08
Friendships						
1984	5.96	5.65	6.10	5.71	−.35*	−.04
1987	5.86	5.56	6.18	5.72	−.38*	−.08
How often depressed(1–5)**						
1984	1.95	2.39	2.19	2.41	.34*	−.11
1987	1.89	2.40	2.08	2.46	.44*	−.06

* Statistically significant ($p < 0.05$).
** Item scoring reversed

difficulty in remembering things, than did the self-reports. (The latter difference holds only for women; wives' reports on their husbands' memory indicate at least as frequent problems as do the husbands' self-reports.)

It is with respect to evaluative items that we might be most reluctant to accept proxy reports as a substitute for self-reports, and Table 3 indicates that there are significant differences between proxy reports and self-reports on satisfaction with each of three domains of life. In each case, the spouses report lower average levels of satisfaction than do the respondents in describing themselves. Moreover, respondents report that their spouses are more frequently depressed than do those spouses.

Quality of Spouse Data: Validity Estimates. Table 3 shows that there are statistically significant, and substantively important, differences in the distributions of self-reports and proxy reports on a large proportion of the variables examined. This indicates that the decision to use proxy data should not be made casually, but by itself does not tell us whether the proxy reports are either worse or better than self-reports. Moreover, the data in Table 3 tell us nothing about the impact of proxy data on measures of bivariate or multivariate relationships. Even if there were a systematic bias in proxy reports, this might have no effect on measures of association; indeed, if there were less random error in proxy reports than in self-reports, measures of association based on proxy data could be more accurate than measures based on self-reports.

One evaluation of the relative validities of self-reports and proxy reports is a simple comparison of the correlations of each type of measure with outside sources of information about the same characteristics. For example, the interviewers made observations at the end of each interview with respect to the functional health of the respondents: apparent difficulties with vision, hearing, mobility about the house, and memory. The respondents were also given a standard vision test, and toward the end of the interview were asked a series of questions about that interview as a test of their memory. Whether the respondents voted in several recent elections was ascertained from local voting records. Date of birth was obtained from drivers' license and voting records, although those dates were originally supplied to the agencies by the respondents and so may contain errors that are correlated with those in their interview reports.

The correlations of these external measures with self-reports and with proxy reports are shown in Table 4 (1984 data only). Four of the 11 correlations with the self-reports are significantly higher than the corresponding correlations with the proxy reports. These are with respect to reports of mobility, date of birth, and whether the person voted in one of three elections. On the other hand, the proxy reports correlate at least as highly with scores on the vision and memory tests as do the self-reports, and the average correlation across the 11 comparisons is only faintly lower for the proxy reports than for the self-reports.

The pattern observed for the entire sample is replicated quite closely for each of two age groups: those under age 60, and those age 60 and older.

Stability of Self-reports and Proxy Reports. As another indicator of the relative quality of self-reports and proxy reports, the stability of these reports between the

Table 4. Correlations of self-reports and proxy reports with external records or interviewer reports, 1984

Variable	Total		Under Age 60		Age 60 and over	
	Self-report	Proxy report	Self-report	Proxy report	Self-report	Proxy report
Functional health						
Mobility						
Interviewer	0.690 *	0.601	0.769 *	0.688	0.645 *	0.527
Memory						
Test	.118	.139	.033	.051	.146	.125
Interviewer	.169	.133	.122	.065	.174	.165
Vision						
Test	.304	.305	.358	.372	.381	.365
Interviewer	.289	.264	.224	.207	.386	.339
Hearing						
Interviewer	.288	.358	.199	.226	.337	.390
Birthdate						
Voting records	.999 *	.991	.999 *	.981	.986 *	.910
Driver's license	.998 *	.990	.999 *	.984	.950 *	.893
Voting (versus records):						
1980	.674	.653	.693	.668	.509	.527
1982	.640 *	.581	.642 *	.576	.540	.518
1983	.332	.346	.339	.355	.264	.286
Average	.500	.488	.489	.470	.483	.459

*Difference is statistically significant ($p < 0.05$).

Table 5. Cross-time correlations of self-reports and proxy reports

	N	Self-report	Proxy report	χ^2
Date born	371	0.999	0.998	0.67
Whether employed	373	.584	.563	.09
Frequency of:				
Doctor visits	377	.323	.244	1.39
Visiting friends	377	.570	.416	8.08*
Religious services	383	.780	.742	1.09
Rate health	383	.585	.584	2.79
How often depressed	350	.353	.328	1.38
Functional health (difficulties):				
Vision	385	.382	.334	2.05
Hearing	384	.571	.644	1.59
Mobility	386	.576	.436	10.36*
Memory	385	.386	.554	4.67*
Satisfaction with:				
House	359	.357	.431	.36
Life	333	.482	.384	5.04*
Friends	329	.339	.277	.20
Vote in:				
1980	375	.723	.810	.42
1982	330	.599	.648	.30
Average		.538	.524	

* Statistically significant ($p < 0.05$)

1984 and 1987 interviews is considered. This analysis is necessarily restricted to the 192 couples, both of whom participated in each data collection.

The cross-time correlations are shown in Table 5. Across the 16 characteristics asked of both respondents in both years, there were 4 on which the stabilities of self-reports and proxy reports were significantly different ($p < 0.05$). Self-reports of the frequency of visiting friends were more stable than proxy reports. Similarly, self-reports of difficulty in getting around (mobility) and satisfaction with life were more stable than the corresponding proxy reports. Proxy reports of memory problems, on the other hand, were more stable than self-reports. More impressive than the four statistically significant differences is the overall similarity of the two sets of stability coefficients. The average value of the stability coefficients for self-reports, at 0.538, is only slightly higher than the average for the proxy reports, at 0.524.

In estimating the stability of these self-reports and proxy reports separately for those age 60 and older, the balance tips slightly in the other direction: proxy reports are somewhat more stable, on the average, than are the self-reports of the elderly. The average stability coefficient is 0.509 for the 16 self-reports by those age 60 and older, compared to an average value of 0.527 for the proxy reports on them. For those under age 60 the stability coefficients of the self-reports and proxy reports are 0.535 and 0.504, respectively.

In assessing these stability coefficients as indicators of the quality of the measures, it should be remembered that stability is affected by real change in the characteristics as well as by measurement errors in the reports. For the preliminary inspection of the data shown in Table 5 it is assumed that the measurement errors are random, and in particular that they are uncorrelated over time.

Conclusion

The data from the Study of Michigan Generations, as described in this and previous papers (Rodgers & Herzog, 1987a, 1987b), are not overly reassuring with respect to the quality of survey data. Those papers report frequent, and often substantial, discrepancies between survey reports and data from external sources (that is, administrative records, maps, and so forth) on a variety of variables, including whether the respondents voted in recent elections, the assessed value of their homes, and distances to neighborhood facilities (for example, the nearest drug store). Moreover, these discrepancies do not appear to be random, as is often assumed in the absence of external criteria against which to assess the self-reports. On most of the variables examined, there is a significant bias in the self-reports relative to the external records: for example, respondents tended to overreport voting in each of three recent elections and the amount of property taxes they paid in the preceding year. Even more serious for many types of analysis is the finding that the discrepancies are not independent of other substantive variables that are typically included in multivariate analyses, indicating that sample estimates of bivariate and multivariate parameters derived from survey data are often subject to bias.

As confirmed in the present paper, differences between nonrespondents and respondents are an impor-

tant source of measurement error. Moreover, there are frequent, and often sizable, discrepancies between survey measures and outside criteria, and as seen in this paper, between self-reports and proxy reports. These discrepancies, moreover, are not randomly distributed, and often may introduce biases into univariate distributions and into measures of bivariate and multivariate relationships. While this message is not new, it is an important one, reminding us of the importance of renewing our efforts to improve our methods of data collection and to increase our understanding of measurement errors, where we cannot eliminate them, so that they can be properly taken into account in the estimation of population parameters.

One method to reduce errors introduced by nonresponse is to use proxy respondents. Our evaluation of data from the Study of Michigan Generations suggests that responses of married persons about their spouses are about as valid as their responses about themselves, even with respect to subjective and evaluative questions that are generally considered out of bounds for proxy interviews. There seem to be some biases in proxy reports relative to self-reports, but it should be possible to compensate for such biases to permit analysis of data from a combination of self-interviews and proxy interviews. The design of this particular study limits the extent to which we can generalize: only spouses were considered as potential proxies, and self-reports and proxy reports can be compared only for those able and willing to be interviewed. With these caveats in mind, it is concluded that the use of proxy reporters should be given serious consideration both to obtain at least some information about nonrespondents and as an additional source of information that is useful in assessing and enhancing the quality of self-reported data.

Proxy respondents may be particularly useful in enhancing the quality of data about elderly populations. The reasons for this include the lower response rate of the elderly (increasing the potential for bias) and the fact that reasons for nonresponse by the elderly may be directly related to substantive objectives of the survey. Moreover, to the extent that the quality of data collected from respondents is lower for the elderly than for younger respondents (which seems quite likely at least for subgroups such as the oldest old group and those with physical health problems or cognitive impairments), this would also enhance the attractiveness of proxy respondents.

References

Cornoni-Huntley, J., Brock, D. B., Ostfeld, A. M., & associates. (Eds.). (1986). *Established populations for epidemiologic studies of the elderly.* (NIH Publication No. 86-2443). Bethesda, MD: National Institute on Aging.

Fitti, J., & Kovar, M. G. (1987). The supplement on aging to the 1984 National Health Interview Survey. *Vital and Health Statistics.* (Series 1, No. 21. DHHS Pub. No. (PHS) 87-1323). Washington: U.S. Government Printing Office.

Havlik, R. J., Liu, B. M., Kovar, M. G., & associates (1987). Health statistics on older persons, United States, 1986. *Vital and Health Statistics.* (p. 22). (Series 3, No. 25. DHHS Pub. No. (PHS) 87-1409). Washington, DC: U.S. Government Printing Office.

Moore, J. C. (1988). Self/proxy response status and survey response quality: A review of the literature. *Journal of Official Statistics,* 4, 155-172.

Rodgers, W. L., & Herzog, A. R. (1987a). Interviewing older adults: The accuracy of factual information. *Journal of Gerontology,* 42, 387-394.

Rodgers, W. L., & Herzog, A. R. (1987b). Covariances of measurement errors in survey responses. *Journal of Official Statistics,* 3, 403-418.

Rodgers, W. L., Herzog, A. R., & Andrews, F. M. (1988). Interviewing older adults: The validity of self-reports of satisfaction. *Psychology and Aging,* 3, 264-272.

Surveying Older Adults

Graham Kalton

Introduction

The increasing interest in the health status and health care of older adults, and in their social and economic conditions, has led to a rapid growth in surveys of the older population in recent years. This growth in survey activity has generated an increased concern for the methodology of such surveys. These five papers make valuable contributions to the developing literature on how to conduct surveys of older adults.

The past half century has seen major advances in virtually all aspects of the methods for conducting surveys of the general population. The survey methodology literature now contains reports of numerous studies on almost every component of the survey process. Reports abound on such topics as sample design, question wording and questionnaire design, interviewer effects, coder effects, mode of data collection, nonresponse and noncoverage, the effects of self-reports versus proxy reports, the effects of appointments and incentives, and conditioning and attrition effects in panel surveys. While much remains to be learned, there is a substantial body of knowledge about the survey process for general population surveys. When interest centers on older adults, the question arises as to whether the findings from the many methodological studies of the general population apply equally to this subset of the population.

A variety of reasons have been advanced for why older adults may be more difficult to survey than the general population. These include concerns that older people may have lower levels of comprehension and concentration; that they may be less willing to interact with strangers (interviewers); that they may be more fearful of crime and hence of strangers; that, because of hearing loss, they may be reluctant to participate in telephone interviews; that many may be unwilling to participate in a survey because of ill-health; and that other household members may act as gatekeepers who protect older adults from being interviewed. It is easy to generate special concerns about almost any segment of the population, but such concerns often turn out to be groundless or of little significance. Empirical studies, like those reported here, are needed to establish whether there is a real basis for a particular concern about any special section of the population.

In assessing methodological concerns for a survey of older adults, an important issue is how an "older adult" is defined. If an equal probability sample is selected for a survey of the older population aged 55 and over, about 2 out of 5 members of the sample will be aged between 55 and 64, and only about 1 in 5 will be 75 or older. If, however, the survey is concerned with the elderly aged 65 and over, about 2 out of 5 members of the sample will be 75 or older. Both the size of the older population and its composition depend on the definition of "older adult" adopted. Since the methodological concerns in surveys of older adults relate most forcefully to the oldest old, the age composition of the survey population is an important consideration.

The following sections discuss some specific issues raised by these five papers, namely sampling, total and item nonresponse, self-reports versus proxy reports, and panel surveys of older adults. The discussion ends with some concluding remarks.

Sampling Older Adults

As with any survey, the initial step in the sampling process for a survey of older adults is to define the survey's target population. The definition should be based on the survey objectives and the statistical inferences that are to be made. Age limits and geographical boundaries clearly need to be specified. For a general survey of older adults, a decision also needs to be made on whether the population is to be restricted to the nonin-

Graham Kalton is with the Department of Biostatistics and the Survey Research Center, University of Michigan, Ann Arbor.

The preparation of this paper was supported in part by National Institute for Aging Grant No. P01AG05561.

stitutionalized population or whether older people in institutions, such as nursing homes, are to be included. Since, as Rodgers and Herzog note, over 20 percent of the population aged 85 and over are in institutions, the decision of whether to include the institutionalized population is a particularly significant one for surveys of the oldest old. The inclusion of the institutionalized population makes the sampling procedures more complex and can also cause significant problems in data collection. However, if the sample is restricted to the noninstitutionalized population, the survey estimates relate only to that population. Since institutionalized older adults are likely to be in poorer health than the noninstitutionalized, estimates from samples of the noninstitutionalized population will tend to understate the health problems of the population of all older adults.

Sampling noninstitutionalized older adults is an example of what is termed "sampling a rare population." A variety of methods have been developed for sampling rare populations in an efficient manner (Kalton & Anderson, 1986; Sudman & Kalton, 1986; Sudman & others, 1988). One obvious approach is to seek a list of members of the rare population or a list on which the members of the rare population can be identified. Kingery gives an example of this approach using the voter registration lists as a frame for sampling older adults. However, as she notes, the voter registration lists are incomplete, with the consequent risk of noncoverage bias. In general, before a list is used as a sampling frame, a careful assessment needs to be made of the noncoverage of the list, of the extent to which members of the rare population are correctly identified on the list, and of the extent to which sampled individuals can be contacted from location information provided on the list. If access can be gained to the Medicare lists (as was the case for the survey reported by Thomas), they can be an attractive frame for sampling the population aged 65 and over. Their coverage appears to be high, and they contain address information that is generally accurate and up to date.

When no adequate list is available, a common method of sampling a rare population is to conduct a large-scale screening of the general population by some economic procedure, and to survey the members of the rare population identified in the screening. The screening may be done in a large-scale survey, as is the case with the Supplement on Aging survey that was added to the National Health Interview Survey, described by Kovar. The applicability of screening depends on its cost, its effectiveness in correctly identifying members of the rare population, and the degree of rarity of that population. The extent of screening required rises rapidly as the rarity increases. Thus, a screening sample of about 4,700 persons would be required to produce a sample of 1,000 persons aged 55 and over (with about 21 percent of the population being aged 55 and over), whereas a screening sample of about 20,000 persons would be needed to produce the same size sample of persons aged 75 and over (with about 5 percent of the population being aged 75 and over).

As Kingery describes, screening can be readily performed as part of a random digit dialing (RDD) telephone survey. In most countries, telephone household noncoverage increases when the head of the household is 65 or older (Trewin & Lee, 1988), giving rise to concerns about noncoverage bias. However, this situation does not pertain in the United States, where the noncoverage is lowest among households with heads aged 65 or more. Thornberry and Massey (1988) report that in 1985 to 1986 only about 3.2 percent of persons aged 65 and older lived in nontelephone households. While there are substantial differences between persons aged 65 and over in telephone households and those in nontelephone households, in terms of such factors as having private health insurance and receiving public assistance, the extremely low proportion of persons aged 65 and over in nontelephone households means that the noncoverage bias is not great.

Disproportionate stratified sampling can sometimes be used to sample a rare population in an efficient manner. The technique involves sampling at higher rates those strata identified as having high concentrations of the rare population. The gains from disproportionate sampling are generally minor unless two conditions apply: The strata to be oversampled need high concentrations of the rare population and they must contain a substantial proportion of the rare population (Kalton & Anderson, 1986). These conditions will not generally hold for sampling older adults in geographically defined strata, and hence disproportionate stratification will be of little benefit for this situation. An alternative is to target the sample to only those areal strata with higher concentrations of older adults. As Kingery notes, this cut off method increases the productivity of the screening, but at the severe cost of restricting the population of inference to the sampled strata.

Nonresponse

There is a concern that total nonresponse rates are higher among older adults than among the general population, and in general population surveys there is evidence that the concern is justified (Herzog & Rodgers, 1988; Herzog & Kulka, 1989). In face-to-face interview surveys, response rates are often lower for older adults, but are only appreciably lower for those aged 75 and over, and markedly lower only for those aged 85 and over. With telephone surveys, the decline in response rate with age is much more substantial. Various explanations for the low response rate by older adults to telephone surveys have been suggested, but there is little concrete evidence available to support or refute them (Groves & Lyberg, 1988). With the increasing use of telephone surveys, research to discover the causes of the low response rates of older adults is urgently needed. The information Kovar provides on the most productive times to call older respondents is helpful to improve the efficiency of the survey and to reduce nonresponse.

Although it appears that older adults are less cooperative than others in general population surveys, this should not be taken to imply that they will always be less cooperative. On the contrary, Hoinville (1983) suggests that the elderly are very cooperative in surveys that are

specifically about them and of direct interest to them. He provides several examples in which high levels of cooperation were secured in surveys among the elderly.

Concerns about higher levels of nonresponse among older adults apply also to item nonresponse. Item nonresponse rates may be hypothesized to be higher among older adults for a variety of reasons, including fatigue with long interviews, difficulties reading show cards, and less ability to remember information. It seems likely that the importance of such factors depends heavily on what population of older adults is being surveyed. If the population is defined as the noninstitutional population aged 55 and over, the major factors affecting item nonresponse are likely to be quite different from those applying with a noninstitutional population aged 80 and over or with a nursing home population.

In their nursing home study, Garrard and others not surprisingly found that elderly residents with severe cognitive impairment are more likely to fail to provide answers to survey questions. Even when such residents do respond, the validity of their responses needs to be considered carefully. In designing a study of a nursing home population in which significant cognitive impairment is common, careful attention clearly needs to be paid to what information can be validly collected from the residents, the form in which it is collected, and whether some information might be better collected from physicians or the nursing staff.

Proxy Informants

As Rodgers and Herzog point out, one way to reduce nonresponse is to accept responses from proxy informants. The acceptance of proxy informants, however, introduces the risk that they may provide less accurate responses. Although this issue has received a good deal of attention in the survey literature, no clear-cut picture has emerged. In part this situation reflects the difficulty of conducting rigorous tests of the hypothesis that there is a difference in the accuracy of self-reporting and proxy reports. A number of studies have simply examined the differences between the survey results obtained from self-respondents and those obtained from proxy informants. However, any such differences observed cannot be safely interpreted as the result of differences between self-reporting and proxy reporting. They may instead just reflect the self-selection of sampled persons to respond for themselves or to have informants respond for them. Sound methodological studies of self versus proxy effects are few in number and they do not provide convincing evidence of quality differences between self-respondents and proxy informants. Based on his literature review of this subject, Moore (1988) concludes that there is "little support for the notion that self-response survey reports are of generally better quality than proxy reports."

Rodgers and Herzog's paper compares self-reports and proxy reports, where the proxy is the spouse of the self-reporter and where both husband and wife provide data on each other. By comparing the results of husbands and wives reporting for themselves with the results when they report for each other, Rodgers and Herzog identify some appreciable differences between self and proxy reports. Moreover, by making comparisons with other data, they are able to show that proxy reports are somewhat less reliable in some cases. The findings from this study are an important addition to the literature on self-reporting and proxy reporting, but the limitations of the study need to be recognized. The results relate only to the variables measured in this survey, to the use of a spouse as a proxy informant, and to cases in which both husband and wife collaborated in the study. Further research is needed to examine the quality of proxy reporting when the proxy informant is someone other than the spouse (for example, a child or a nurse), and when the sampled person is unwilling or unable to respond for himself or herself.

Panel Surveys

Four of the five papers presented here include a panel survey component in which a sample of older adults was interviewed on more than one occasion. This concern with panel surveys follows a general tendency to incorporate a time dimension in population surveys and is a natural outcome in light of the recent interest in the dynamics of aging.

The key advantages of shifting from a cross-sectional to a panel survey design are the abilities to analyze gross changes (that is, changes at the individual level), the durations of events (for example, spells of illness or lengths of stay in hospital), and the relationships between variables measured at different points of time. These advantages are gained, however, at the price of the added complexity of a panel survey and additional nonsampling error concerns that arise in panel surveys, especially panel attrition and conditioning effects.

One aspect of the complexity of panel surveys is that not only do individuals' values on the variables of interest change over time (the raison d'etre for panel surveys), but the population changes over time. New entrants may occur because of "births" (passing, say, age 55 in a survey of older adults), immigration from abroad, and persons returning from institutions (for a panel survey of the noninstitutional population). Leavers may occur because of deaths, emigration, and persons entering institutions. Furthermore, some individuals may move into and out of a population one or more times over the course of a panel survey; for instance, at the start of the survey a person may have been in a household, that person may enter a nursing home at a later point, may subsequently return to the household, and then die sometime later.

These population dynamics need to be carefully considered both in the design of a panel survey and in its analysis. The issue in the analysis is what is the population of inference for the survey results. Some analyses from panel surveys are cross-sectional in nature, producing estimates from a single wave of the panel. When there are significant numbers of new entrants to the survey population since the start of the panel, the panel sample may need to be replenished with a sample of new

entrants to produce estimates for the desired population of inference. With longitudinal analysis, the definition of the longitudinal population needs to be addressed (Goldstein, 1979). Often the longitudinal population may be defined as those who remain in the population throughout the period of interest, but this definition is not always appropriate.

The sample for a panel survey may change over time to take account of changes in the population, for example, by adding samples of new entrants and through losses from death. It will also change because of nonresponse. A clear distinction needs to be made between these two sources of change because they have different implications for analysis. Thus, for example, losses from death lead to questions about how the population of inference should be defined, whereas losses from nonresponse attrition give rise to concerns about nonresponse bias and what type of nonresponse adjustments might be employed.

In practice, it is not always possible to determine whether those sample losses that occur because the individuals could not be traced are in fact nonrespondents or leavers from the population. In panel surveys of older adults, leaving the population occurs often through death. As Kovar describes, the National Death Index can be an effective tool for identifying panel members who have died.

Attrition of a panel sample through nonresponse is a serious concern, but fortunately experience suggests that the problem need not be as serious as might be feared. In panel surveys of the general population, nonresponse rates do generally increase from one wave to the next, but with well-conducted surveys the increases are often relatively modest (Kalton & others, 1989). The very low nonresponse attrition rates reported by Thomas for persons aged 65 and over in New York City suggest that this general finding also applies for older adults.

Although the wave-to-wave increases in nonresponse rates may be small, the overall nonresponse rate after a number of waves may nevertheless become large. There is, however, a mitigating factor at work. A great deal is known about those who leave a panel after the first wave of data collection from their questionnaire responses in the waves in which they did participate. This information can be used to compare respondents and nonrespondents, and used in the development of adjustment procedures that aim to reduce any nonresponse bias that exists (Kalton, 1986; Lepkowski, 1989).

Many other methodological issues arise in the design and analysis of panel surveys, but they cannot be treated here. Discussions of these issues are to be found in Kasprzyk and others (1989), for panel surveys in general, and in Lawton and Herzog (1989), for panel studies of older populations.

Concluding Remarks

The papers presented have addressed a variety of issues that arise in surveying older adults. Although they have been discussed separately, there are interrelationships between them. For example, as Rodgers and Herzog point out, decisions about the use of proxy informants should take into account not only the quality of the data reported by the informants but also the increased level of nonresponse that might occur if the survey were confined to self respondents.

The overall strategy for survey design is to minimize total survey error for the resources available. This strategy involves an economic control of errors from all sources, including sampling error, noncoverage bias, nonresponse bias, and response error. The implementation of this approach requires an extensive knowledge of the magnitudes of errors arising from various sources and how they vary with different survey design features. Studies like those reported in these five papers contribute to our understanding of the survey process. As such, they will help those planning the designs of future surveys of older adults.

References

Goldstein, H. (1979). *The design and analysis of longitudinal studies*. London: Academic Press.

Groves, R. M., & Lyberg, L. E. (1988). An overview of nonresponse issues in telephone surveys. In R. Groves, P. P. Biemer, L. E. Lyberg & associates (Eds.). *Telephone survey methodology* (pp. 191-211). New York: Wiley & Sons.

Herzog, A. R., & Kulka, R. A. (1989). Telephone and mail surveys with older populations: A methodological overview. In M. P. Lawton & A. R. Herzog (Eds.) *Special research methods for gerontology* (pp. 63-89). Amityville, NY: Baywood Publishing Company.

Herzog, A. R., & Rodgers, W. L. (1988). Age and response rates to interview sample surveys. *Journal of Gerontology,* 43, S200-205.

Hoinville, G. (1983). Carrying out surveys among the elderly. *Journal of the Market Research Society,* 25, 223-237.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics,* 2, 303-314.

Kalton, G., & Anderson, D. (1986). Sampling rare populations. *Journal of the Royal Statistical Society,* A, 149, 65-82.

Kalton, G., Kasprzyk, D., & McMillen, D. B. (1989). Nonsampling errors in panel surveys. In D. Kasprzyk, G. J. Duncan, G. Kalton & associate (Eds.). *Panel surveys* (pp. 249-270). New York: Wiley & Sons.

Kasprzyk, D., Duncan, G. J., Kalton, G., & associate. (1989). *Panel surveys*. New York: Wiley & Sons.

Lawton, M. P., & Herzog, A. R. (1989). *Special research methods for gerontology.* Amityville, NY: Baywood Publishing Company.

Lepkowski, J. M. (1989). Treatment of wave nonresponse in panel surveys. In D. Kasprzyk, G. J. Duncan, G. Kalton, & associate (Eds.). *Panel surveys* (pp. 348-374). New York: Wiley & Sons.

Moore, J. C. (1988). Self/proxy response status and survey response quality. A review of the literature. *Journal of Official Statistics,* 4, 155-172.

Sudman, S., & Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology,* 12, 401-429.

Sudman, S., Sirken, M. G., & Cowan, C. D. (1988). Sampling rare and elusive populations. *Science,* 240, 991-996.

Thornberry, O. T., & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. Groves, P. P. Biemer, L. E. Lyberg, & associates (Eds.). *Telephone survey methodology* (pp. 25-49). New York: Wiley & Sons.

Trewin, D., & Lee, G. (1988). International comparisons of telephone coverage. In R. M. Groves, P. P. Biemer, L. E. Lyberg, & associates (Eds.). *Telephone survey methodology* (pp. 9-24). New York: Wiley & Sons.

DISCUSSION PAPER

Collecting Data from Samples of Older Adults and Nursing Home Populations

Steven B. Cohen

Introduction

Many of the critical health policy initiatives of the next decade will be directed toward addressing the needs of our elderly population. The recent landmark legislation that provides Catastrophic Health Care benefits to Medicare beneficiaries is symptomatic of the national attention that is directed toward the aged. Many of the national health care surveys that have been conducted in this decade, or are planned for the 1990s, have specially targeted samples of older adults and nursing home populations.

More specifically, national nursing home surveys have recently been conducted by both the National Center for Health Services Research and the National Center for Health Statistics to obtain detailed profiles for nursing home residents on their demographic and health status characteristics, their health insurance coverage, health care utilization, and medical expenditures. Furthermore, many of the household-based national health care surveys that are conducted by the Federal government include questionnaire supplements directly addressed to the elderly. In addition, a number of these national household surveys, such as the National Medical Expenditure Survey, include an oversample of the elderly to improve the precision of survey estimates that characterize this policy-relevant population subgroup.

As the demand grows for data necessary to form health policy initiatives targeted to the elderly, it will be even more critical to ensure the collection of reliable, accurate, high quality data in a timely and cost-efficient manner. Organizations responsible for sponsoring, collecting, and/or analyzing the data will be looking for new advances in the field of survey research that will have an impact in the reduction of total survey error, and for data collection strategies that yield reductions in survey costs without any complementary loss in the quality of the resultant data. Further research in the area of nonresponse bias may result in new data collection approaches to minimize survey nonresponse in surveys of elderly populations, or in improved estimation strategies to correct for nonresponse bias as it impacts on survey estimates of the elderly.

The papers in this section have been primarily directed to the issues of data quality, to components of total survey error, and to alternative sampling techniques as they relate to surveys of older adults and nursing home populations.

Surveying Older Adults

The paper by Kingery was directed toward an examination of the performance of two alternative sampling schemes to survey older adults. In the analyses that she presents, two rather disparate surveys have been compared on the dimensions of efficiency and sample representativeness. The differentials across the two surveys with respect to the targeted populations, the different study sites, the variant sample sizes, and the differing study goals appeared to have a confounding effect on the study comparisons. Nevertheless, the paper provides a good framework in which to understand the advantages and limitations of the sampling strategies employed for each of the surveys in their own right.

For the random digit dialing (RDD) survey, an inefficiency in the sampling technique was noted with respect to identifying the study's eligible population. This result is not unexpected, however, because any randomly selected sample will be affected by additional survey costs associated with screening to identify the target population. One option offered by the author, to increase the efficiency of the RDD sample, requires exclusion of telephone exchanges where the sample representation is low. I would strongly advise against adopting this strategy as a consequence of the serious detrimental effects of resultant undercoverage on survey

Steven B. Cohen is with the National Center for Health Services Research, Rockville, Maryland.

estimates. A more balanced strategy would allow all exchanges a chance of selection in the survey but selectively oversample the exchanges with the higher "hit" rates.

Alternatively, the efficiency of directly identifying a sample of elderly individuals from a voter registration list is clearly illustrated in the paper. This advantage, however, has to be weighed against the potential limitations of the list that serves as the sampling frame. Of primary concern is the representativeness of the sampling frame with respect to the target population. In the example presented, to the extent that a significant portion of the elderly population have not registered to vote, the survey estimates will be affected by a component of survey error due to undercoverage. Even when it can be demonstrated that the set of individuals that are represented in the sampling frame mirror the proportionate representation of the target population for a select set of demographic indices, those individuals not represented on the frame may significantly differ with respect to the critical measures of interest for a given study.

Other concerns that are associated with a list sample relate to the timeliness of the list. As indicated in the paper, a number of the older respondents identified in the list were determined to be deceased at the time of the survey contact. However, unlike the RDD sampling restriction that requires a sample respondent to have a telephone, a list sample that provides both addresses and telephone numbers will allow for a mixed mode data collection strategy and facilitate sample representation of individuals without telephones.

In summary, the paper serves to direct our attention to the competing survey design goals of minimizing total survey error and reducing survey costs. Good survey practice requires a design specification that indicates a population coverage rate that must be achieved by the adopted sampling scheme. Similar design requirements should be specified for expected survey response rates. The most efficient survey design that simultaneously satisfies the survey design requirements should be considered for adoption.

Nonresponse and Attrition

Surveys that are characterized by longitudinal or panel designs are subject to two distinct forms of nonresponse, complete nonresponse and partial nonresponse. Complete nonresponse occurs when an eligible sampling unit does not participate in the study. Alternatively, partial nonresponse occurs in longitudinal designs when initially responding participants provide data for only a part of the time they are eligible to respond. For example, initially responding persons refuse to participate in later rounds of data collection. Furthermore, an inability to locate participants who change residences also results in partial response. As a consequence of the generally higher levels of attrition experienced in surveys of the elderly, there is a serious concern regarding the impact of nonresponse on survey estimates that characterize this population.

Thomas's paper provides a excellent framework to assess the potential impact of both components of nonresponse on survey estimates. The data for the study come from a panel survey of elderly individuals selected from a list of Medicare enrollees, who were contacted every 6 months over a 3.5-year period, to obtain data on their health, health care, and aging. As noted, 73 percent of the eligible sample completed baseline interviews, with an average wave- specific sample diminution of 2 percent due to nonresponse and 3 percent due to death.

What is particularly notable about this study is the inclusion of a recontact sample to represent the initial survey nonrespondents. Even more remarkable is the capacity to obtain required study information for 86 percent of the targeted nonrespondents. The data facilitated a comparison of the characteristics of survey respondents with the nonrespondents on both sociodemographic characteristics and the health care measures of primary analytical importance to the study. When resources permit, most surveys should include such a resurvey component for a representative sample of the nonrespondents. The resultant data will not only facilitate an assessment of the potential level of nonresponse bias in survey estimates, but also serve to reduce nonresponse bias by an improvement in the overall survey response rate.

In Thomas's comparisons of the study respondents and initial nonrespondents, significant differentials were noted across the two study groups on a number of health and sociodemographic characteristics. The nonrespondents were in poorer health, more likely to be women, to live alone, and to have lower annual incomes than their counterparts. In this study, however, the nonrespondents were markedly similar on key analytic measures of hospitalizations, source of medical care, and age. In an attempt to determine whether study estimates would significantly shift with the inclusion of the recontacted nonrespondents, the estimates from the combined sample were compared with the results from only the wave four respondents. As expected, the measures for which no significant differentials were noted between study groups did not shift survey estimates with inclusion of the nonrespondents. In addition, the author notes that for the remaining measures that distinguished the two groups, there were no apparent differentials in estimates of whether the nonrespondents are included. This finding is not typical of many of the health care surveys that are directed at the elderly. Not only do the nonrespondents differ significantly on many of the key analytic measures for a given survey, but their inclusion, when possible, in the derivation of survey estimates will have a significant effect in moving the mean.

Since the author did not go into detail with respect to the statistical test that was considered for this analysis, it is difficult to assess why the measures that distinguished the nonrespondents were not significant here. Perhaps the test statistic that was considered for the comparisons did not take into account the expected correlation in survey estimates that was a consequence of including the respondents in both of the estimates that were compared. If the author had assumed independence in survey estimates for the comparisons, the var-

iance of the estimate of the difference between estimates would have been overstated, resulting in a loss in power to detect significant differences.

An additional analysis was conducted to determine the characteristics of the survey dropouts and the effect of their exclusion on survey estimates. Study results indicated that the voluntary survey dropouts more closely resembled the survey respondents than those participants that died. The analyses focused on wave- specific comparisons which understate the overall impact of attrition over the entire course of a panel study. Given that the overall attrition rate for the survey exceeded 35 percent, I was looking for an additional discussion that also contrasted the initial sample with the resultant sample at the end of the 3.5-year study. Referring back to the results in Table 1, it appeared that the sample remaining at the end of the seventh wave differed noticeably from the wave one sample for a number of key analytic measures. When panel surveys are conducted over a lengthy time period, and the survey suffers significant levels of attrition over time, consideration should be given to the selection of an independent supplemental sample at the end of the study to help benchmark survey estimates. Much like the recontact survey of initial nonrespondents, this independent survey of the same target population at a later date will inform studies of attrition bias and facilitate the application of improved nonresponse adjustment strategies.

Finally, the author adds another dimension to the concerns regarding sample loss due to total nonresponse and sample attrition by focusing on its impact with respect to multivariate analyses. Even when no differentials are noted in the parameter estimates that characterize a survey as a consequence of these potential sources of survey error, they may seriously influence the dynamics of multivariate analyses. Consequently, adjustments for nonresponse bias must be directed toward both the derivation of point estimates and multivariate analyses.

Item Nonresponse from Elderly in Nursing Homes

The paper by Garrard deals with another important component of survey error in data collection efforts targeted to the elderly, that of item nonresponse. Focusing on surveys targeted to the elderly population in nursing homes, it raises a number of important questions regarding the feasibility of conducting interviews directly with the sampled residents. Most surveys that are directed to obtaining data on health-related characteristics of nursing home residents rely primarily on responses from either the nursing home staff, data in institutional records, and interviews with next of kin. Using data from a small- scale survey on quality of care of residents in nursing homes, the author examined levels of item nonresponse for different subgroups of the study population.

Study findings indicated no significant differentials in the item nonresponse profiles for two distinct groups of residents, the new admissions and the long-stay residents, both overall and further distinguished by age, gender, and health status disaggregations. However, this may be a function of the limited statistical power of the small sample that was selected for the study. As expected, a noticeable differential was observed in the resultant levels of item response in the comparisons across resident mental status. The relatively low levels of item nonresponse that characterized the cognitively intact provides support for direct interviews with this subset of elderly nursing home residents. Furthermore, if the primary research query was directed to assessing the capacity of nursing home residents to respond for themselves in a survey of the quality of care, it is unclear what is gained by including the questions that were gathered from nursing home records or obtained by the interviewer.

The other concern I had regarding this investigation was the dependence on the level of item nonresponse as the only measure of quality. More generally, the achievement of high item response rates should be viewed as a necessary but not sufficient requirement to minimize bias in estimates due to nonresponse. Additional protocols must be put in place to ensure the validity of the data that are obtained.

Proxy Respondents for Elderly Populations

Rodgers and Herzog provide a good discussion of the interactive nature of two sources of error encountered in surveys of elderly populations, errors due to nonresponse and to the use of proxy respondents. As data collection efforts are directed to household surveys of elderly populations, there is a reliance on the acceptance of proxy respondents to improve survey response rates. Often the targeted elderly respondent is incapable of directly providing the required data, and this course of action is necessary. However, the authors caution against the acceptance of proxy reports to improve response rates without first assessing the impact of this component of error on survey estimates.

Analyses in this paper examine differentials between respondents and nonrespondents in addition to proxy versus self-reporting, using data on married couples from the Survey of Michigan Generations. It should be noted, however, that the analyses are not strictly limited to the elderly, with approximately one third of the survey respondents under the age of 60. Furthermore, the survey response rate was problematic, with less than half the eligible households meeting the definition of complete respondents.

Study findings revealed that significant differentials exist between the respondents and the nonrespondents on the dimensions of gender, age, and socioeconomic status, which are strong correlates of a significant number of the criterion measures of the survey. Furthermore, the comparisons between self-reports and proxy reports of key study measures revealed significant reporting differentials on the dimensions of health care utilization, functional health, and personal satisfaction. The self-reported data also exhibited greater agreement

patterns with external records in the validity analyses that were conducted.

I was looking for a framework to assess which of the two components of error in this survey is most serious from a total survey-error perspective. Clearly, gains in precision are achieved by enhancing the sample size of the survey through the inclusion of the proxy respondents. It remains unclear, however, as to whether a reduction in the mean square error of survey estimates was achieved. It can be demonstrated that when errors in measurement due to proxy response are dominant, the overall mean square error could be reduced by a survey design that accepted a lower response rate with the inclusion of only self-respondents. Since data from external sources exist on a number of criterion measures obtained from the survey, and were used to inform the validity study, I would suggest their incorporation into an additional analysis of the mean square error that is obtained from these alternative design strategies.

In this study, the authors have allowed the one-respondent households to represent the nonrespondents. It should be noted that relative to the households with either complete or partial nonresponse, the one-respondent households represent less than 40 percent of the total. To the extent that differentials exist between these partial and complete nonrespondents, the study results should be interpreted with caution. If data from external sources were also available for the total nonresponding households, they should have been incorporated into the analyses.

Longitudinal Study of Aging

Kovar nicely complements the self-analysis versus proxy analysis of the previous paper. Using data from the Longitudinal Survey of Aging, the analyses are directed to an examination of the feasibility of conducting a telephone interview for a survey of the elderly. In addition, the author attempts to examine the quality of data obtained from self-respondents versus proxy respondents.

More specifically, the study attempted to determine whether a nationally representative sample of the civilian, noninstitutionalized population aged 70 or older could be recontacted through telephone interviews over a 4-year period. Telephone contacts were scheduled in 2-year intervals. Study findings indicate that the conditional response rate to the first follow-up recontact in 1986 was 92 percent. The results appear to support the claim that a telephone interview, as the primary mode of data collection for a longitudinal survey of the elderly, would yield respectable response rates. Here, I was curious as to what percent of the completed interviews were actually conducted over the phone.

Alternatively, the analysis of self-respondent versus proxy respondent focused on responses to the 1984 Supplement on Aging. It was noted that data for 8.7 percent of the sampled persons aged 65 or older were obtained from proxy respondents. Study findings revealed the persons with proxy responses had a higher representation of individuals aged 85 and older, and who need help or have difficulty with three or more activities of daily living (ADL). Consequently, exclusion of the data from proxy respondents in the derivation of national estimates would have underestimated the number of individuals with ADL difficulties. The results support the author's recommendation to press for a self-respondent but to accept proxy reports when this is unobtainable. I was curious, however, as to why the sampling weights developed for the national estimates derived from the self-only respondents did not incorporate a poststratification adjustment by age.

Data from Medicare files were used to evaluate the quality of data obtained from self-respondents versus proxy respondents. The results indicated that the self-respondents were characterized by a higher level of data quality, as measured by the match rates for the Medicare claims numbers.

Clearly, more research which examines the quality of data obtained from self-respondents versus proxy respondents for surveys of elderly populations is necessary. Ideally, the design considered in the previous paper, which obtained self-responses and proxy responses for the same person, and had access to more accurate administrative record data as a third source, would be a model to consider for future studies of this type.

Summary

By focusing on topics related to sample design, data quality, and components of total survey error, these papers have covered a broad area of survey research directed to surveys of older adults and nursing home populations. The authors are to be complimented for their contributions to improving our understanding of the survey process that affects this special target population.

Obligations Attending Gaining Information: A Moral Question for Health Survey Researchers

Joan C. Callahan

Introduction

Some years ago a physician from a research group approached me to talk about a case with which he and his colleagues had been struggling.[1] They were conducting an information-gathering study on victims of Huntington's Chorea and their offspring. Huntington's Chorea is a hereditary degenerative disease of the nervous system, with symptoms usually appearing between the age of 40 and 60. It is initially characterized by personality changes (such as obstinacy, moodiness, and lack of initiative) followed by increasing involuntary movements, leading through severe personality changes to profound dementia and extreme motor disabilities, including the inability to walk. Four of the subjects in the study were 3 brothers in their 20s and their father. The father, who was experiencing early symptoms, had recently been diagnosed as having the disease. One son was married with one child, one was engaged, and the third was single, but in a long-term relationship with a woman whom he expected he would one day marry. The men had entered the study hoping that their participation might help in the search for criteria allowing early diagnosis of the disease, and ultimately might help in the search for a cure for the disease. Fearing that they might pass on the disease, the sons had all undergone vasectomies.

In the course of doing various tests on these subjects, the researchers discovered that the married son could not have been the offspring of the father. At the time he told me the story, the physician was deeply troubled and completely unsure what his team should do with this information. All of the sons had read virtually all the available literature on the disease, and all knew that the state of medical understanding at the time was such that there were no reliable indicators for establishing whether an offspring of a person afflicted with the disease had inherited it or had the capacity to pass it on. Thus, although telling the young man that he was not at presumed risk would relieve him of considerable anxiety regarding his own medical status and that of his young son, telling would virtually certainly lead to his asking how the researchers could be certain that he was not at presumed risk.

At the direction of the majority of the team members, the physician called the mother and told her about the team's discovery. He asked if she would tell the son that he was not at presumed risk for the disease and why this was the case. She adamantly refused, saying that telling the son that he had been fathered by someone other than the man he believed to be his father would destroy her relationship with her son. Further, she said she was virtually certain that her son would feel compelled to tell her husband, and that this would be irreparably damaging to her relationship with her husband. She went on to explain that this son's conception had been the result of a terrible mistake she had made in a single night's liaison at a time when she was about to initiate divorcing her husband because of an infidelity on his part, shortly after which they had reconciled. She said she had considered aborting the pregnancy, but her religious commitments had prevented her doing so. The pregnancy and birth of this child had subsequently drawn them closer together, she said, and their marriage had been nearly ideal ever since. She said she realized that withholding the information had serious implications for the well-being of her married son, his wife, and her grandson, but that her first concern needed to be for her husband, who was still lucid and would be crushed by the information at the very time when he most needed the comfort and security of their extremely good relationship. Finally, she said that losing her husband to this disease was all that she could manage—that

Joan C. Callahan is with the Department of Philosophy, University of Kentucky, Lexington, Kentucky.

[1] A number of facts about this case have been modified to protect the identities of the researcher and subjects.

she simply would not be able to bear losing her relationship with her son.

I do not know what the researchers in this case finally decided to do, but the physician who told me the story was having great difficulty accepting the view taken by one of his colleagues that their moral obligations as researchers were strictly limited to carefully gathering and accurately reporting the information sought by the study—no more, no less.

Strong Role Differentiation in the Professions

The position taken by the physician's colleague supports what is known as strong role differentiation in a profession, and it is a position that is commonly defended in and for a number of professions. This view holds that professionals acting in their professional roles have clearly circumscribed moral obligations that exempt them from other obligations we might ordinarily think a person has toward other persons or society more generally. In legal practice, for example, the view that attorneys are to be strongly role-differentiated holds that once an attorney takes on a client, his or her moral obligations regarding that client are strictly limited to serving as a zealous advocate of that client's legal interests. Thus, concerns about the interests of other parties or society and even concerns about the nonlegal interests on the client are not to occupy the attorney. It is this position that attorneys appeal to in justifying the zealous representation of clients thought to be the perpetrators of heinous crimes.[2]

Strong role differentiation in the practice of criminal legal representation is argued for on a number of grounds, which include the claim that attorneys are simply not in a position to judge the guilt or innocence of a client. That judgment, it is argued, can only be made after both the prosecution and defense have made their cases as strongly as possible. The defending attorney has one job in that process, the prosecutor another, the judge another, and the jury yet another. If attorneys were to presume to make a judgment on a client's guilt or innocence and conduct a defense according to that judgment, they would have arrogated to themselves a role they are simply not entitled to nor in a position to assume. After all, it is rightly pointed out, clients sometimes misremember facts; and they sometimes lie—not only about their innocence, but about their guilt as well. Thus, it is contended, the values of truth and substantive justice are best served by defense attorney's functioning in clearly limited ways as zealous advocates for their clients. See, for example Freedman (1975a) and the discussion of legal practice in Goldman (1980).

Strong role differentiation is also often taken to serve as a justification for what are sometimes thought to be morally outrageous actions, policies, or decisions in business; for example, a corporation's closing a plant with devastating effects on a community, or Ford Motor Company's decision to produce the Pinto with a known and lethal defect, or a number of corporations' decisions to continue operations thought to be supportive of profoundly unjust governments. Part of the argument is that business managers do not have the expertise to competently decide questions of social good and social policy and that sacrificing profits in the name of moral values amounts to managers' spending their shareholders' money to forward their own values. Milton Friedman (1962) has captured the position in his often-quoted assertion that the single social responsibility of business managers is to increase profits just as long as deception and fraud are avoided (see also Carr, 1968). Attempting to forward any other moral values, Friedman claims, has the business manager arrogating to himself a role that exceeds his moral rights.

Similar arguments are made for strong role differentiation in medical practice. For example, it is frequently argued that physicians should limit their concerns to the pursuit of the health of their patients. Thus, it is not the business of the physician to make treatment decisions on the basis of concerns about other persons, such as a patient's family, or on the basis of considerations like the high cost of a preferred treatment modality.

Strong Role Differentiation in Health Survey Research

One of the most pressing moral questions facing health survey researchers is the question of strong role differentiation in the profession. In gathering the information required for certain studies, health survey researchers, like the physician collecting information on victims of Huntington's Chorea and their offspring, sometimes find themselves with unexpected information that creates moral dilemmas for them. Other times, the very nature of the information to be gathered in a study raises hard moral questions for researchers. Currently, many studies conducted in research on acquired immunodeficiency syndrome (AIDS) can create exquisite moral dilemmas for health survey researchers.

For example, it may happen that in gathering required information, a researcher learns from a male homosexual that he has tested negatively for exposure to the human immunodeficiency virus (HIV), so he has continued to maintain his preferred lifestyle, which includes changing his sexual partners frequently. If the researcher learns that the subject has had more than 10 homosexual contacts within the past year and knows that this is more likely to indicate exposure to the virus than a negative result in an initial screening for exposure, does she or he have a moral obligation to make this known to the

[2]The view that what is morally permissible and/or morally required in professional life differs substantially from what is permissible and/or required in ordinary conduct is discussed in many places and in regard to many professions. See for example, Hughes (1937), Friedman (1962), Carr (1968), Veatch (1972), Walzer (1973), Freedman (1975a, 1975b), Wasserstrom (1975), Howard (1977), several papers in Hampshire (1978), Freedman (1978), Goldman (1980), Freedman (1981), Martin (1981a, 1981b), several papers in Beauchamp and others (1982), Ellin (1982), several papers in Robison and others (1983), several papers in Luban (1984), Gewirth (1986), and Callahan (1988), particularly Chapter 3.

subject? Or suppose another subject indicates that he has recently had a similar number of contacts, has not been tested, has no plan to be tested, and has no intention of cutting back on contacts because he is resigned to dying from the effects of AIDS. Does the researcher have an obligation to attempt to advise the subject in any way? Does he or she have a moral obligation to breach confidentiality and alert some authority if a subject manifests a cavalier attitude toward infecting others? Or suppose a study calls for testing blood from a cross section of HIV antibodies. Does the researcher have a moral obligation to inform subjects whose blood is used of the results of their tests?

For each of these scenarios, the proponent of strong role differentiation in the profession will hold that the researcher's moral obligations are limited to accurately gathering and reporting the information required by the study; and the argument for this position is likely to be much the same as it is for strong role differentiation in other professions, namely, that the professional's training does not give him or her the expertise necessary to justify assuming the roles of educator or advisor or protector of the public. Those roles, according to this view, are not the researcher's to assume, and assuming them would amount to arrogating to himself or herself authority researchers are not morally entitled to take, any more than an attorney is entitled to assume the authority that belongs to a judge or a jury.

The question of strong role differentiation in health survey research is one that confronts every practitioner. Like every other genuinely morally dilemmic question, it is one that no practitioner can escape answering in one way or another, simply because no researcher can escape deciding whether he or she has any moral obligations beyond merely acquiring and reporting the information given in conducting a study. An invariable feature of such unavoidable decisions is that they can be made well or badly. Making them badly involves an unreflective acceptance of an answer because it is (say) the standard answer within the profession. Standard answers can be morally mistaken, and when they are, individuals who accept them cannot escape moral accountability for being mistaken, because every moral agent is morally accountable for the standards of actions he or she accepts. Whether we like it or not, then, no professional can escape moral responsibility for the standards he or she accepts to govern his or her professional practice. This is not to suggest that responsible professionals will never disagree about what standards should govern their practice—it is the nature of morally dilemmic questions to resist quick resolutions that will satisfy all morally conscientious persons. But it is to suggest that no professional can escape answering the question of strong role differentiation in professional practice, and that because this it the case, the question needs to be carefully considered by each practitioner.

We have already touched on some of the considerations that can go into making the case for strong role differentiation in health survey research—considerations having to do with the accuracy of information given to a professional (for example, just as an attorney's client might lie, a research subject might lie), and considerations concerning the limits of expertise and rightful authority. One of the most attractive features of strong role differentiation in health survey research is that functioning according to this standard gives professionals precise guidance on how they are to conduct themselves in all cases. The task is limited to careful information gathering and reporting, and the professional virtues to be cultivated (for example, diligence) can be set out clearly. But this very feature of strong role differentiation is also just the feature that makes strong role differentiation in professional practice morally troubling, since it simplifies the professional's moral universe, obliging the professional to refrain from intervening when an intervention (be it educating, advising, or in some circumstances, breaching confidentiality) might easily prevent serious harm or injustice from being done (Wasserstrom, 1975). In holding practitioners to a single simple moral standard, the doctrine of strong role differentiation faces the problem that confronts any monistic (that is, single-principled) theory of moral obligation, namely, that the theory is not complex enough to allow for the prevention of certain avoidable harms and violations of moral rights. On the other hand, giving up strong role differentiation opens the practitioner to a professional life fraught with moral hazards and hard decisions. Left without a single, clear principle to prevent moral dilemmas from arising, the professional must now weigh competing moral concerns and make judgments regarding which concern takes priority in a given case. Such a professional world is a far more complicated and morally uncertain one than is one governed by the principle of strong role differentiation.

To acknowledge this, however, is not to resolve the question in favor of strong role differentiation in professional practice generally or in the practice of health survey research in particular. Outside of our professional roles, few of us function according to some single moral principle, and although we often do not have the luxury of certainty regarding the many moral judgments we are forced to make, we can always be certain that the decision-making procedures we use have carefully taken into account the competing moral values that come together to make a question a hard moral one. In formulating the first full Western philosophical treatise on ethics over 2,000 years ago, Aristotle warned his students more than once that morality is an area of inquiry importantly unlike other areas of inquiry, and that they would be making a mistake in moral inquiry to expect the kind of certainty and precision that can be expected in the mathematical or the empirical sciences (Aristotle, Bk. 1, chaps. 3, 7). Many moral philosophers today believe that Aristotle's advice remains correct, and that the search for a single principle to govern any area of our moral lives is misguided (e.g., Nagel, 1979; Williams, 1985); and this includes the principle of strong role differentiation in professional practice.

Conclusion

The purposes of this discussion have been (1) to suggest a way of conceptualizing the question of what moral

obligations accompany gaining information in health survey research and (2) to suggest that this question, once conceptualized as the question of strong role differentiation in the profession, is an open one that no health survey researcher can escape answering. Despite the initial plausibility of some of the arguments for strong role differentiation in the professions, my own view is that conducting professional practice according to this single principle is seldom justified in a profession, and that health survey researchers sometimes will have moral obligations beyond accurately collecting and reporting information. If this view is correct, then the profession itself has a duty to do as much as it can to clarify the extent of such obligations. Accomplishing this may call for the profession's having a strong professional association and a useful code of ethics to help guide practitioners. Minimally, since this view allows that hard moral questions do arise for health survey researchers, it calls for continuing, careful discussions of those questions.

References

Aristotle, *Nichomachean Ethics*. Many editions.

Beauchamp, T. L., Faden, R. R., Wallace, R. J. Jr., & associate (Eds.). (1982). *Ethical issues in social science research.* Baltimore: Johns Hopkins University Press.

Callahan, J. C. (Ed.). (1988). *Ethical issues in professional life.* New York: Oxford University Press.

Carr, A. Z. (1968). Is business bluffing ethical? *Harvard Business Review,* January-February, 143-53.

Ellin, J. S. (1982). Special professional morality and the duty of veracity. *Business and Professional Ethics Journal,* 1(2), 75-90.

Freedman, B. (1978). A meta-ethics for professional morality. *Ethics,* 89(1), 1-19.

Freedman, B. (1981). What really makes professional morality different: Response to Martin. *Ethics,* 91(4), 626-630.

Freedman, M. H. (1975a). *Lawyers' ethics in an adversary system.* Indianapolis, IN: Bobbs-Merrill.

Freedman, M. H. (1975b). Personal responsibility in a professional system. *Catholic University Law Review,* 27, 191-206.

Friedman, M. (1962). *Capitalism and freedom.* Chicago: University of Chicago Press.

Gewirth, A. (1986). Professional ethics: The separatist thesis. *Ethics,* 96(2), 282-300.

Goldman, A. H. (1980). *The moral foundations of professional ethics.* Totowa, NJ: Rowman and Littlefield.

Hampshire, S. (Ed.). (1978). *Public and private morality.* New York: Cambridge University Press.

Howard, K. (1977). Must public hands be dirty? *Journal of Value Inquiry,* 11(1), 29-40.

Hughes, E. C. (1937). Institutional office and the person. *American Journal of Sociology,* 43, 404-413.

Luban, D. (Ed.). (1984). *The good lawyer: Lawyers' roles and lawyers' ethics.* Totowa, NJ: Rowman and Allanheld.

Martin, M. W. (1981a). Professional and ordinary morality; A reply to Freedman. *Ethics,* 91(4), 631-33.

Martin, M. W. (1981b). Rights and the meta-ethics of professional morality. *Ethics,* 91(4), 619-25.

Nagel, T. (1979). The fragmentation of value. In *Mortal questions.* (pp. 128-41). New York: Cambridge University Press.

Robison, W. L., Pritchard, M. S., & Ellin, J. S. (Eds.). (1983). *Profits and professions: Essays in business and professional ethics.* Clifton, NJ: Humana Press.

Veatch, R. M. (1972). Medical ethics: Professional or universal? *Harvard Theological Review,* 65, 531-559.

Walzer, M. (1973). Political action: The problem of dirty hands. *Philosophy and Public Affairs,* 2(2), 160-180.

Wasserstrom, R. A. (1975). Lawyers as professionals: Some moral issues. *Human Rights,* 5(1), 1-24.

Williams, B. (1985). *Ethics and the limits of philosophy.* Cambridge, MA: Cambridge University Press.

SESSION SUMMARY

Collecting Data from Samples of Older Adults and Nursing Home Populations

Doris Northrup, Recorder, and Jack Elinson, Chair

The discussion focused on several major issues in designing and executing studies of the older population—sampling issues; nonresponse issues (including the desirability of using proxy respondents); special measurement issues; and the issue of the relationship between university-based researchers studying the elderly and the Federal Government.

Definitions of the elderly vary greatly from one study to the next—some studies of the elderly include persons 50 and older while others are interested only in the oldest-old, usually described as 85 and older. For this reason, caution must be taken in generalizing about research on the elderly because the study samples vary greatly in their range of inclusion.

With respect to the issues of sample design, three general approaches to designing samples of the elderly were considered: telephone screening, list frames, and piggybacking on a larger study. Each method has certain problems associated with it. Random digit dialing telephone screening to find elderly respondents can be expensive, especially as the age of the population surveyed rises. In addition, although telephone coverage among the elderly is high in the United States, there is still some problem of nontelephone households and the need to consider physical factors of the elderly such as hearing loss.

The use of list frames has some advantages, and the most satisfactory frame seemed to be the Medicare eligibility list. However, little documented evidence exists concerning the completeness or currency of its coverage or its quality in other ways. Since the Medicare eligibility list is used to make payments to the Medicare-eligible population, it was suggested that HCFA makes every effort to keep it accurate and up to date. However, two other problems present barriers to using the HCFA files for sampling: (1) it is not clear that HCFA will make the file available as a sampling frame except to researchers working on studies sponsored by HCFA and certain other agencies, and (2) telephone numbers are not provided, which would almost certainly lead to problems of nonresponse in telephone surveys.

Voter registration lists were also discussed, but coverage problems of such frames may be of greater concern for certain elderly subgroups.

The idea of piggybacking surveys of the elderly on to large population-based health surveys (an approach used in obtaining the sample for the Longitudinal Study on Aging from the Health Interview Survey sample) was seen as advantageous in many ways; however at least two barriers to its use exist. First, it is an option that is available to very few researchers. Second, unless special measures are taken at the time the sample is drawn for the original survey, it is likely to produce very small samples of the oldest-old, a problem in some studies.

Related to the sample design issue was the potential problem of noncoverage when the institutionalized are omitted from studies of the elderly. The institutionalized will be of increased importance as the oldest-old become the target of investigation. Studies of the institutionalized suffer from many of the same problems of nonresponse discussed below.

With respect to response rates, contradictory evidence was presented as to whether they are more problematic for the elderly than for other respondents. Participants also expressed some difference of opinion as to whether the elderly like to be interviewed or not. Also discussed were the possibility of a decline in response rates with age from the young-old to the oldest-old as well as the need to better understand the use of proxies and the problems of attrition.

The voluntary attrition or dropout problem exists in all longitudinal surveys; but it seems to be of greater concern in studies of the elderly, given the increased chances of losing the most unhealthy respondents because of institutionalization or death. It was pointed out that one should distinguish death from attrition as a cause of sample loss since, strictly speaking, death is not attrition. At least in certain studies, death can be viewed as a finding or outcome, depending on the analytic objectives of the study.

Doris Northrup is with CODA, Inc., Silver Spring, MD. Jack Elinson is with the Institute for Health Care Policy, Columbia University.

A possible solution proposed for these attrition-related longitudinal study problems was the idea of drawing a new cross-sectional sample to be interviewed concurrently with the last wave of the longitudinal study so that the results could be compared with the data for the survivors. Of course, it must be noted that if one is looking at a particular age cohort, there would be as much loss due to death and institutionalization in a new cross-sectional sample as in the survivors of the longitudinal study.

Much discussion was devoted to the use of proxy respondents, and there seemed to be a general willingness to reconsider the conventional wisdom that proxy data are of lower quality than self-response data. A more positive view of the advantages of using proxies was espoused even by some who had previously felt that proxy data were undesirable. It was pointed out that for studies of the elderly, a frequently cited reason for using proxies is that the sampled person is not mentally competent to respond; and there seemed to be fairly general consensus that in these circumstances, it was better to get proxy data than no data at all. Also, the use of proxies for reasons of respondent incompetence increases greatly with age of the sampled person and is also probably related to health status. The importance of developing better measures to use to determine when respondents are not mentally competent for self-response was emphasized.

Proxies—or perhaps more appropriately multiple respondents—might also come into play when data are needed that cannot be obtained from or are not available to the respondent, such as medical record data. It seemed to be generally agreed that this was a desirable procedure.

It was suggested that proxy data might be more satisfactory for factual data than for evaluative or subjective data (although no clearcut method for categorizing items along these lines was offered). It was also suggested that when proxies are needed, it might be better to get the same data from more than one proxy. Good candidates for proxies might be spouses, other household members (particularly when there is no spouse), relatives who are nonhousehold members, caregivers, or neighbors. Finally, it was also suggested that in studies that rely heavily on proxy data, it would be a good practice to try to get proxy data for everyone, including those from whom direct responses were obtained, so that the relationship between proxy data and self-response data could be examined and the problem of self-selection that exists in many studies of proxies could be avoided.

Also discussed was a research effort underway at the Survey Research Laboratory at the University of Illinois to study cognitive aspects of proxy reporting. The ZUMA Center for Survey Methods and Analysis is also participating in this study, which is being done under a grant from the National Science Foundation. Pairs from the same households—spouses or roommates—are being asked to respond to a broad range of typical survey behavior and attitude questions about themselves and the other member of the pair and to think aloud about the methods they used to answer the questions. Additional experiments are planned to explore two key dimensions—level of involvement and salience.

Several special measurement issues were also discussed, including test-retest issues in longitudinal studies, the possible advantages of unstructured interviews, and determining the optimal length of time between measurement points in longitudinal surveys. It was felt that more research was needed on how to measure change. For example, when changes in disability level are detected, how can the researcher determine whether real change has occurred or the apparent change is simply due to measurement error? It was suggested that one needs a study design with at least three points in time to examine this problem.

Also discussed was the observation that when elderly respondents are interviewed, they often want to tell their experiences in their own way; it is difficult to get them to follow the prescribed order of the interview. Related to this point is the fact that in recent years, there have been several attacks from outside the field of survey research on the strict adherence to structured interviews (see Mishler, 1987; Briggs, 1986; Jordan and Suchman, 1988; and Riessman, 1989). In addition, some survey researchers have called for a reevaluation of this issue. Current evidence, however, for moving away from structured interviews is very anecdotal.

The question was raised as to whether the relationship between the Federal Government and university-based researchers should be reexamined. The suggestion was made that university-based researchers are at a disadvantage in carrying out their surveys for two reasons. First, they cannot obtain the high response rates that are obtained in government studies because respondents may not consider the study as important when the interviewer cannot mention government sponsorship. Second, gaining access to large government list frames or piggybacking on a large government study is more difficult for university-based researchers.

In response, it was pointed out that government sponsorship can be a two-edged sword. In studies involving questions on sensitive topics, respondents may feel the government is intruding into areas inappropriately. Also, government studies involve many layers of approvals, and often constraints are put on the researcher in terms of what may be asked.

Considerations for Further Research

Several issues for future research on elderly populations were discused:

1. To design and execute studies of the older population it is essential to clearly specify the age groups under study. The studies described in this volume, for example, vary greatly in their definitions of the elderly. There is a need for consensus with respect to terms of reference for the various segments of the older adult population, such as young-old, old-old, oldest-old, frail elderly, etc.

2. In using telephone surveys with elderly respondents, researchers must consider their physical limitations, deterioration in health status, and lifestyle differences. For example, research experience from the

national Longitudinal Study of Aging suggests that desirable times for conducting telephone surveys among the elderly are different from those for younger adults.

3. Because rates for the elderly in the noninstitutionalized population decline with age, greater efforts and more resources are necessary to improve response rates, especially among the oldest-old. The relative success of the National Center for Health Statistics in achieving high response rates compared with the efforts of academically based research centers requires further investigation. Among the issues to be addressed are the quality of the individual information obtained when response rates are very high; the attitudes on the part of respondents with respect to alleged government interference and the uses to which the information may be put; and privileged access to governmentally maintained rosters, such as Medicare eligibility lists.

4. The possibility of improving response rates by proxy respondents also needs further study. For example, in nursing homes—where response rates to individual survey questions decline with age—proxy respondents, including nursing personnel, may provide useful information especially with regard to questions of objective fact. Other proxy-related issues for research are the reasons for using proxies, the type of data for which they are most useful, and the interaction of these factors with the type of proxy respondent.

5. The optimum timing of intervals between waves in longitudinal studies of the elderly also needs investigation. The tradeoff seems to be that the longer the interval, the greater sample attrition; but the shorter the interval, the less change observed. Although no single answer to this problem may exist since it depends very much on what one is trying to study, research in this area could improve survey design for all studies of the elderly.

6. Surveys of older people have reawakened interest in allowing respondents to answer questions in a more unstructured way rather than to force responses into predetermined answer categories. To meet the objectives of large-scale quantitative surveys, research that takes into account idiosyncratic response patterns of respondents is necessary.

References

Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research.* Cambridge, England: Cambridge University Press.

Jordan, B., and Suchman, L. (1988). Interactional troubles in survey interviews. Unpublished manuscript.

Mishler, E. (1987). *Research interviewing.* Cambridge, MA: Harvard University Press.

Riessman, C.K. (1989, August). Life events, meaning, and narratives: A case of infidelity and divorce. *Social Science and Medicine* 29(6), pp. 743-751.

Samples for Studies Relating to AIDS

Measuring Behavior Related to Risk of AIDS

Introduction by Floyd J. Fowler, Jr.

Although the conference as a whole addressed general methodological issues, the specific problem of how to use survey research methods to conduct acquired immunodeficiency syndrome (AIDS)-related research clearly was an overriding theme. Consequently, two sessions were devoted to the general topic of AIDS. The first focused on sampling concerns; the second focused on data collection issues.

The challenges posed by AIDS are, in fact, not new. There is a long and reasonably well-developed literature on the collection of data about sensitive topics. A literature also exists on doing surveys of rare or hard-to-find populations which provides a helpful background for research on AIDS. But AIDS-related research brings together in a single area very difficult methodological problems. Surveys have been used to estimate contraceptive practices, rates of abortions, drug use, excessive alcohol use, and other personal or socially stigmatized behaviors. However, the activities that affect the risk of contracting AIDS are almost all illegal or personal, and the degree of detail that researchers want in order to truly assess vulnerability is very high. In addition, sampling the groups in which AIDS researchers have a special interest—intravenous drug users, homosexual and bisexual men, prostitutes, prisoners, and persons with AIDS—clearly stretches the limits of current methodological knowledge and techniques.

The feature papers presented in the two sessions devoted to AIDS reflect a wide diversity of research designs, objectives, and problems. The first session's papers include presentations relevant to how one finds and enlists the cooperation of samples of people needed for the study of behaviors relevant to the spread of AIDS. The papers in the second session address ways to ask questions and induce respondents to give accurate and meaningful answers about behaviors critical for assessment of their risk of AIDS.

One fundamental premise of the conference—that survey methodology must be viewed from a total design perspective—is synthesized in the concluding discussion paper which puts the methodological issues in a broader perspective and provides a framework for examining sampling, nonresponse, question design, and data collection methods in an integrated way. Because all of these are potential sources of survey error, undue emphasis cannot be placed on any one feature of a survey to the exclusion of others. Data emanating from a survey are no better than the worst feature of the methodology, and researchers must consider the variety of decisions that affect their data.

FEATURE PAPER

Efficiency of General Population Screening for Persons at Elevated Risk of HIV Infection: Evidence from a Statewide Telephone Survey of California Adults

Frank J. Capell and Greg Schiller

Introduction

Most data on behaviors associated with transmission of human immunodeficiency virus (HIV) have been obtained from convenience samples. This owes both to the understandable research emphasis on so-called "sentinel" settings, such as sexually transmitted disease clinics where high risk persons are likely to go for treatment, to difficulties associated with probability sampling of rare or sensitive events. Increasingly however it is being recognized that population-based data on risk factor prevalence are essential for effective acquired immunodeficiency syndrome (AIDS) prevention planning and program evaluation. California has conducted two statewide probability sample telephone surveys of HIV risk behavior prevalence and a number of other states, including Minnesota, Arizona, and Washington, are planning similar projects.

The chief purposes for such surveys are to provide a basis for estimating the number of persons in defined populations engaging in HIV transmission-associated behavior; to obtain descriptive information on specific behaviors, for example, on sex partner acquisition rates; and to assess population knowledge levels, attitudes, and beliefs regarding HIV and AIDS. The acronym KABB, for Knowledge, Attitudes, Beliefs and Behavior, has been coined to refer collectively to this group of survey topics. It should be noted that these different purposes, while often rolled into a single survey project, are not all equally well served by the same approach to sampling. Probably the most difficult survey objective is to develop a large enough probability sample of high risk persons to permit analysis of risk behavior and factors that may be associated with it.

Frank J. Capell and Greg Schiller are with the Office of AIDS, California Department of Health Services, Sacramento.

This work was partially supported by Cooperative Agreement No. U62/CCU902019 from the Centers for Disease Control to the California Department of Health Services, Office of AIDS.

This paper focuses on the behavioral component of population-based KABB surveys. In particular, it describes results from an attempt to design a probability sample telephone survey so as to improve the efficiency with which individuals reporting high risk behavior are captured in the sample. The questions to be addressed are: (1) to what extent is it possible to increase the proportion of respondents in a general population survey who qualify as members of one or more high risk groups; (2) what specific survey procedures contribute most to risk group sampling efficiency; and, (3) what recommendations can be made and what questions remain for future surveys seeking to capture respondents who may be at elevated risk for HIV infection?

Method

Prior information on the geographic distribution within California of persons engaging in high risk activities was obtained from a statewide general population telephone survey of AIDS KABB fielded in late 1987. Of 2,012 respondents, this initial survey yielded roughly 375 adults who admitted engaging in one or more high risk activity. Using reported amount and kind of risk behavior, an ad hoc system was developed for assigning "risk scores" to sample respondents. Especially high scores were assigned to those respondents reporting male homosexual activity, intravenous (IV) drug use; or heterosexual contact with bisexual males, IV drug users, HIV-positive persons, or multiple partners with low or unknown risk. Respondents reporting recreational drug use (RDU) or who were in an "open couple" relationship (where one or both members could have sexual contacts outside the relationship) were assigned intermediate scores; all remaining respondents were assigned scores of zero. To reflect the interest of the study in sampling as many high risk women and minorities as possible, risk scores for these subgroups were doubled (tripled for those both female and minority).

Table 1. Selected population and survey sample* characteristics

	Stratum		
	High	Medium	Low
1980 Census characteristic			
Total population	4,488,162	6,943,443	12,313,561
Cumulative AIDS cases through 7/88	4,173	3,918	4,207
Cumulative AIDS cases per 100,000	92.98	56.43	34.17
Proportion of nonfamily households	0.41	0.41	0.37
Percent nonwhite	51.61	41.41	22.39
Ratio single to all males 20 to 44 years	0.65	0.62	0.58
Sample characteristic			
Sample size	1,482	1,663	1,516
Percent male	44.53	40.77	44.00
Gay or bisexual men (%)	2.02	1.26	.99
All HRH men (%)	8.20	8.16	6.95
Men w/high risk partner (%)	2.02	1.08	.92
All HRH women (%)	4.24	3.96	3.94
Women w/high risk partner (%)	1.75	1.44	1.72

HRH = high risk heterosexual
SOURCE: California Department of Health Services

Risk scores were then aggregated by ZIP code, and areas where the average score for respondents exceeded an essentially arbitrary cutoff were assigned to the first of three strata, referred to as the high risk stratum. The second, or medium risk, stratum was composed of all ZIP codes with average risk scores greater than zero but less than the cutoff and any California ZIP codes not already accounted for which, according to the 1980 census, contained relatively high proportions of minority telephone households (≥40 percent black, ≥40 percent Hispanic, or ≥15 percent Asian). The final low risk stratum included all remaining areas of the State. The low risk stratum therefore consists of ZIP codes for which prior information suggested a relative absence of high risk persons, or for which no prior information existed, that is, which were not sampled in the 1987 survey. These ZIP code groupings were forwarded to Survey Sampling, Inc., of Fairfield, CT, where sample phone number lists were generated. The goal was to oversample numbers in the high risk and medium risk strata by factors of three and two, respectively, relative to telephone numbers in the low risk stratum. However, with the additional constraint that numbers from Los Angeles County not comprise more than one third of the total pool, it was only possible to achieve a ratio of sampling fractions for the three strata of approximately 2.7:2:1.

Table 1 displays selected census characteristics for the three strata. Just under one-fifth of the State's 1980 population resided in ZIP code areas included in the high risk stratum, while over half were in the low risk stratum. The strata contain very nearly equal portions of the State's cumulative mid-1988 AIDS cases with known ZIP codes (about 92% of all cases). However, when AIDS cases are expressed as cumulative incidence rates per 100,000 population the strata are markedly different, with the rate in the high risk stratum nearly triple that for the low risk stratum. The stratum populations also differ in ethnic composition, which is the only census characteristic explicitly incorporated into stratum definitions, and to a lesser degree in density of single males.

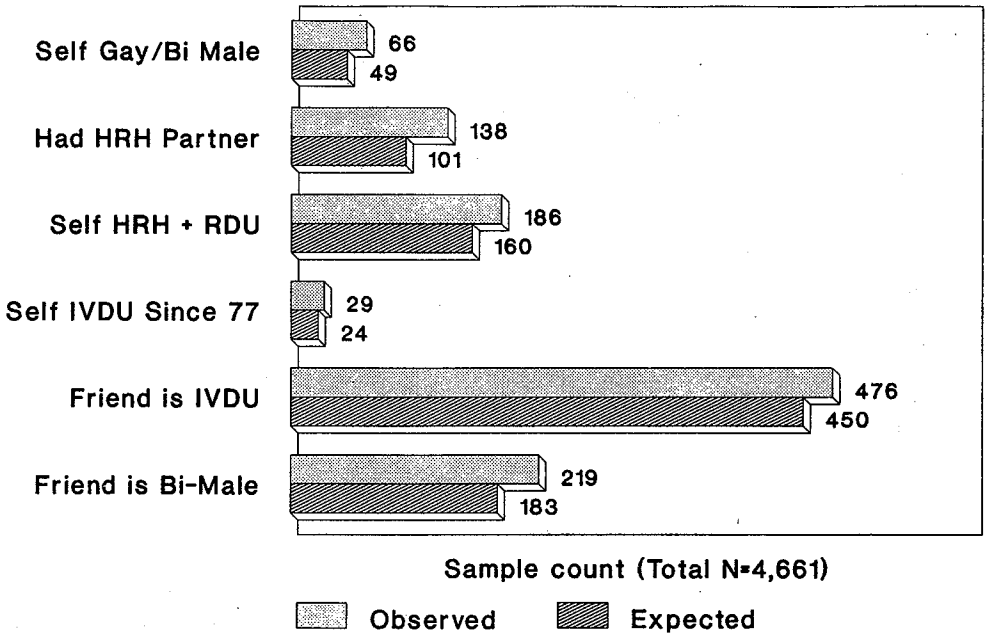
The survey was fielded between October and December 1988. After up to 20 attempts, an interview was completed with a randomly selected adult in approximately 68 percent of contacted households, for a total sample size of 4,661. All respondents were administered a brief screening interview; 800 of these reported one or more of the targeted high risk activities and were administered an additional series of questions. Among outright refusals, whether before or after respondent selection, 65 percent were female.

Results

The lower portion of Table 1 shows selected characteristics of the obtained sample, by stratum. The medium risk stratum contains slightly more than its share of the sample as well as a significantly smaller percentage of male respondents. The last five rows in the table present raw stratum percentages of respondents qualifying as gay or bisexual men, high risk heterosexual (HRH) men and HRH women; also shown is the percentage of respondents qualifying as high risk heterosexual specifically through heterosexual contact with a member of a primary AIDS risk group. Overall for high risk men and women there is a tendency for the stratum percentages to be in line with what would be expected, that is, highest in the high risk stratum and lowest in the low risk stratum. For gay or bisexual men the difference between the high and low risk strata in raw percentage of qualifying respondents is statistically significant.

To provide a comparative base for assessing the efficiency of the sample, "expected" numbers of persons reporting behaviors or characteristics of interest were derived by applying three weights to the sample data: (1) a correction for unequal selection probabilities (CUSP) to adjust for differences in household size and

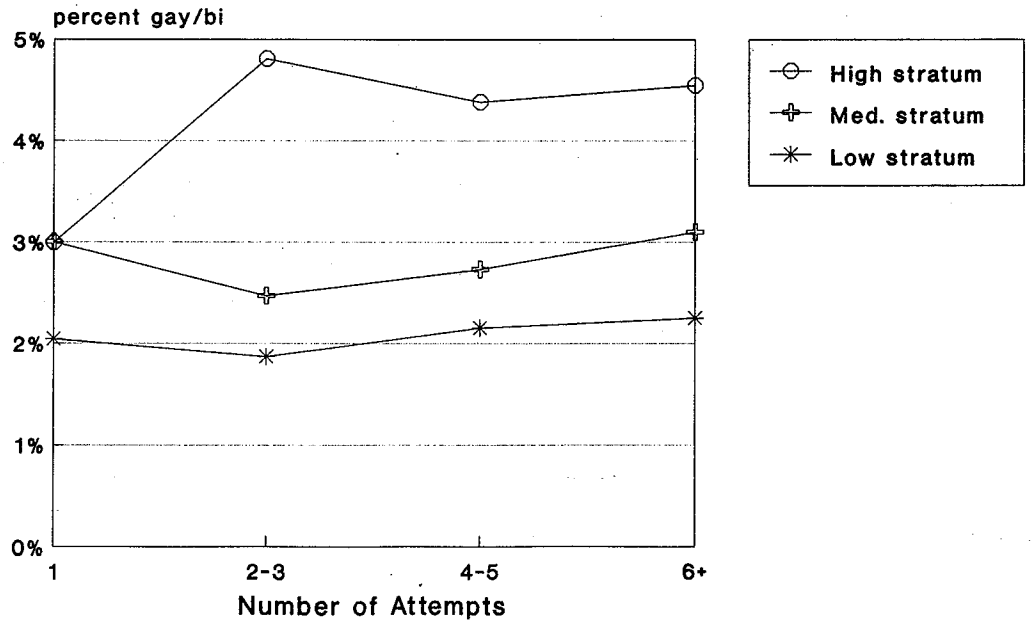
Figure 1. Observed and expected* counts for selected HIV-related characteristics (California 1988)



* "CUSP", stratum & post-stratif. wts.

number of telephone lines; (2) a stratum adjustment to correct for disproportionate sampling within strata; and (3) a poststratification adjustment to align the sample age-race-sex composition with the 1988 California adult population. The resulting weighted counts are interpreted as numbers of persons who would be expected to report the targeted behavior or characteristic in a representative sample of individuals.

Figure 2. Cumulative percent men self-identifying as Gay/Bisexual by stratum and number of interview attempts



SOURCE: CA DHS Office of AIDS, 1989

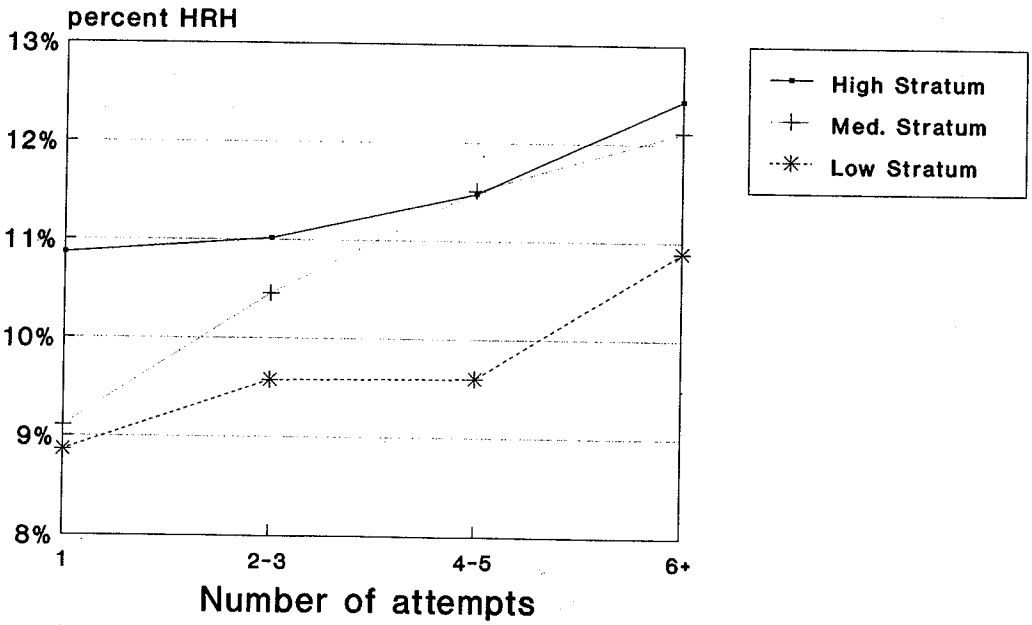
Figure 1 shows the observed and expected counts for several HIV-related characteristics. In each case the unweighted count exceeds the corresponding weighted figure, suggesting that the sample has yielded larger numbers of high risk persons than might be expected under representative sampling. The median improvement in relative efficiency is about 20 percent, ranging from a low of 6 percent for persons reporting an IV drug user among their close friends to a high of 35 percent for self-identified gay or bisexual men. Before examining these findings further, it is necessary to consider a related aspect of the sampling plan—its efficiency in capturing ethnic minorities. Compared to 1980 census estimates of nonwhite households for the three strata, the unweighted sample ethnic composition is a little disappointing. The relative increases in efficiency for black and Asian respondents are just 9 percent and 2 percent, respectively; for Hispanics, however, the sample is actually less efficient, by 8 percent, than would be expected for a representative design. Whether the undersampling of Hispanics results from differential nonresponse or from discrepancies between current ethnic and minority residential patterns and those estimated from the census, it must be kept in mind as other aspects of the survey are discussed.

Figures 2 and 3 present additional information on the efficiency of the stratified sampling plan for capturing persons at elevated risk of HIV infection. Figure 2 displays within-stratum prevalence estimates for homosexual activity among male respondents as a function of number of interview attempts, somewhat coarsely grouped. This activity is reported twice as often among men sampled from the high risk stratum as among those from the low risk stratum. In addition, estimated prevalence of men reporting sex with other men tends to increase steadily, without regard to stratum, as respondents reached after larger numbers of interview attempts are added to the sample. This suggests that these men are harder to reach than other men.

Figure 3 shows a similar pattern for respondents qualifying as high risk heterosexual. Differences among strata in prevalence of HRH activity are apparent; but equally apparent are the increases in HRH prevalence with increased persistence in seeking to complete an interview. To check whether these stratum differences in gay and bisexual men and HRH prevalence might possibly be due to differential diligence on the part of interviewers, Figure 4 plots the cumulative proportion of completed interviews after varying numbers of callbacks within each stratum. There is no indication of unequal allocation of callback resources among sample strata, as might be inferred were the cumulative proportion of interviews ultimately completed within one stratum to approach unity more rapidly or slowly than in the other strata.

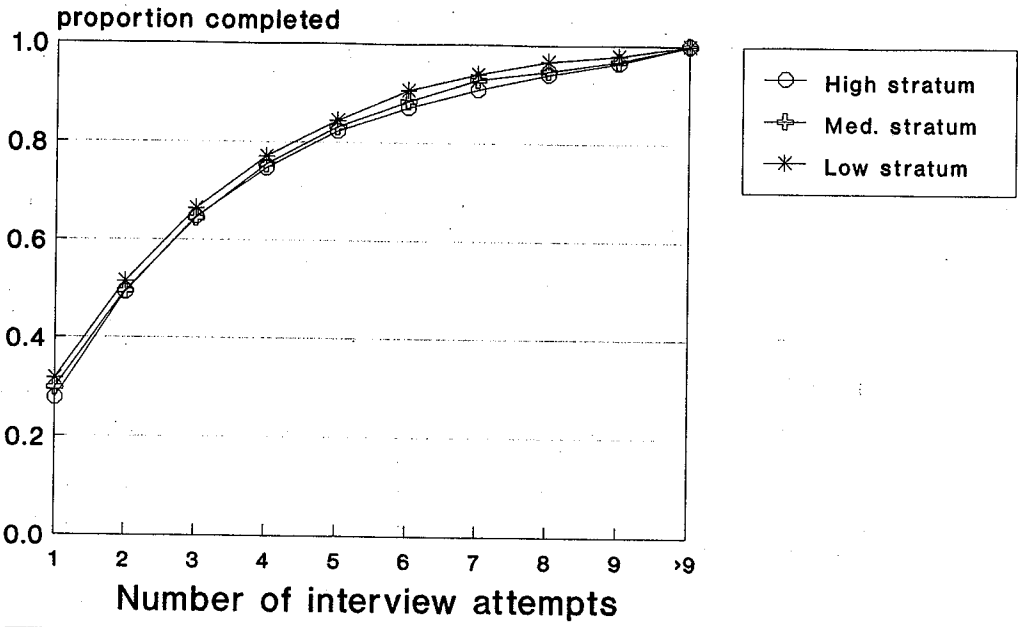
Figure 5 focuses exclusively on the impact on risk group prevalence estimates of increasing numbers of callbacks, and plots the relative bias in risk group prevalence for gay and bisexual men and HRH men and women after varying numbers of interview attempts. Relative bias is defined as the difference between prevalence estimated for respondents reached after a given number of callbacks and prevalence estimated from the total sample divided by total sample prevalence. To assist in interpreting these relative differences, coefficients of variation for final sample risk group prevalence esti-

Figure 3. Cumulative percent high risk heterosexual (HRH) respondents by stratum and number of interview attempts



SOURCE: CA DHS Office of AIDS, 1989

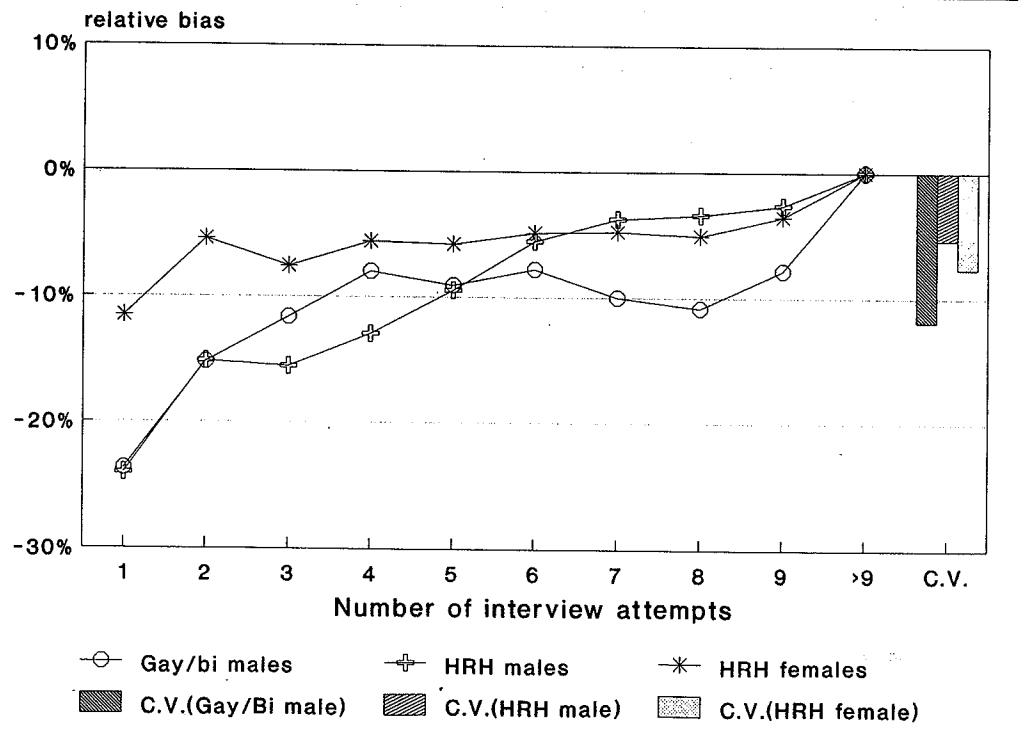
Figure 4. Cumulative proportion of completed interviews by stratum and number of interview attempts



SOURCE: CA DHS Office of AIDS, 1989

mates are shown as bars at the right of the figure. Clearly, harder-to-reach men are more likely to report high risk activity. For both men and women, the largest reduction in bias comes from the first callback, that is, the second interview attempt. For women, however, the gradient of bias reduction from subsequent callbacks is quite flat, suggesting rather minor increases in HRH prevalence. It is tempting to link this to the somewhat

Figure 5. Relative bias in risk group prevalence estimated after varying numbers of interview attempts



SOURCE: CA DHS Office of AIDS, 1989

larger initial refusal rate for women; for example, hard-to-reach female respondents may be more likely to successfully refuse the interview. Among gay and bisexual respondents, sample prevalence remains biased 10 percent downward even after eight interview attempts.

Discussion

The survey described in this presentation sought, among other things, to develop a minimum-size sample of persons, screened from the general population, who reported one or more activities that may place them at elevated risk for HIV infection. Unfortunately, fielding of the survey coincided with a period of heated public debate in California over a controversial ballot proposition that would have mandated reporting of HIV-positive persons as well as those "suspected of being HIV positive." This latter aspect of the proposition was construed by many to include persons engaging in high risk activities. (One suggestion was to forward the entire San Francisco phone book to public health authorities!) The survey itself was the target of investigations by several San Francisco journalists during the time it was being fielded. With this backdrop, it is not surprising that even with the measures described for improving risk group sampling efficiency, the overall "hit rate" barely equaled that attained a year earlier with an essentially random sampling plan.

A key question in assessing the success of the approach to forming sample strata is the extent to which geographic clustering of high risk persons actually occurs. The answer appears mixed: For gay and bisexual men in California there is a clear geographic concentration of risk group members. The high risk stratum in the 1988 California KABB survey sample produced a 4.5 percent prevalence of men reporting sex with other men, a rate significantly higher than other areas of the State. If the objective had been to screen exclusively for gay and bisexual men, for example by focusing on census tracts corresponding to gay districts in cities such as San Francisco, prevalence could have been increased to about 10 percent (L. L. Bye, personal communication, 1989). For HRH men and women, the observed differences among strata were less marked. Whether this is due to the essentially incomplete nature of the prior information used to form strata or to a relative lack of geographic clustering among high risk heterosexuals, or both, is not clear.

We intend to iterate through the stratification process again in 1989, treating data from both the 1987 and 1988 surveys as prior information. The 1988 survey also included a series of questions on close friends who engage in high risk activities (Figure 1); estimates of the density of high risk persons in respondent social networks may provide a more stable basis for targeting sampling. In addition, by the time the next statewide sample survey is conducted, other more crucial information on the geographic distribution of HIV infections among selected groups, for example childbearing women, should be available to further specify areas of the State where survey data collection should be concentrated.

The 1988 California KABB survey also provided useful information on the value of persistence in seeking to complete interviews with high risk persons. For self-identified gay and bisexual men, the statewide risk group size is underestimated by more than 60,000 when the sample is restricted to those reached on the initial interview attempt, and by 30,000 after as many as 8 attempts. Similarly, had callbacks been discontinued after 5 attempts, the size of the high risk heterosexual subpopulation of the State would have been underestimated by roughly 200,000.

FEATURE PAPER

Use of Telephone Surveys in AIDS-Related Community Research

Howard E. Freeman, Kathleen Montgomery, Charles E. Lewis, and Christopher R. Corey

Introduction

In mid-1988 a telephone survey of men living in Los Angeles County was undertaken to estimate the size and characteristics of the population at risk of being human immunodeficiency virus (HIV)-infected. This survey was part of a research-demonstration project that had the following two objectives: (1) to develop a measure of HIV risk status and (2) to evaluate whether males identified at high risk on the basis of the survey information, when informed of their risk status, would follow recommendations that they seek counseling, HIV testing, or medical care.

Although other approaches were considered, data collection by computer-assisted telephone interviews was chosen because of the special need to "score" individuals on an algorithm during the conduct of the interview, as well as the general advantages of this method that are often advanced. This paper reports on this experience on the use of a telephone survey in AIDS-related community research, particularly for studies directed at identifying the high risk population.

Background

Since Gottlieb and associates (1981) described the acquired immune deficiency syndrome (AIDS) less than a decade ago, AIDS has been found among persons in virtually every country in the world and properly may be regarded as the first worldwide epidemic of life-threatening proportions in modern times. Consequently, it has provoked extensive laboratory research and has been the foci of numerous drug trials. Despite the frustrations of limited progress from these biomedical efforts, the speed and breadth of their initiation has been truly remarkable.

There is a conspicuous absence of information, however, about AIDS-related risk behavior among community members (Turner & associates, 1989). The database consists almost exclusively of information from longitudinal studies of volunteers for the Multi-Center AIDS Cohort Study (MACS) of gay men, and studies of special populations such as patients seen in sexually transmitted disease clinics, clients of drug rehabilitation programs, hemophiliacs, blood donors, and others (Burke & associates, 1987; Chaisson & associates, 1987; Cleary & associates, 1988; Kaslow & associates, 1987; Quinn & associates, 1988).

Community surveys are essential to obtain epidemiologic data and the information necessary to design and implement effective public health and medical care programs. In particular, survey information is required to estimate HIV risk status and to target community members at high risk. In response to the current lack of information, several national surveys are in the planning and pilot stages; moreover, data collection in a number of statewide and large-city studies has been completed and analyses and preliminary reports are emerging.

Given the unavoidable resource constraints, both the design and implementation of community studies—whether national, state, or local—in almost all cases require making compromises between generalization potential, breadth and depth of content covered, and precision of estimates. In addition, it is essential to identify the most cost-beneficial data collection approach for risk estimation surveys.

Each of the three major data collection methods, face-to-face interviews, telephone interviews, and questionnaires, has advantages and limitations, and each approach has its advocates and antagonists. For obtaining risk-related information, however, there are strong reasons to argue that the telephone interview is the most

Howard E. Freeman, Kathleen Montgomery, and Christopher R. Corey are with the Department of Sociology, University of California at Los Angeles. Charles E. Lewis is with the Department of Medicine, University of California at Los Angeles.

The work was supported by Cooperative Agreement U62/CCU 90108-01-1 from the Centers for Disease Control.

appropriate data collection method. Costs are lower compared with face-to-face interviews. Additionally, required branching and other tailoring of interview schedules are easily implemented, screening for eligible interviewees is effectively and efficiently accomplished, consistency checking between items can be undertaken easily, and interview quality control procedures and interviewer supervision are readily routinized.

Of course, data collection by telephone has disadvantages; for example, compared with face-to-face interviewing, data collection by telephone may result in a lower cooperation rate, it does not allow presenting choice categories visually to interviewees, and the presence of other persons near the telephone may be detrimental to respondent cooperation. Questionnaires also may have advantages over data collection by telephone: they are the least expensive data collection approach, they may inspire the strongest feelings of confidentiality by interviewees, and they may allow obtaining the most detailed information.

Unfortunately both time pressures and cost constraints limit the amount of "pure" methodological work that can be undertaken on data collection in AIDS-related community studies. Consequently, to understand the tradeoffs involved in the choice of data collection methods, it is important to accumulate reports on the experiences of different investigators. These reports then can be collated and generalized by formal and impressionistic metaanalysis. It is in this spirit that this report on efforts to use a telephone interview to obtain information on HIV risk measures among the Los Angeles County male population is made.

The Los Angeles Survey

Given the paucity of AIDS-related community research at the time the Los Angeles survey was being planned, it was necessary to make a number of design decisions without the benefit of analogous research. These decisions centered around sample design and selection and interview content.

Sample Design and Selection

The survey budget permitted approximately 1,600 interviews of about 30 minutes each.[1] A number of procedures were implemented to maximize sampling efficiency:

1. The sample was selected by random-digit dialing; however, a Waksberg procedure was implemented to reduce the number of nonworking numbers that were dialed. At least six callbacks were made to nonanswering households. In total, 1,610 interviews were conducted.
2. In Los Angeles the vast majority of AIDS cases are males between 18 and 60 years of age. Since the intent was to identify high risk persons and refer them to counselors, the sample was confined to this group. Consequently, screening for eligible interviewees was required in each household. Whoever answered the telephone was asked if males of the required ages lived in the household; if the respondent reported more than one, the "next birthday" procedure was used to select the interviewee. If present, the designated male was interviewed; if not immediately available, at least six additional attempts were made to interview him.
3. In addition to restricting the sample to males, households in the 18 census tracts with unusually high historical prevalence rates of AIDS (averaging about 19 per 1,000) were oversampled 14 to 1. The decision to oversample at this ratio was based on the necessity to include a sufficient number of high risk males, taking into account the interviewing budget. This weighting resulted in samples of about even sizes in the high and low prevalence areas. The percentages in the tables appropriately weight for sample stratification and interviewee selection procedures.

Interview Content

Decisions on interview content include human subject requirements and confidentiality, item wording, and the need for fine-grain sexual experience and lifestyle information:

1. Strict compliance with human subject guides is required in the case of AIDS-related community studies, and review groups invariably carefully scrutinize protocols. In this survey, it was required that an informed consent statement be read to potential interviewees. This statement explained the purpose of the study, informed interviewees that they were selected by random-digit dialing, and assured them that they could not be identified by the interviewers. All interviewees at the outset were informed that the survey was on AIDS and that they did not have to answer any questions they felt were too sensitive.
2. A decision was made to use nontechnical but not colloquial terminology for the sexual behavior items. No pretest was made of whether responses would have differed if colloquialisms had been used or available as substitute items. Similarly, terms defining sexual orientation (for example, homosexual, bisexual, and heterosexual) were not used in any of the items that ascertained sexual orientation. Rather, interviewees were asked to provide the gender of their partners.
3. To estimate the probability of risk, detailed information on sexual behavior was required. The items were constructed in consultation with several infectious disease clinicians and epidemiologists to cover those behaviors that were viewed at the time of the survey as risk indicators. Since these behaviors are interdependent, considerable branching was necessary. It began by establishing whether the person had been sexually active during the past 2 years, determining number of partners, and then the sex of the partners. Subsequently, items on specific sexual activities were asked depending on the sex of partners.

[1] Interviewing was undertaken by UCLA's Institute for Social Science Research. Data collection was computer-assisted. The trained interviewers placed calls between 9 A.M. and 7 P.M. on all days of the week from May through July of 1988. Interviews were conducted only in English.

Table 1. Data collection results: Study group sizes

	Reported cases of AIDS		
	High areas	Low areas	Total LA
Completed interviews	798	812	1,610
Refusals	303	383	686
Working numbers, but interviewees not identified or unavailable	679	603	1,282
No answer after six callbacks	278	299	577

Consistency checking was undertaken on number of partners but not on the congruence of reports about different sexual behaviors.

Computer-assisted interviewing clearly was necessary given the complexity of the information collection, a strong argument for a telephone survey. Eventually, the development of computer-assisted personal interviews (CAPI) will eliminate the branching advantage and similar capabilities of telephone interviews.

4. Although homosexual behavior was the key determinant of risk at the time of the survey in Los Angeles, an effort was made to secure information about intravenous drug use.

Findings

As discussed, a key advantage of telephone surveys is cost: The data collection costs of this survey averaged about $30 per interview, about one third of what face-to-face interviews would have cost. However, the decisions made on both sampling design and selection impacted on the outcome of the survey from a method standpoint. A discussion of the most salient of these outcomes follows.

Table 2. Data collection results: Completion rates

	Reported cases of AIDS		
	High areas	Low areas	Total LA[a]
Completed/ completed + refusals	72.5%	68.0%	68.3%
Completed/ completed + refusals + interviewees not identified or unavailable	44.9	45.2	45.2
Completed/ completed + refusals + interviewees not identified or unavailable + no answer after six callbacks	38.8	38.8	38.8

[a]Weighted percentages for total Los Angeles.

Cooperation Rates

The cooperation rates are reported in Tables 1 and 2. Some 4,200 households were contacted, from which 1,610 interviews were obtained. The ratio of completed cases compared with completions plus refusals is around 70 percent, and is not markedly different by density of HIV cases in a census tract (Table 2).

However, this percentage underestimates bias. In approximately an additional 20 percent of the households it was not possible to interview a male between 18 and 60 years of age. This was so because the person answering the telephone either broke off the interview before knowledge of a male living there was ascertained or refused to identify eligible males, or the male was not available at the time of the call or at subsequent callbacks. The proportion of these households that should be included in the noncooperation rate is unknown.

The most stringent estimate of noncooperation includes working telephone numbers that were not answered even after six callbacks at various times of the day and week. The cooperation rate (approximately 40 percent), including all telephones that did not answer despite six or more callbacks, is undoubtedly an underestimate (Groves & Kahn, 1979); a reasonable estimate is in the 50-percent range.

If this were an ordinary survey, a fair question is whether the bias is too high to warrant use of the information. However, the sensitivity of the questions, the human subjects requirement that the survey content be disclosed at the outset, and the scarcity of AIDS-risk information from other than convenience samples must be taken into consideration. It should be noted that the cooperation rate is lower but not markedly so compared with the unpublished Chicago community study (Murphy & Binson, 1988). It is much lower, however, than the cooperation rates discussed at this meeting by groups that are planning national surveys.

It is critical, however, to develop ways to increase the cooperation rate. First, the opening explanation of the purpose of the survey and its sponsorship may affect the cooperation rate. Trials using variations in these respects, consistent with human subject requirements, should be attempted. Second, some of the refusals were from women and persons ineligible by age who initially answered the telephone. It may well be that the advantages of screening noted earlier are overshadowed by the increased noncooperation rate as a result of gatekeeper refusals. Third, experiments should be undertaken on whether either paying interviewees or, for listed telephones, providing advance letters explaining the survey would increase cooperation.

Sample Size

As noted, 1,610 males were interviewed in this survey. This is not a particularly small sample size for community surveys. However, to a considerable extent interest focused on a small proportion of the study group, namely gay men. In total, 172 homosexual or bisexual males were included in the sample, when weighted equaling 5 percent of 18- to 60-year-old males in the county.

Table 3. Sexual behavior patterns placing heterosexuals at high or very high risk of HIV infection

Number of partners[a]	+	Body fluid risk[b]	+	Partner(s) at risk[c]	+	Respondent prostitution[d]	=	Risk level	*N* (unweighted)	Percent of group (weighted) (1,273)
2-3		low		no		yes		high	1	*
2-3		low		yes		yes/no		high	87	7.4
4+		low		no		yes		high	2	
4+		low		yes		no		high	164	13.1
4+		low		yes		yes		very high	1	*
										20.5

[a] Number of partners = number of sexual partners reported by respondent within the past 2 years.
[b] Body fluid risk: low = vaginal intercourse, oral sex, anal insertive intercourse; high = anal receptive intercourse.
[c] Partner(s) at risk: yes = partner has engaged in prostitution in the past 2 years OR partner has had multiple sex partners in the past 2 years OR partner has used intravenous drugs.
[d] Respondent prostitution: yes = respondent has engaged in prostitution within the past 2 years.
* Percentage (weighted) equals 0.1 or less.

NOTE: Risk category combinations with empty cells are not shown.

Inasmuch as gay men constitute a critical core subsample for identifying the high risk population, and since they constitute only a small minority of male community members, large samples are required to make estimates with reasonable confidence limits and to disaggregate estimates by relevant social and demographic characteristics. In general, reasonable estimates cannot be made without samples of, say, 10,000 males in communities where the primary transmission of the virus is by sexual relations. For example, in Los Angeles County, an unweighted sample of 10,000 males would yield about 500 gays; this survey demonstrates that, within these 500, the proportion of men who are sexually active, at high risk, and who currently are not taking prophylactic measures is extremely low. Nevertheless, it is essential to have a reasonably precise estimate of such persons because it is imperative to estimate the spread of the virus.

Moreover, in communities where either the risks of being infected via sexual intercourse are common or increasing among females or there is a high rate of infection from intravenous drug use, surveys must include both sexes; thus sample sizes several times greater than the 10,000 estimate may be necessary.

Estimation of Sex-related Risk

As discussed, a number of items on specific sexual behaviors were asked about interviewees' past and present sexual activities, with different item paths for heterosexual, homosexual and bisexual males, as well as for monogamous and nonmonogamous men. Collection of data in cases in which there were multiple partners is extremely tedious because it is important that data be comprehensive. Information about either a sample of an individual's sexual episodes or an incomplete roster of a person's sexual partners may not uncover his or her high risk sexual experiences.

The results on risk status are shown in Table 3 (for heterosexuals) and Table 4 (for homosexuals). The risk

Table 4. Sexual behavior patterns placing homosexuals at high or very high risk of HIV infection

Number of partners[a]	+	Body fluid risk[b]	+	Partner(s) at risk[c]	+	Respondent prostitution[d]	=	Risk level	*N* (unweighted)	Percent of group (weighted) (1,273)
1		low		yes		yes/no		high	5	4.0
1		high		yes		yes/no		high	4	4.3
2–3		low/high		no		no		high	18	13.0
2–3		low		yes		no		high	6	4.5
4+		low		no		yes/no		high	11	6.9
2–3		high		yes		yes/no		very high	7	5.9
4+		low		yes		yes/no		very high	35	13.7
4+		high		yes/no		yes/no		very high	51	36.3
										88.6

[a] Number of partners = number of sexual partners reported by respondent within the past 2 years.
[b] Body fluid risk: low = vaginal intercourse, oral sex, anal insertive intercourse; high = anal receptive intercourse.
[c] Partner(s) at risk: yes = partner has engaged in prostitution in the past 2 years OR partner has had multiple sex partners in the past 2 years OR partner has used intravenous drugs.
[d] Respondent prostitution: yes = respondent has engaged in prostitution within the past 2 years.

NOTE: Risk category combinations with empty cells are not shown.

levels are based on expert clinical opinion. Among heterosexuals, some 20 percent are classed as of high risk; among homosexuals and bisexuals, some 56 percent are judged at very high risk and an additional 33 percent at high risk. Although the risk estimates obtained are high, there are convenience samples with even higher proportions at risk (Winkelstein & associates, 1987).

Classifying community populations either as in this study, or on the basis of even more gross classification schemes, is typical in the community studies that have been recently undertaken (Capell & Schiller, in this volume). Subsequent to data collection, the critical need to conceptualize risk in more specific terms became clear. The three important dimensions of risk are:

1. acquiring the virus because of ongoing high risk behavior;
2. being HIV seropositive because of past behavior;
3. transmitting the HIV virus because of past and ongoing behavior.

The relevant concept of risk depends on the purposes behind estimating risk status. To develop educational programs directed at minimizing new cases, it obviously is important to estimate the numbers of persons whose current lifestyles place them at high risk of acquiring the virus. In contrast, efforts at encouraging HIV testing and at anticipating the demand for medical care for patients requires estimates of the size of the already HIV-infected group in the community. Within the group who are seropositive there is urgent need to identify those whose current behavior places them at high risk of being the transmitters of the virus to others in the population. These persons, obviously, should be the targets of intensive behavioral change interventions.

In the Los Angeles survey, no collected information was on the frequency of sexual encounters and durations of relationships except in the case of monogamous persons. In addition, information on sexual practices over a longer time than 2 years and regularity and circumstances of condom use is needed to refine risk measures. The investigator is confronted with the need for extended interviews in the case of highly active, multipartner persons; consequently, studies of maximum length of telephone interviews compared with tolerance of interviewees for other modes of data collection are essential for optimizing community research related to AIDS.

Ethnicity and Sex-risk Level

Table 5 shows sexual orientation and rates of risk by ethnicity. The results suggest that there is a bias against black males being included in the study group because in Los Angeles County the proportion of black males of these ages is considerably higher. Also, a much smaller proportion of black gays interviewed are at high risk compared to whites, and whether this is a reporting bias or a behavioral difference is unknown. Similarly, these findings suggest that there may be an underreporting of homosexual behavior by Hispanics.

A cogent explanation is that the oversampled area with a high density of gays was predominantly white. Thus, minorities had an opportunity, for the most part, to be selected only from the general community segment of the sample. The small number of minorities may be the result of the low probability of being selected because of the 1:14 sample ratio. In general, this ratio is too lopsided; however, decreasing the difference in selection weights would require considerably larger samples than recommended earlier unless a sharp reduction in the number of gays in the study group is tolerable.

Parenthetically, the inordinately high sampling ratio may account for the finding that approximately one in five persons classified as gay because of homosexual experiences during a 2-year period reported heterosexual experiences as well. If this is the case, given the high proportion at risk of being HIV positive, there is considerable reason to be concerned with transmission to the heterosexual population. However, bisexual males living outside the densely populated gay area are weighted 14 to 1, and the addition or subtraction of a few such persons could shift the estimate considerably. This is another reason to be wary of such extreme sample weighting in risk-related surveys.

Drug-related Risk

As part of the survey, an effort was made to obtain information on intravenous (IV) drug use. This sequence began by inquiring if interviewees had ever used IV drugs. Only about 4 percent of the population replied in the affirmative and when asked about drug use in the past 2 years the numbers were reduced to less than 2 percent. These percentages clearly are low compared with impressions of the prevalence of drug use in the county. Moreover, information on needle-sharing was not gathered, because at the time of the survey any history of drug use was deemed sufficient to classify persons at high risk.

Given the illegality of drug use and the stigma attached to it, it is unlikely that reasonable estimates of drug-related risk are possible to obtain in telephone surveys. Moreover, to the extent that the drug-using popu-

Table 5. Sexual orientation and sex-risk level by ethnicity

	Percent of total sample	Percent homosexual	Percent at high or very high sex-risk	
			Homosexual	Heterosexual
White	59.8	6.9	93.1	19.9
Black	9.2	4.5	30.0	25.5
Hispanic	21.3	2.1	91.7	21.2
Asian	8.2	3.3	96.7	15.5

Table 6. Responses to sexual behavior questions[a]

	Percent engaging in behavior		Percent refusing to answer	
	Homosexual	Heterosexual	Homosexual	Heterosexual
French kissing	95.2	89.5	0.6	2.2
Oral sex	89.9	72.6	.3	2.7
Vaginal sex		91.2		1.9
Female anal sex		12.5		2.5
Male anal insertive sex	67.4		.6	
Male anal receptive sex	57.5		.6	
Frequency of sex				
Daily	13.9	7.0	.3	1.5
Several times/week	34.4	54.4		
Weekly or less	51.4	37.0		

[a] Excludes bisexuals

lation is residentially, occupationally, and otherwise unstable, probability of contact in a telephone survey is low. Finally, unlike the gay population, who cluster in particular parts of the county, oversampling is unlikely be a successful strategy to increase the population of interest.

Quality of Data

The quality of data on sexually sensitive items is an additional issue that needs to be considered in any examination of the utility of telephone surveys for collecting risk-related information. Table 6 shows responses to sexual behavior items by sexual orientation. There are relatively similar responses for the two "unisex" items (french kissing and oral sex) although gays are more sexually active than nongays, which is the general impression in the literature. While the refusal rates to these items are low, heterosexuals refuse to answer items somewhat more often than homosexual males. However, data quality does not seem to be an issue, at least in terms of refusal to answer sexually sensitive questions. As shown in Table 7, 96 percent of persons provided answers to the five items included in the survey.

There is reason to be less sanguine about data quality, however, than the above discussion suggests. Figure 1 shows number of partners over 2 years. Among sexually active males, about 13 percent of the heterosexuals and 57 percent of the homosexuals reported four or more partners; some 5 percent of the homosexual and the 2 percent of the heterosexuals reporting 25 or more partners during the 2-year period. Although the number of cases is small, frequency of sexual activity and number of partners was cross tabulated. There is some reason to question whether at least some of these "outliers" were fantasizing. For example, a few among homosexual and heterosexual males report 25 or more partners but sexual activity only once a week or less. These men either have peculiar episodic sex lives, or there is reason to question their reports.

Table 7. Refusals to sexually sensitive questions[a]

	Percent of sample
Not refusing	96.4
Refusing 1 item	1.2
Refusing 2 items	0.4
Refusing 3 items	0.2
Refusing 4 items	1.2
Refusing 5 items	0.5

[a] Five sexually sensitive items for heterosexuals: Have you done French kissing with her? Have you had vaginal sex with her? Have you had oral sex with her? Have you performed anal sex on her? How frequently would you say that you generally have sex with her?

Five sexually sensitive questions for homosexuals: Have you done any French kissing with him? Have you had oral sex with him? Have you performed anal sex on him? Have you had him perform anal sex on you? How frequently would you say you generally have sex with him?

Excludes bisexuals.

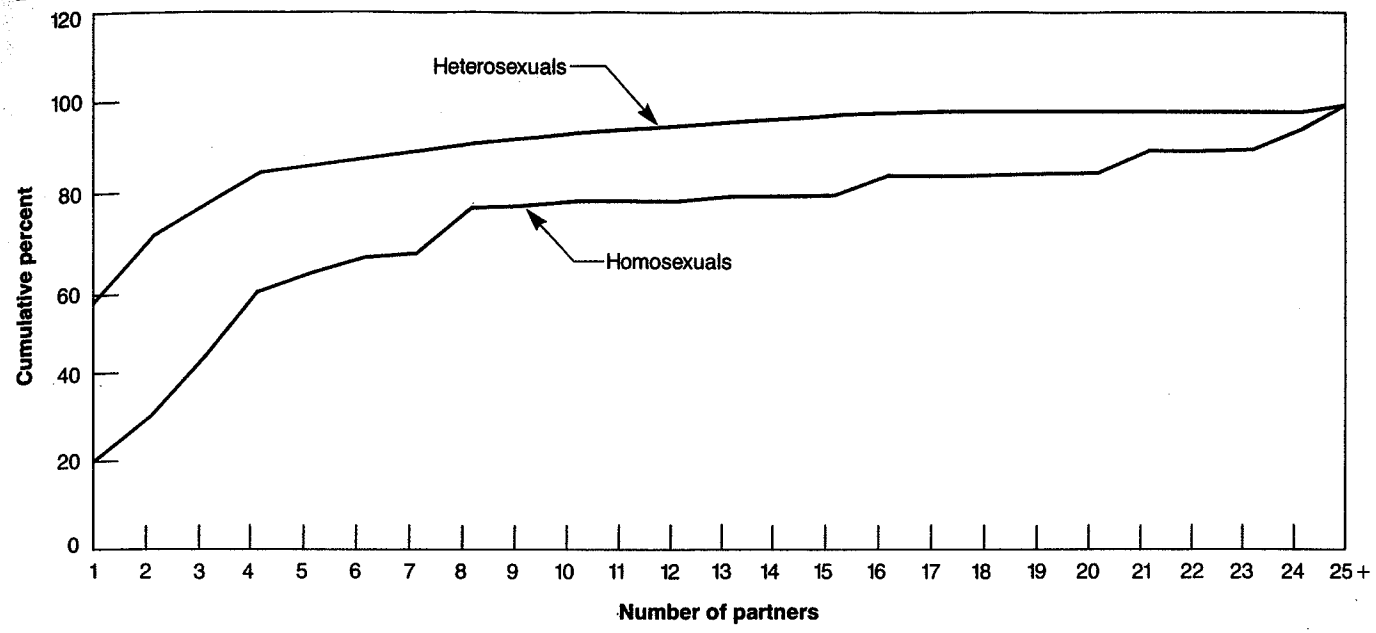
Conclusions

There is no doubt that compared with surveys on more conventional topics, this one is flawed. An inordinate noncooperation rate occurred, and probably a bias against inclusion of minorities in the completed interviews. The estimate of bisexuals in the population is questionable and there is doubt that reports of IV drug use are reasonable. Further, efforts to develop estimates of HIV infection suggest that risk needs to be conceptualized along multiple dimensions and that considerably more information than obtained here is required to appropriately classify individuals. Finally, this sample is much too small to undertake the detailed analyses necessary to identify characteristics of persons at high risk.

In AIDS-related community surveys there is a very special information collection problem. Considerable "technology transfer" occurs, of course, between surveys in other fields and surveys such as this one. However, merely adopting procedures that have proved successful in routine surveys is inappropriate. Rather, the accumulation of wisdom about maximizing the rigor and quality of surveys in this area is critical.

Moreover, the state of basic research and clinical knowledge about AIDS and HIV infection is enlarging rapidly. Thus, sampling frames and contents of surveys will have to be modified continually.

Figure 1. Sexual preference and number of partners over two years



Excludes males sexually inactive for two years

Finally, as much as it offends our methodological sensitivities, we may have to reduce the "standards" we have come to use in appraising traditional surveys. Given the urgent demand for information, more often than we wish, we probably will have to settle for "good enough" efforts. If we do not, we will be unable to provide many data-driven epidemiologic estimates, or any of the empirical information required for the development of preventive and community treatment programs.

References

Burke, D.S., Brundage, J. F., Herbold, J. R., & associates. (1987). Human immunodeficiency virus infections among civilian applicants for United States military service: October 1985-March 1986. *New England Journal of Medicine,* 317 (3), 131-136.

Chaisson, R. E., Moss, A. R., Onishi, R., & associates. (1987). Human immunodeficiency virus infection in heterosexual intravenous drug users in San Francisco. *American Journal of Public Health,* 77 (2), 169-172.

Cleary, P. D., Singer, E., Rogers. T. F., & associates. (1988). Sociodemographic and behavioral characteristics of HIV antibody-positive blood donors. *American Journal of Public Health,* 78 (8), 953-860.

Gottlieb, M. S., Schroff, R., Schanker, H. M., & associates. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: Evidence for a new severe acquired cellular immunodeficiency syndrome. *New England Journal of Medicine,* 305 (21), 1425-1431.

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

Kaslow, R. A., Ostrow, D. G., Detels, R., & associates. (1987). The multicenter AIDS cohort study (MACS). *American Journal of Epidemiology,* 126 (2), 310-318.

Murphy, P. A., & Binson, D. (1988, May). Who says no to whom: Respondent-interviewer interactions in refusals to sensitive questions. Paper presented at the meeting of the American Association for Public Opinion Research, Toronto, Ontario.

Quinn, T. C., Glasser, D., Cannon, R. O., & associates. (1988). Human immunodeficiency virus infection among patients attending clinics for sexually transmitted diseases. *New England Journal of Medicine,* 318 (4), 197-203.

Turner, C. F., Miller H., & Moses, L., (Eds.). (1989). *AIDS: Sexual behavior and intravenous drug use.* Washington, DC: National Academy Press.

Winkelstein, W., Lyman, D. M., Padian, N., & associates. (1987). Sexual practices and risk of infection by the human immunodeficiency virus. *Journal of American Medical Association,* 257 (3), 321-325.

FEATURE PAPER

Sampling and Accessing People with AIDS: A Study of Program Clients in Nine Locations

John A. Fleishman, Joan S. Cwi, and Vincent Mor

Introduction

Over 85,000 cases of acquired immunodeficiency syndrome (AIDS) have been reported to the Centers for Disease Control, and it is estimated that for every case of AIDS there are 10 to 15 people infected with the human immunodeficiency virus (HIV). An epidemic of this size raises important policy issues concerning the organization, delivery, and use of health services. Informed discussion of policy options requires data concerning health practices, preferences for types of care, and demand for services. The opinions and preferences of people with AIDS themselves, who are actual or potential consumers of AIDS-related health and social services, should be an important input into policy analyses. Although a number of general population surveys have included questions assessing knowledge about AIDS and reactions to people infected with HIV, far fewer studies have systematically collected data from people who have been diagnosed as having AIDS or AIDS-related complex (ARC).

As part of an evaluation of the Robert Wood Johnson Foundation's AIDS Health Services Program, The Center for Gerontology and Health Care Research at Brown University, in conjunction with Survey Research Associates, is conducting interviews with people with AIDS-ARC (PWAs) in nine communities: Atlanta, Dallas, Ft. Lauderdale, Jersey City, Miami, Newark, New Orleans, Seattle, and Nassau County (NY). A total of 1,031 interviews with people with AIDS-ARC were completed.

Data collection for this project sheds light on several obstacles in sampling and gaining access to people with AIDS. These obstacles include (1) the comprehensiveness of administrative records; (2) contacting clients through intermediaries, rather than directly; and (3) respondent cooperation and comprehension. This paper discusses these obstacles and describes our efforts at overcoming them.

John A. Fleishman is with the Department of Community Health, Brown University, Providence, Rhode Island. Joan S. Cwi is with Survey Research Associates. Vincent Mor is with the Center for Gerontology and Health Care Research, Brown University, Providence, Rhode Island.

This research was supported in part by Grant No. 12044 from the Robert Wood Johnson Foundation.

Background

The Robert Wood Johnson Foundation began its AIDS Health Services Program in 1987. The Program's goals include (1) developing a coordinated network of agencies providing health and social services to people with HIV-related conditions, and (2) facilitating continuity and comprehensiveness of care via case management of individual clients. Nine program sites in 11 communities were selected from a number of applicants in cities with the highest numbers of reported AIDS cases. In each program site, a consortium of participating agencies has been established; these consortia typically include a community-based organization (CBO), which provides case management and other social services, and a large public hospital, which is the location for an outpatient clinic serving people with AIDS-ARC.

For purposes of evaluating the AIDS Health Services Program, the opinions and evaluations of the actual consumers of services cannot be ignored. Therefore, the evaluation design includes interviews with Program clients. The interview gathers data on health services utilization, satisfaction with health services and with case management services, perceived needs for social services, and preferences for palliative versus aggressive medical treatment.

Management Information System as a Sampling Frame

To facilitate the evaluation of the AIDS Health Services Program, and to permit a description of its client

Table 1. Client characteristics from MIS

Site	Characteristic					
	Percent male	Percent gay or bisexual	Percent IVDU	Percent white	Percent job	Percent alone
New Jersey (Total)	75.1	13.9	63.4	36.0	16.2	21.5
Jersey City Med Ctr.	67.0	9.9	66.8	53.2	9.7	21.8
VA	98.1	7.2	71.0	23.3	22.6	30.3
St. Michaels	62.4	17.1	65.9	43.2	17.2	22.3
Nassau (Total)	78.2	40.0	42.9	71.9	18.1	22.6
CBO	83.3	49.3	32.9	84.4	20.0	27.2
Hospital	69.4	22.1	62.3	50.0	14.6	15.1
Atlanta (Total)	94.2	78.7	8.0	64.5	34.4	44.5
CBO	97.5	88.6	3.4	80.9	44.5	54.8
Hospital	90.1	66.2	13.7	44.0	22.5	32.0
New Orleans (Total)	93.5	77.6	6.9	74.6	26.8	29.7
CBO	96.8	85.5	4.6	81.5	33.2	33.3
Hospital	80.1	68.8	9.5	66.2	20.5	24.2
Dallas	98.7	83.1	3.1	91.0	21.5	36.3
Seattle	95.4	76.9	6.0	89.7	16.6	37.1
CBO	97.5	87.6	2.6	93.5	11.4	38.7
Hospital	93.8	69.3	8.4	86.9	20.1	29.1
Ft. Lauderdale	82.0	53.3	19.1	68.4	16.7	32.2
Miami	74.9	28.3	16.7	48.8	37.1	—
CBO-No Hospital	79.5	29.3	12.1	48.7	79.6	—
Outpatient	67.9	11.2	4.0	52.3	64.3	—

population, a computerized database program was developed and distributed to each site to serve as the basis for a management information system (MIS). For each client, at the time of initial entry into the Program, an intake form is to be completed and entered into the database. The intake form records client background data, including age, gender, race, risk group, and employment status.

In most sites there are multiple entry points for clients, at a minimum the community-based organization and the hospital. Many clients of the community-based organization do not receive outpatient care through the clinic at the public hospital, and many patients at the public hospital's outpatient clinic are not clients of the community-based organization. The intake site for clients—community-based organization versus public hospital—defines important subgroups of clients.

Table 1 presents summary data from the management information system, breaking down the data by intake site. It is clear that, compared to the community agency, clients who enter at the hospital are more likely to be intravenous (IV) drug users, nonwhite, female, and unemployed.

The Program did not require a uniform management information system across all sites. Sites vary in the degree to which MIS data are subject to quality control. Moreover, sites vary in the degree to which a comprehensive management information system has been implemented. Nonuniformity in the definition of a Program client gives rise to a lack of coverage in the management information system. In some sites patients at the outpatient infectious disease clinic are automatically considered to be Program clients and are in the MIS database. In other sites, only those people who received case management are in the management information system; many clinic patients are therefore not represented in the management information system. This is particularly true in Newark and in New Orleans.

Choice of a sampling frame presumes a definition of the population being studied. In the present case, the population of interest is HIV-infected people receiving services from agencies participating in the AIDS Health Services Program. Consequently, defining Program clients as people whose intake data were in the management information system omits significant numbers of people who actually are receiving services in some sites. The lack of uniform client definition across sites reduces the comprehensiveness of the management information system as a sampling frame. Our experience provides a cautionary tale regarding the utility and comparability of databases maintained by separate agencies.

Procedures for Sampling and Respondent Access

Obtaining access to people with AIDS-ARC is difficult, in part because of strong pressures to protect the confidentiality of people with AIDS or HIV-related conditions. Confidentiality restrictions severely limit the possibility of gaining access to people with AIDS-ARC

through a State registry. In any circumstance, it would be awkward to contact people with AIDS-ARC directly, without the mediation of some representative of the health care system.

The strategy in this study has been to work through existing service providers, such as staff at outpatient clinics or community-based agencies. Staff at such agencies have large caseloads of potential respondents. Moreover, agencies participating in the Program have agreed to cooperate with evaluation efforts. Our role as Program evaluators has greatly facilitated access to clients.

Despite our status as Program evaluators, most agencies were unwilling to provide us directly with the names of their clients with AIDS. The study was limited to two basic strategies: (1) working with clients' Program identification numbers from the management information system and (2) selecting clients as they waited for their appointments at the outpatient clinic. In both cases, access to potential respondents was mediated through service providers.

The initial sampling frame was based on MIS data. Each Program client in each site received an identification number at intake, which was recorded in the management information system. In each site, client identification numbers were randomly sampled from the management information system, with the restriction that the client be at least 18 years old and had been served by the Program for at least 1 month. Clients were stratified by intake site (hospital versus community-based organization). In four sites (Atlanta, Dallas, New Orleans, and Seattle) the clientele was predominantly gay and bisexual males; sampling was limited to this subgroup of clients in these sites. In the other sites, the sample was also stratified by risk group—gay and bisexual males, IV drug-using males, and females.

In each site, one person was designated to serve as a liaison with the evaluation. This person was usually the supervisor of client case managers at the community-based agency. Liaisons agreed to receive lists of sampled identification numbers, link the numbers with client names, and distribute the lists of sampled clients to the appropriate case manager. The case manager then called the client and solicited participation in the study. Names and telephone numbers of clients who agreed to participate were given to the on-site research interviewer. Those people who wanted to participate without revealing their names were contacted using a pseudonym.

This procedure has the advantage of affording maximal protection of client confidentiality. The research team sees only the names of people who agree to participate. In addition, case managers usually insisted on having the option to screen out beforehand any client too physically ill or demented to be interviewed.

There are several disadvantages to this procedure: (1) Case managers can remove a client from the sample at their discretion, introducing unknown biases; (2) case managers initially present the study to respondents, potentially increasing the refusal rate compared to a trained research interviewer; (3) poorer respondents in some areas do not have telephones, have moved, or are difficult to locate.

Perhaps the greatest liability of this approach is that direct service staff are already overburdened with their regular duties. Caseloads of case managers range from around 30 to over 200. Large caseloads make it difficult to maintain contact with all clients. Further, direct service providers often view research as an intrusion and an interference with their job performance. Taking time to contact clients for a research interview is often a low priority and may be postponed or done half-heartedly (for example, some failures to locate may result from reluctance to make several telephone calls at different times of day to a single respondent).

Table 2 illustrates this problem. Of the 2,693 identification numbers sent to liaisons across all data collection sites, 46.8 percent had no disposition. This means that direct service staff provided no information about these clients to the field interviewers. The identification numbers, in essence, sat on someone's desk. The number of no dispositions is inflated, in part, because clinic sampling was instituted at most hospitals and the staff there stopped trying to contact specific clients. Nevertheless, losing nearly half the sample through the inaction of intermediaries poses a severe challenge to sampling persons with AIDS-ARC.

Table 2. Disposition of cases by case managers

Site	No. ID's sent to liaisons	Disposition							
		Refused	Too ill	Moved	Not located	Language	Deceased	No disposition	Agreed to interview
Atlanta	400	2.7	2.7	2.7	18.2	3.5	14.7	37.2	18.0
Dallas	318	3.5	3.5	7.5	9.4	6.9	16.3	9.1	43.7
Nassau	391	7.6	2.5	3.6	0.8	0.8	12.8	42.2	29.7
Newark	548	1.6	0.1	2.2	0.7	10.4	14.6	60.5	9.3
New Orleans	247	6.9	5.3	8.1	7.6	6.9	7.3	35.2	14.6
Miami	220	3.2	3.2	1.4	22.3	0.9	5.9	58.1	5.0
Ft. Lauderdale	64	1.6	0	1.6	6.2	6.2	3.1	64.1	17.2
Seattle	324	1.2	0.9	0.9	2.5	0.6	0.3	76.5	17.0
Jersey City	181	1.1	6.1	1.6	16.6	7.7	17.6	44.7	4.4
Total	2,693	3.4	2.5	3.4	8.2	5.0	11.4	46.8	18.5

Table 3. Numbers of cases obtained through MIS, clinic, and flood samples

Site	Sample type		
	MIS	Clinic	Flood
Atlanta	72	150	6
Dallas	139	0	17
Nassau	116	80	8
New Orleans	36	55	33
Miami	11	10	27
Ft. Lauderdale	11	51	0
Seattle	55	0	2
Jersey City	8	34	0
Newark	51	56	43

Initially there was concern that case managers would screen out a large proportion of clients on the basis of illness. However, the actual proportions of clients eliminated from the sample on this basis were smaller than expected. Client mortality was a more serious problem. Substantial numbers of clients (11.4 percent) were deceased by the time the sample was selected; this is a continuing problem in obtaining samples of people with AIDS-ARC, who have a median life expectancy after diagnosis of approximately 18 months.

Client mobility was another problem. Overall, 3.4 percent of the sampled clients were known to have moved out of the area. An additional 8.2 percent could not be located by case managers. It is unclear whether mobility differs by risk group.

Working through case managers resulted not only in substantial nonresponse, but also in very slow sample accrual. Moreover, it became apparent that case managers at the community agency often did not have close contact with many clinic patients. The fact that the management information system did not adequately represent the population of clinic patients has already been discussed. Therefore, in addition to working through case managers at community-based agencies, a second sampling procedure was instituted. To access clients at outpatient clinics that specialize in treating people with HIV illness, names were sampled randomly from lists of patients scheduled for clinic appointments. Clinic staff then identified these people and asked them if they were willing to participate in a confidential interview; those who agreed were introduced to the research interviewer. As with the MIS sample, clinic staff could eliminate patients who they judged to be too ill to participate.

This approach has the advantage of enabling a large number of respondents to be contacted in an efficient manner. On the negative side, patients often do not appear for scheduled clinic appointments; some clinics have waiting areas not exclusively designated for patients with HIV illness, thereby making solicitation for an interview difficult without breaching confidentiality; and many clinics lack space to conduct an interview.

Eventually, to increase respondent accrual it was necessary to institute a third procedure, the flood sample, in which case managers asked clients to participate, without any prior sampling. The hope was that case managers would find it easier to solicit participation in the study during the ordinary course of their contacts with clients. However, the yield from this procedure was less than expected. Although asking direct service providers to make a special effort to contact specific clients increases their workload, it also creates a structure that can establish accountability for client contact. Table 3 shows the numbers of respondents obtained by sampling

Table 4. Characteristics of respondents

Site	Characteristic						
	N	Percent male	Percent gay or bisexual	Percent IVDU	Percent white	Percent job	Percent alone
New Jersey	120	73.3	26.7	63.3	35.8	16.7	19.2
Jersey City	22	68.2	27.3	63.6	63.6	13.6	23.8
Newark	98	74.5	26.5	63.3	29.6	17.3	18.6
Nassau	125	83.2	47.2	44.8	79.2	24.8	25.6
CBO	68	82.4	55.9	35.3	85.3	29.4	20.6
Hospital	57	84.2	36.8	56.1	71.9	19.3	31.6
Atlanta	95	100	96.8	1.1	89.5	24.2	26.3
CBO	51	100	98.0	0	98.0	33.3	33.3
Hospital	44	100	95.5	2.3	79.5	13.6	18.2
New Orleans	54	98.1	85.2	5.6	66.7	27.8	29.6
CBO	23	100	87.0	4.3	78.3	8.7	39.1
Hospital	31	96.8	83.9	6.5	58.1	41.9	22.6
Dallas	103	100	94.2	1.9	95.1	27.2	36.6
Seattle	20	100	100	0	95.2	15.0	65.0
Ft. Lauderdale	19	94.7	89.5	5.3	100	31.6	21.1
Miami	20	85.0	65.0	20.0	80.0	15.0	25.0

Table 5. Comparison of MIS and sample

			Percent male	Percent gay	Percent IVDU	Percent white	Percent employed	Percent alone
MIS	Total	(N=6316)	83.5	53.7	30.1	58.1	24.0	30.9
	CBO	(N=2179)	92.7	77.1	10.0	81.6	33.3	40.2
	Hospital	(N=4081)	78.4	34.3	42.7	45.1	18.8	25.8
Survey	Total	(N=537)	89.6	54.4	18.7	74.7	23.3	28.0
	MIS	(N=343)	91.0	57.1	13.4	80.5	27.4	30.4
	Clinic	(N=153)	84.3	48.4	25.5	65.4	18.3	20.5
	Flood	(N=61)	95.1	54.1	31.1	65.6	13.1	33.3

from the management information system, by recruiting patients at the clinic, and by the flood sample.

A rough sense of the biases introduced by the sampling procedures can be obtained by comparing sample characteristics with MIS data. Table 4 displays, for each site, respondent gender, risk group, race, employment status, and living arrangement analogous to Table 1. These data are based on the 557 cases that have been processed through data cleaning to date. These cases are a mixture of the three sampling procedures.

Table 5 aggregates both the management information system and the interview data across sites. The obtained sample mirrors the pattern of differences between clients of community-based organizations and hospitals. That is, clinic respondents are more likely to be IV drug users, nonwhite, living with others, and unemployed. The fact that the sample reproduces the pattern of differences found in the management information system is reassuring.

Comparison of the magnitudes of corresponding entries for the total management information system and the total sample shows that, in general, the obtained sample has a lower proportion of IV drug users and a higher proportion of whites. Apparently, gay, white clients are more likely to be accessible and cooperative. The proportions of clients who are gay, live alone, or are employed are similar in the overall management information system and in the survey. Comparing management information system data from clinic clients with the clinic sample (Table 5, rows 3 and 6) reveals a similar pattern. Comparing MIS data from the community-based organization with respondents sampled through the management information system, who are predominantly from the community-based organization, shows that the sample had a lower proportion of clients who were gay, employed, and lived alone. Clients who are employed may be less accessible than those who are unemployed, since the latter may have more available time to be interviewed.

Client Cooperation and Comprehension

Once contacted, most clients were willing to be interviewed. Only 2.3 percent refused to participate when asked by the case manager. When subsequently recontacted by the field interviewer, 3.4 percent changed their minds and refused.

Another indicator of client cooperation is reflected in willingness to sign informed consents. To protect respondent confidentiality, and to avoid refusals based on respondents' reluctance to disclose their names, the interview was conducted anonymously whenever a client so preferred. Before beginning the interview, respondents were read and given a copy of a form outlining their rights and assuring them of confidentiality. To preserve anonymity, respondents did not sign the informed consent form.

At the end of the interview, after rapport between respondent and interviewer had developed, respondents were asked to sign two new informed consents. One asked for permission to recontact the respondent within a year for a second interview; the other asked for permission to examine medical records. Signing the consent forms precludes anonymity. Overall, 95 percent of respondents agreed to be recontacted, and 85 percent agreed to permit access to medical records. These data show that, overall, people with AIDS are willing to participate in research interviews and to release their names, with appropriate assurances of confidentiality.

The procedure of administering the interview anonymously and subsequently asking for further informed consents has proven to be very valuable. Compared to administering all consent forms before the interview, the current procedure not only avoids losing respondents who wish to be anonymous, but it also presumably increases the proportion of those who sign consents.

Comprehension. The HIV attacks the central nervous system, producing AIDS-related dementia. Potential cognitive impairments must be taken into consideration when surveying people with AIDS. For example, people with dementia may be unable to give accurate responses to questions concerning utilization of health services over a period of several months.

As noted, direct service providers could screen out any person who they felt was too cognitively impaired to participate. This occurred infrequently. As a further check for cognitive impairment, respondents were given a brief cognitive screen, adapted from the Mini Mental State Exam.

Preliminary analyses of these data have been completed for 297 respondents. Only four people could not immediately recall three words. Sixty-two (20.7 percent) people had one or more errors in delayed recall of the three words. Fifty-two people (17.4 percent) made one or more errors in naming the months of the year in

reverse order. Finally, respondents were asked to perform a series of simple subtractions (that is, subtract 7 from 100, 7 from the remainder, and so on). Forty-nine percent made one or more subtraction errors.

These data suggest that few respondents had such severe cognitive impairment that their interview could not be completed. Most respondents exhibited no gross cognitive dysfunction. However, a percentage of respondents (at least 15 percent) may be suffering from some impairment.

Conclusion

It is extremely difficult to achieve a probability sample of people with AIDS. Confidentiality requirements virtually prohibit developing a sampling frame from a register with people's names and addresses. Some sort of catchment-area-specific random digit dialing sample could be attempted, but such a procedure might underrepresent drug users with no telephone or stable residence. Given the necessity of protecting client confidentiality, access to people with AIDS is perhaps best done through direct service providers. Clients often have learned to trust nurses and case managers and will respond positively if a study has their approval.

However, working through direct service providers introduces substantial problems of nonresponse, and it is important to determine what biases this introduces. Contrasting characteristics of the obtained sample with those of agency clients in general, as reflected in the management information system, is one approach to estimating sampling bias. Once fieldwork is complete, MIS data from respondents and nonrespondents will be compared to get another perspective on sampling bias. Clearly, further work is needed on assessing the magnitude of bias introduced when one tries to contact people with AIDS through direct service providers.

FEATURE PAPER

Area Samples of Male Street Prostitutes Richmond, Virginia, 1988

Judith Bradford and Scott Keeter

Background

Despite the spread of human immunodeficiency virus, (HIV) infection throughout the intravenous drug-using and general populations, men who have sex with other men continue to be the group most likely to be infected. (Virginia Department of Health, 1989). Because of the incidence of seropositivity within the gay male community, male prostitutes who provide service to other men are a group who need to protect themselves from infection and to prevent the transmission of HIV to their customers. Although no data could be located about HIV seroprevalence among male prostitutes in the United States, a 1986 study of 50 male prostitutes serving other men in Amsterdam found that 13 percent were seropositive, a rate comparable to that of the city's gay male population as a whole. (Coutinho & associates, 1988) Eighteen percent of the Amsterdam sample were also IV drug users. There is a high rate of intravenous drug use and HIV infection among female prostitutes in the United States (Becker & Joseph, 1988); and if this is also true for males, a second probable route of HIV infection exists for this population.

The importance of providing education about HIV risk reduction methods to individuals who engage in high risk sexual and drug use behaviors cannot be overstated. Uninfected persons cannot count on the physical appearance of potential partners to provide trustworthy evidence of health. Nor can it be expected that all or even most of those who are infected will be aware of their HIV status. Many individuals still have little to go on when making decisions about who to have sex with or whether it is necessary always to use new syringes when injecting drugs. In the absence of certainty about what to do and when to do it, many individuals can be expected to use risk reduction methods sporadically or to initiate change that they cannot maintain. (Becker & Joseph, 1988; Emmons & associates, 1986).

However, findings from a number of behavioral and psychosocial studies indicate that groups of gay men, female prostitutes, and IV drug users are eager for risk reduction information (Barton, & associates, 1987; National AIDS Network, 1987; Stein & Bransom, 1987) and have adopted new behaviors in an effort to protect themselves and their partners from transmission of the virus (Becker & Joseph, 1988; Wartzman, 1987; Watters & associates, 1988; Bradford & Honnold, 1988). Knowledge has been found to be a consistently important variable in these studies, with education about AIDS and HIV described as the most important predictor variable for protective behavior change (Emmons & associates, 1986). When adequate attention is paid to the collection of sound data about behavioral responses, results can be trusted to reflect what is happening with high risk groups. (Sisk & associates, 1988; Communication Technologies, 1987). Valid results provide a rationale for investing resources in educational outreach and the distribution of risk reduction materials, particularly when these are targeted to the demonstrable needs of individuals within a planned area of intervention.

Specific methods have been found to be helpful in planning and implementing educational outreach to special populations. It is of primary importance that health educators and other outreach workers understand the environment in which a targeted group operates. A number of studies with drug users have made successful use of this approach (National Academy of Sciences, 1989; Williams, 1986; National AIDS Network, 1987; Wartzman, 1987). These efforts all emphasize that such special populations all must be assured that those with whom they deal can be trusted.

Others have attested to the value of a social exchange approach, or "value given for value received" (Alexander & McCullough, 1981; Kotler & Levy, 1969). This

Judith Bradford and Scott Keeter are with the Survey Research Laboratory, Virginia Commonwealth University, Richmond, Virginia.

Data for this paper were collected as a part of the Virginia Statewide AIDS Needs Assessment, conducted for the Virginia Department of Health, with funds provided by the Centers for Disease Control.

approach has been valuable in the Pitt Men's Study, where gay men were successfully recruited for research participation through the use of "social marketing strategies" (Silvestre & associates, 1986) as well as in the New Orleans project with male prostitutes (Watters & associates, 1988). Variations have been applied in several projects with IV drug users, where research participants have been rewarded with money, information, bleach, paraphernalia, or a combination of these.

If possible it is also important to tailor educational and risk reduction efforts to the directly assessed needs of the target group. Street outreach methods have been shown to be quite cost effective (Watters & associates, 1988; Newmeyer, 1988) and have been used to recruit participants for research and evaluation projects (H.J. Osofsky, personal communication, March 1989; Watters & associates, 1988) or for on-site data collection, as in Richmond. Using proven outreach methods to accomplish needs assessment and monitor change can also be a cost-effective way to collect descriptive information about targeted groups. Needs data gathered in this manner can be used as a basis for program planning as well as for facilitating interaction with members of the target group. These data can also be quite useful in soliciting program funds if a need for intervention can be demonstrated.

The project reported here was designed for a population of particular concern to the city of Richmond in its efforts to prevent the transmission of HIV.[1] It is part of a larger study being conducted by Virginia Commonwealth University's Survey Research Laboratory on behalf of the Virginia Department of Health AIDS Program, and will assess the knowledge, attitudes, and HIV risk behaviors of individuals throughout the State. The purpose of the study reported here has been to provide information to the Virginia Department of Health about how to allocate its HIV educational resources, as well as to monitor the effectiveness of funded programs, when appropriate. A major methodological interest has been the collection of data from various groups difficult to reach through ordinary means but who are essential links in the chain of HIV prevention, such as street prostitutes, drug users, and the homeless.

During the first year of data collection from street populations, primary emphasis was placed on learning how to collect valid data and on exploring the extent of data collection that could be accomplished. During the second year of data collection, revised instruments for recording survey data have been developed, follow-up data are being collected in Richmond, and projects are being initiated in two additional urban areas of the State. This paper explores what has been learned from the first year's effort and presents baseline data collected from male street prostitutes during the spring of 1988 in Richmond.

[1]The projected 1990 population for Richmond is 214,300). Although CDC reports fewer than 100 cases of HIV infection assigned to Richmond, from 350 to 400 patients are actually being treated (Charlotte Syran, personal communication, March 1989), and a growing proportion of gay men in central Virginia are reportedly seropositive (Bradford & Honnold, 1988).

Methodology

The project design was twofold: First, to collect descriptive information about the HIV-related needs of male street prostitutes in the city, and second, to establish educator and advisor relationships with them, through which their behavior could be influenced. This information and influence will be used to design and implement an effective plan for educational outreach. In addition, a transferrable store of knowledge and methodological expertise will be developed which we can be shared with groups interested in doing similar work in other urban areas of the State. The primary feature of the design was a merger of data collection and educational outreach, through interviews that served to fulfill both functions at once.

Data were collected on the streets where participants work, and during their working hours. All but five data collection interviews were conducted by the director of Richmond Street Outreach Project (RSOP), a social worker with extensive experience in street work, and provided an opportunity for participants to discuss HIV and acquired immunodeficiency syndrome (AIDS) with him. A second RSOP staff member was also present, to provide assistance with outreach activities and to receive on-site training from the director. During these conversations, participants were asked a standard set of questions about themselves and about the variables of interest. Information collected in this manner was intended to establish a baseline of information about the knowledge, attitudes, and risk-associated behaviors of this population. Two serendipitous purposes were also achieved: A demographic portrait of male street prostitution in Richmond was constructed and research staff developed confidence in their ability to use the data collection process as a mode of intervention with this population. Limited resources available for this effort were thus expended in a manner that had a positive effect at several different levels: Participants received needed information and risk reduction materials (condoms, bleach, and printed materials); funding sources and other political constituencies received information about the population and its practices; and a rationale for continued outreach and sharing of prevention materials was established.

Other Studies of Male Prostitutes

Certain aspects of this study are similar to those of other projects identified, although no directly comparable effort is known. Only one published report of a similar data collection with male prostitutes was identified through a systematic search of the literature (the Amsterdam study mentioned above) (Coutinho & associates, 1988). Two other studies of male prostitute samples are known to be in progress, one in New Orleans (Osofsky, 1989) and a second in Atlanta (personal communications, E. Morse, New Orleans; K. Elifson, Atlanta). Each of these two studies has made use of street contacts to recruit research participants, but in all cases subjects have reported to a clinic setting to be

Table 1. Methodology of male prostitute studies

Study	Method of sampling	Sample size	Representativeness
Bradford & Keeter (this paper)	Area population	95	98%
Coutinho & associates	Convenience, in brothels	50	Unknown
Boles & Elifson	Snowball	240	High
Osofsky & associates	Snowball	240	High

interviewed. Research protocols for these studies have been more extensive than those reported here, and have included the collection of blood samples for laboratory analysis, making street data collection inappropriate.

Methodological details of the New Orleans, Atlanta, and Amsterdam studies are compared in Table 1 with these findings. This comparison emphasizes the unique characteristics of the work accomplished in Richmond. The other studies were not designed as educational projects; their similarity to the reported study has been in the selection and recruitment of participants.

Projects with Similar Goals but Different Populations

Street outreach as an educational approach has been used in several cities other than Richmond, where the target groups have been IV drug users, youth at risk, and minority group members (Friedman & associates, 1986; Williams, 1986; National AIDS Network, 1987; Wartzman, 1987; Watters & associates, 1988). Other street outreach programs have also provided information to male prostitutes (personal communications, J. Izzo, Washington, DC; B. Jones, Baltimore). This approach has proved to be successful as an educational method and it seemed reasonable to expect it would provide a feasible method for data collection, as well.

Community Preparation

Early in program planning, Richmond Street Outreach Project and Survey Research Laboratory (SRL) staff informed the City Police Department about the Project and described what was planned. A letter was sent from the Needs Assessment Principal Investigator to the Chief of Police, explaining the validity of these activities and asking for official support from the department. The RSOP director spoke with several police department officials on the telephone and met with a senior officer in person. Written materials about the Project were distributed to the department.

Outreach staff members were equipped with picture identification cards, which gave two emergency numbers, and with information about the Project and its connection with the Survey Research Laboratory. In addition, staff carried with them during working hours a copy of a letter from the SRL director, which acknowledged their affiliation with the university. These materials were carried in the back pocket of staff clothing, to be accessible if needed for identification to police but not obvious to others. Police officers have not interfered with the outreach workers and no problems of this nature have ever arisen.

Training and Orientation

The importance of careful staff selection and adequate training cannot be overemphasized when work will take place in a natural environment, such as the streets. Street work requires that staff be able to interact with individuals who live within a subcultural context and who may have little contact with ordinary social institutions. To do this work, staff must be accepting of the individuals they interact with on the street and must understand as much as possible what happens in that context. To be effective, street outreach workers must be able to establish their own credibility and be able to build rapport with members of the targeted population (Friedman & associates, 1986).

Initial training in outreach methods and appropriate orientation toward the targeted populations were received from staff of the HERO Project in Baltimore and from the Whitman Walker Clinic in Washington, DC. Indigenous workers were hired or accepted as volunteers; most staff who worked in the male prostitution areas were gay men. This strategy has been recognized as valuable when the primary focus is on educational efforts and behavioral influence and staff can be thought of as positive role models for those with whom they work (Friedman & associates, 1986; National Academy of Sciences, 1988).

Education, Outreach, and Data Collection

RSOP activities began with establishing a presence for staff on the streets where outreach and data collection would take place. Key prostitution areas were identified by gay men on the staff and others who work with the Survey Research Laboratory; two areas were selected for intensive coverage because they were the locations where most activity takes place. One of the intensive areas is worked primarily by men and the second by transpersons, or cross-dressers. Four other areas of the city are also sites where male prostitutes work; but onsite observation of these areas during the first 3 months of program operation revealed very little activity and no regular individuals who did not also work in the two primary areas. As a result, RSOP outreach was focused in the two main areas, with periodic visits to provide educational and risk reduction materials in the others. Data collection took place primarily in the two main areas.

Staff spent the first 3 months of program operation learning the environment and getting to know the prostitutes who worked there. From the beginning, staff were very open about their goals for being on the street.

Table 2. Population characteristics N=95

		Gay	Bisexual	Heterosexual	Male	Transpersons
White	56%	75%	23%	2%	62%	38%
Black	44	60	21	19	88	12

Age		Education		Marital status	
Under 21	28%	Less than h.s.	27%	Single	88%
21-25	35	H.S. graduate	52	Married	3
26-30	23	1-4 yrs. college	15	Divorced	8
30+	14	Graduate degree	1		
Children		**Likelihood of Getting AIDS**			
Any at all	7%	No chance	19%		
Under 6 yrs.	3	Little chance	37		
		Some chance	33		
		High chance	10		
		Don't know	1		

They carried educational literature and risk reduction materials and gave these to anyone who would take them. They attempted to engage prostitutes in conversation about AIDS and answered questions about testing, symptomatology, and services for those who are infected.

Some uncertainty was expressed from the beginning about the process of data collection. Because staff had no previous experience of this kind, they were concerned that asking questions would be perceived as inappropriate and would interfere with the relationships they were building on the streets. Great care was taken to address these legitimate concerns and to establish reasonable criteria by which to evaluate the quality of data collection methods. Staff were encouraged to believe that the key to collecting valid data would be their own ability to interact with those they interviewed and to enter into a mode of questioning that resembled a dialogue rather than a question-and-answer session. The value of this attitude has been recognized by other researchers, because it allows the interviewer to be vigilant for contradictions and to explore answers that might otherwise have to be recorded as missing or incomplete data. (Huang & associates, 1988). Participants in these interviews become active learners, (personal communication, C. Syran, March 1989) a benefit that was immediately appreciated by staff.

Data were recorded by the interviewer on a 4" x 6" sheet of paper that was then folded and kept in the interviewer's back pocket. Completed records were turned in to the Survey Research Laboratory the next morning. Data collection took place on alternating nights, during a 6-week period, typically between 11 p.m. and 3 a.m.[2] During this time, the interviewers approached every individual whose behavior indicated he was being paid to have sex and attempted to engage him in conversation about the needs assessment topics. Interviewers usually waited until they had seen an individual on more than one night before approaching him for an interview. If this attempt failed on a given night, the interviewer would try again at a later date, until successful or until the individual declined to be involved.

Sampling

Initial plans for data collection called for the development of a sampling strategy to be determined on the basis of observation by staff during the first 2 or 3 months of program operation. Options considered included the selection of a subset of prostitution areas, some way of selecting certain individuals but not others, or confining data collection to a narrow time period. However, it became apparent during the observation period that the number of prostitutes working the streets was small enough to permit interviewing everyone, given the fact that staff would be working several nights per week for 6 weeks. The concentration of activity and individuals in two major areas dictated that both be covered; by doing so, staff felt confident that they would reach virtually all male prostitutes who were working the streets regularly in Richmond during the 6 weeks selected.

As a result, no sampling was attempted. With the exception of two men who did not want to be included, most of the regular population of male prostitutes working in known prostitution areas of Richmond during this time were interviewed for the needs assessment. Nineteen other individuals who were not prostitutes were also interviewed in the same areas of town, because RSOP staff did not want to be perceived as interested in prostitutes only. These cases were excluded from the analysis reported in this paper.

Among the 95 males who participated as respondents in the data collection, about 20 (21 percent) were observed to be working almost every night RSOP staff were in the area. Thirty-two percent were observed to be on the streets fairly frequently, perhaps several nights a week, and another 30 percent were there less often but regularly. The remainder of those interviewed (about 17 percent) were observed only once or several times during the 6-week period. If these 95 individuals

[2]On some occasions, data collection followed a two days on, two days off schedule; scheduling was also affected by weather conditions.

Table 3. HIV antibody testing and AIDS/HIV knowledge

	Yes	No	Don't know
Condom use can reduce the chance of getting AIDS	100	—	—
Can catch AIDS by sharing needles	99	—	1
Can be infected without looking sick	100	—	—
Can transmit the virus without looking sick	100	—	—
Using bleach to disinfect works can make you sick	3	67	30
Cleaning needles with bleach kills the virus	67	3	30

Know about the test?		Been tested?		Results		Know where to go for testing?	
Yes	99%	Yes	43%	Negative	88%	Yes	81%
No	1	No	57	Don't know	3	No	18
				Positive	10		

do in fact make up a regular population of male street prostitutes, they would represent a rate of one per 2,263 individuals in the city population as a whole.

Results

Population Characteristics. (Table 2). Nearly two-thirds of those interviewed were 25 years old or younger; 35 percent were between 21 and 25, and 28 percent were under 21. Relatively few (14 percent) were more than 30 years old. A majority (67 percent) had graduated from high school, and 16 percent had at least some college experience. Almost all (88 percent) were single. Seven percent had children, 3 percent had children under 6 years old.

The sample was racially diverse, with 56 percent white and 44 percent black; this breakdown is roughly comparable to that of the city as a whole, which is 48 percent white and 51 percent black. Overall, whites and blacks were demographically similar but they did vary with respect to sexual orientation and gender identity. Whites were more likely to describe themselves as primarily gay and three times as likely to be cross-dressers (or transpersons).[3]

Participants were about evenly split on the question of their own likelihood of becoming infected. Nearly one in five thought they had "no chance" of getting AIDS, and 37 percent thought they had only a "little chance." Forty-three percent thought they had either some chance (33 percent) or a high chance (10 percent). Overall, those interviewed were not particularly concerned about becoming infected.

Knowledge of AIDS and HIV and Antibody Testing. (Table 3). Almost everyone gave correct answers to four basic questions about how the virus is transmitted. However, one third either did not know or gave wrong answers to two basic questions about safe needle use.

All but one participant also knew about the HIV antibody test, although 18 percent did not know where to go to be tested. Fewer than half (43 percent) had actually had the test. Eighty-four percent (36 individuals) who had been tested shared the results of their test; 35 were negative and 1 did not know the outcome. There was a statistically significant relationship between age and having been tested. Prostitutes who had been tested were older than those untested.

Sexual Behavior. (Table 4). Participants were asked about their experiences with five specific types of sex and how often they used condoms for each practice. Responses to these questions indicated a general awareness of which behaviors are most likely to result in HIV transmission. Reported behaviors in descending order from most to least common were these: insertive oral sex (practiced by 98 percent), receptive oral sex (82 percent), insertive anal sex (76 percent), receptive anal sex (68 percent), and vaginal intercourse (30 percent). These data were accepted as credible by RSOP staff.

Receptive anal intercourse, considered to be the most likely of the five practices to provide a transmission route for HIV, was reported as the least commonly practiced behavior (except for vaginal intercourse) and the one most likely to involve condom use. Sixty-eight percent said that they engage in this practice, and 89 percent of those who do so said condoms are always used when they are the receptive partner in anal sex. By contrast, almost everyone (98 percent) reported engaging in insertive oral sex, but only 35 percent said that they always use condoms when doing this. Thirty-nine percent of those who do insertive oral sex never use a condom when engaging in this practice.

When condoms are used by these male prostitutes, it is often at their initiation. Forty-two percent reported that they always suggest using a condom when they are paid to have sex; and an additional 53 percent reported that they suggest condoms at least sometimes. Five percent reported that they never suggest using a condom.

Many prostitutes who were interviewed were able to identify circumstances in which they would be willing to have sex without a condom. The most frequent reason given was simply not having condoms available: 22 percent have unprotected sex for this reason. Eighteen percent have unprotected sex if they are paid more to do so; 16 percent if they know the partner; 14 percent when a partner refuses to use a condom; 8 percent if a partner looks healthy; and 7 percent if a partner says he tested

[3]Participants were not asked if they had had surgery to change their gender. Therefore, it is not known how many "transpersons" were actually transsexuals, if any. These individuals were classified on the basis of their mode of dress and presentation of themselves as women.

Table 4. Sexual behavior

	Yes	Use condom? Sometimes	Always	Never
Do you have . . .?				
Vaginal intercourse	30%	32%	36%	32%
Receptive anal	68	11	89	—
Insertive anal	76	19	79	1
Receptive oral	82	25	44	30
Insertive oral	98	26	35	39
How often do *you* suggest using a condom?		53%	42%	5%

Do you have sex without a condom, if . . . ?	Yes
You have no condoms	22%
Paid extra	18
Partner is someone you know	16
Partner refuses to use a condom	14
Partner looks healthy	8
Partner says he tested negative	7
Changed since AIDS?	17

negative for HIV. Since interviewers did not distinguish among types of sex when they asked if respondents would perform for extra money, it cannot be determined from the data which behaviors respondents would participate in if paid more. Fewer than one in five (17 percent) of those interviewed reported making changes in their sexual behavior since learning about AIDS.

Safer Sex Practices. (Table 5). Nearly everyone interviewed (97 percent) had "heard about safe sex." Almost as many reported that they practice safe sex. When asked to describe what they do to have safe sex, 73 percent of the sample volunteered that they use condoms. Whites were much more likely than blacks to say that they use condoms as a way of having safe sex; 85 percent of whites said this, but only 58 percent of blacks did so. Conversely, blacks were more likely than whites to say that they avoided anal sex in order to protect themselves from infection. Twenty-two percent of blacks reported having only oral sex with men, and 25 percent reported that they only have insertive and not receptive oral sex. White men were much less likely to have only oral sex or to decline to give oral sex.

Intravenous Drug Use. (Table 6). Eleven percent of those interviewed reported using IV drugs. All of these used cocaine and about one-third used heroin. About 8 out of 10 drug users reported going to shooting galleries, and about 3 out of 10 had been in treatment for drug abuse. Twenty-one percent of male prostitutes interviewed accepted vials of bleach from RSOP staff, many "for a friend."

Table 5. Safer sex practices

	Yes		
Heard about safe sex?	97%		
Practice safe sex?	92		
What do you do to have safe sex?	All	Whites	Blacks
Use condoms	73%	85%	58%
With men, have only oral sex	13	7	22
Don't *give* oral sex	12	2	25
Masturbate	9	9	8
Look for open sores	2	2	3
Rinse with hydrogen peroxide	2	2	3
What keeps you from having safer sex?			
No condoms	6	7	4
Good looking partner	3	5	—
Don't like condoms	6	2	11

Summary

Data were collected from a street population of male prostitutes about their knowledge of AIDS and HIV transmission, sexual and drug use behaviors, and safe sex practices. These data were gathered during the course of conversational interviews conducted by staff of the Richmond Street Outreach Project, which provides educational outreach and distributes risk reduction materials to street populations in the city.

All male street prostitutes who were working regularly in known prostitution areas during the 6 weeks data collection period were interviewed. Participants were interviewed during their working hours, on the streets where they work. Along with being interviewed, participants were provided with educational literature and risk- reduction materials, including condoms and bleach. They were encouraged to ask questions of RSOP staff members and to discuss any concerns they had about HIV infection and AIDS. If asked, staff shared information about where to go for testing and counseling.

Table 6. Intravenous drug use

	Yes		
All participants			
Accepted bleach?	21%		
Use IV drugs?	11		
Use cocaine?	11		
Use heroin?	3		
IV drug users			
Go to shooting galleries?	80		
Been in treatment?	30		
Changed since AIDS?	30		
	Sometimes	Always	Never
Share needles or works?	60	—	40
Use bleach to disinfect?	10	50	40

The RSOP staff were well received by the targeted population. Interviewers spent 3 months on the streets before they began to collect data for the baseline needs assessment and gained acceptance from the individuals they were working with. The RSOP staff have continued to provide outreach on the streets and are often approached by prostitutes whom they have not met but who have been referred by others. Staff are asked to provide risk reduction materials and are often engaged in conversation by those who work the streets.

Descriptive information gathered during the first wave of data collection in 1988 was useful in constructing a picture of street prostitution among males in Richmond. In addition, this information was quite helpful to the Richmond Street Outreach Project in knowing what interventions were needed and in providing a justification for continued educational outreach. Although many prostitutes who participated reported frequent condom use, it was clear from the data that most do not believe themselves to be at risk. The needs assessment indicated that many if not most male prostitutes in Richmond are taking steps to prevent transmission of HIV.

Male prostitutes are exposed to considerable risk of infection with HIV, and it is important that they receive the information and materials they need to change risky behaviors. An ethical approach to this problem should be based on the acceptance of prostitution as a fact of life and a pragmatic approach devised. Such an approach would make use of a regulatory model rather than a punitive one and would take into consideration the right of adults to make their own decisions about behavior (Alexander & McCullough, 1981). Given the expressed interest of this population in acquiring information and risk reduction materials, their needs should be given greater priority within state and Federal funding decisions. (National Academy of Sciences, 1989; Rosenberg & Weiner, 1988; Huang & associates, 1988).

Street outreach techniques have proven useful in establishing relationships with street prostitutes and in facilitating the collection of information about their behavior and needs. These relationships have enabled sound information about this special population to be made available to those who are responsible for preventing the spread of HIV and who control the resources necessary to provide education and the distribution of risk-reduction materials. These relationships can be used on an ongoing basis to influence protective behavior change among a population at special risk of infection with HIV.

References

Alexander, K., & McCullough, J. (1981). Application of marketing principles to improve participation in public health programs. *Journal of Community Health,* 6, 216-222.

Barton, S. E., & associates. (1987). Female prostitutes and sexually transmitted diseases. *British Journal of Hospital Medicine,* July, 34-39.

Becker, M. H., & Joseph, J. (1988). AIDS and behavioral change to reduce risk: A review. *American Journal of Public Health,* 78(4), 394-410.

Bradford, J., & Honnold, J. (1988). *AIDS-related knowledge, attitudes and behavior of Virginia gay and bisexual men:* (The Virginia statewide needs assessment. Report prepared for the Virginia Department of Health, Office of Epidemiology, AIDS Program). Richmond, VA: VCU Survey Research Laboratory.

Communication Technologies. (1987). *A report on: Designing an effective AIDS prevention campaign for San Francisco: Results from the fourth probability sample of an urban gay male community.* Prepared for the San Francisco AIDS Foundation. San Francisco, CA: Author.

Coutinho, R. A., van Andel, R. L. M., & Rijsdijk, T. J. (1988). Role of male prostitutes in spread of sexually transmitted diseases and human immunodeficiency virus. Letter. *Genitourinary Medicine,* 64 (3), 207-208.

Emmons, C-A., Joseph, J. G., Kessler, R. C., & associates. (1986). Psychosocial predictors of reported behavior change in homosexual men at risk for AIDS. *Health Education Quarterly,* 13(4), 331-345.

Friedman, S. R., Des Jarlais, D. C., & Sotheran, J. L. (1986). AIDS health education for intravenous drug users. *Health Education Quarterly,* 13(4), 383-393.

Huang, K. H. C., Watters, J. K., & Case, P. (1988). Psychological assessment and AIDS research with intravenous drug users: Challenges in measurement. *Journal of Psychoactive Drugs,* 20(2), 191-195.

Kotler, P., & Levy, S. (1969). Broadening the concept of marketing. *Journal of Marketing,* 33, 37-44.

National Academy of Sciences. (1989). *AIDS: Sexual behavior and intravenous drug use.* Washington, DC: National Academy Press.

National Academy of Sciences, Institute of Medicine (NAS/IOM). (1988). *Confronting AIDS—Update 1988.* Washington, DC: National Academy Press.

Newmeyer, J. A. (1988). Why bleach? Fighting AIDS contagion among intravenous drug users: The San Francisco experience. *Journal of Psychoactive Drugs,* 20(2), 159-163.

Ngugi, E. N., Simonsen, J. N., Bosire, M., & associates. (1987, June). Effect of an AIDS education program on increasing condom use in a cohort of Nairobi prostitutes. In: *Abstracts from the III International Conference on AIDS* (p. 157). Washington, DC: U.S. Department of Health and Human Services and the World Health Organization.

Rosenberg, M. J., & Weiner, J. M. (1988). Prostitutes and AIDS: A health department priority? *American Journal of Public Health,* 78(4), 418-423.

Silvestre, A., Lyter, D. W., Rinaldo, C. R., & associates. (1986). Marketing strategies for recruiting gay men into AIDS research and education projects. *Journal of Community Health,* 11(4), 222-232.

Sisk, J. E., Hewitt, M., & Metcalf, K. L. (1988). The effectiveness of AIDS education. *Health Affairs,* Winter, 37-51.

Stein, J. B., & Branson, B. M. (1987). New AIDS prevention strategies for the I.V. drug user. *Focus: A Guide to AIDS Research* 2(10), 1-3.

Street outreach: A new approach (1987). Washington, D.C.: National AIDS Network, 1(4), 1-2.

Virginia Dept. of Health. (April 1989). Acquired immunodeficiency syndrome (AIDS) surveillance report.

Walters, L. (1988). Ethical issues in the prevention and treatment of HIV infection and AIDS. *Science,* 239, 597-603.

Wartzman, R. (1987, November 4). Street-wise teaching tries to stop AIDS in the inner city. *New York Times,* p. 20.

Watters, J. K., Case, P., Huang, K. H. C., & associates. (1988). HIV seroepidemiology and behavior change in intravenous drug users: Progress report on the effectiveness of street-based prevention. In *IV International Conference on AIDS, Stockholm, Sweden.*

Williams, L. S. (1986). AIDS risk reduction: A community health education intervention for minority health risk group members. *Health Education Quarterly,* 13(4), 407-421.

FEATURE PAPER

Developing a Probability Sample of Prostitutes: Sample Design for the RAND Study of HIV Infection and Risk Behaviors in Prostitutes

Sandra H. Berry, Naihua Duan, and David E. Kanouse

Introduction

This paper outlines a preliminary sampling plan for a study of human immunodeficiency virus (HIV) infection and risk behaviors of Los Angeles prostitutes that will be carried out by The RAND Corporation. This study is now in the design phases. Pilot testing is scheduled for summer 1989 and fieldwork for fall 1989 through spring 1990.

Background

At present, acquired immune deficiency syndrome (AIDS) cannot be cured and no vaccine for preventing it has been developed. Consequently, efforts to contain the epidemic must emphasize changing behaviors that allow transmission of HIV, the AIDS-causing virus. The behavior of female prostitutes may significantly affect the epidemic's future, particularly its potential for spreading through heterosexual contact. Yet their behavior has been little studied and is poorly understood.

A study is being designed that will contribute to general understanding of heterosexual transmission by focusing on female prostitutes, their characteristics and behaviors, and the role they may play in the epidemiology of AIDS. Specific aims will include:

1. Developing numerical estimates of the size of the prostitute population in a large metropolitan area and of its distribution according to predominant mode of soliciting customers (street, out-call, massage parlor, escort service, brothel, etc.).
2. Characterizing prostitute career patterns
3. Performing HIV antibody testing to determine the extent of HIV infection in this population and how this varies by mode of solicitation.
4. Measuring the prevalence and incidence of specific risk behaviors (sexual and drug-related) that can transmit HIV infection.
5. Measuring the type and frequency of preventive behaviors (using condoms, disinfecting needles).
6. Examining the relationship between HIV antibody status, prostitute characteristics, and risk and preventive behaviors.
7. Estimating the numbers and percentages of specific sexual acts, both protected and unprotected, that occur between HIV-infected prostitutes and their customers, and the distribution of these acts according to prostitute characteristics.
8. Comparing the characteristics of the entire population of prostitutes with those of subgroups most likely to be recruited in studies of convenience samples (street prostitutes, prostitutes currently in jail, etc.).

The study will develop a statistical sampling frame and use it to identify, interview, and test a sample of 1000 prostitutes in Los Angeles County.

Collaborative research is now being carried out in various U.S. cities to determine how many prostitutes are infected with HIV-1. Virtually all the studies are using samples of convenience. Their results show seroprevalence rates from 0 percent (in Las Vegas, Nevada, and Colorado Springs, Colorado) to 57 percent (in Newark, New Jersey). Such studies provide valuable indications, but their statistical sampling techniques do not permit extrapolation to defined populations of epidemiologic interest. Instead, they provide information about selected groups of women who may differ substantially from those not sampled.

This study will provide unique data about prostitutes, permitting empirically based estimation of important population characteristics for the first time. By reducing uncertainty about this key population's characteristics and behavior, the study will greatly improve our ability

Sandra H. Berry, Naihua Duan, and David E. Kanouse are with the RAND Corporation, Santa Monica, California.

The study is being supported by Grant No. ARR-2(AHR-V)1R01 HD24897-01A1 from the National Institute of Child Health and Human Development.

to construct epidemiologic models and predict the future course of the HIV-1 epidemic. It may suggest intervention strategies and ways to target them to the groups at highest risk of infection. Finally, it may improve our methods for collecting similar data in other geographic areas.

Overview of the Sampling Plan

Illegal markets are notoriously difficult to study because they are covert in nature. Fortunately (unlike gambling and drugs), the prostitution market depends heavily on advertising. Therefore, it may be possible to develop rough estimates of the number of prostitutes and to describe their market's characteristics. Further, previous studies suggest that despite their need to be secretive, many people in the business have cooperated when anonymity is guaranteed.

The central feature of this study design is the use of randomized sampling methods to produce unbiased estimates and to assess the estimates' precision, using standard statistical methods. Successful application of this sampling approach will have considerable research value, because it will dramatically improve estimates of the size and characteristics of the prostitute population.

Because no simple enumeration of all prostitutes is available, this population cannot be randomly sampled from a convenient list. Previous studies suggest that appropriate first-round sampling units are most easily constructed by stratifying prostitutes according to their means of soliciting clients.

We can distinguish five major solicitation media: (1) advertisements in mass media (newspapers and magazines); (2) listings in yellow pages (for massage parlors and escort services); (3) street signs (massage parlors, strip joints); (4) personal referrals (e.g., through bell captains, taxi drivers, and organizers of entertainment for events such as trade shows and conventions); and (5) personal solicitation (streetwalkers).

A combination of list sampling and area probability sampling will be used to construct an overall study sample including prostitutes who use each solicitation method. Based on previous research these can be grouped into three broad subpopulations: (1) streetwalkers; (2) sex industry workers; and (3) call girls.

Street Prostitutes

These women solicit primarily through physical presentation and are best studied using an area probability sample. The approach in this study will be one that has been used successfully in studies of the homeless—that is, using information in the first round from police and providers of social services to estimate the population density of prostitutes by block. These estimates will be checked and updated through field observations, and sampling will be employed to select blocks representing varying levels of nonzero density. Blocks will be sampled on a probability-proportional-to-size basis, where size equals density. During the second round, selected blocks will be sampled and attempts will be made to interview a specified sample of the prostitutes working there. The distribution of street prostitutes in an area changes by time of day and season of the year, so density estimates will have to be time-specific as well as area-specific.

Sex Industry Workers

Customers for sex industry workers are solicited in their places of employment (for example, massage parlors and clubs). To sample these work locations, a combination of list sampling and area probability sampling will be used, compiling lists from advertisements in yellow pages, newspapers, and magazines. High density areas will be located through informants, and will be identified through advertisements such as street signs. For each location, information about the number of women who work there and the percentage who are prostitutes will be gathered. Based on these estimates, locations for interviews will be selected and sample sizes defined at each location. Because gaining access to each location involves substantial fixed costs, the sampling will be allocated in clusters to reduce the number of locations.

Call Girls

This segment of the prostitute workforce may be the most difficult group to sample with traditional techniques, because they are more difficult to identify, count, and interview. This category includes women who work for or through escort services, as well as self-employed call girls. A list sampling approach will be used for this population, working with lists from various sources. First we will compile lists of advertised call girl services and their telephone numbers from the yellow pages and other published sources. These lists will be matched to eliminate duplicate telephone numbers. Then a random sample of telephone numbers will be drawn, again using a probability- proportional-to-size approach.

From lists of call girls obtained from taxicab drivers, bell captains, entertainment organizers, and others whose work puts them in a position to make such referrals, sample women will be contacted by telephone and a meeting arranged. Again, cluster sampling of women in services will be used to reduce costs.

Naturally, every effort will be made to minimize nonresponse, but some will occur; however, the sampling procedure will be adapted to minimize its effects. The strategy to be applied will ensure that persons or institutions that refuse to participate are replaced with others that are as similar as possible. Elements in each sampling frame will be stratified according to characteristics judged important and that can be measured in advance (for example, neighborhood characteristics, type of publication in which advertising appears). Refusals will be replaced with those elements in the sampling frame that are most similar with respect to a vector of such characteristics.

Overview of Data Collection

Size and Composition of the Prostitute Population. Information will be collected that will allow us to deter-

mine for each prostitute (1) all of the methods clients use to contact her (and thereby the sampling frames through which she might have been recruited), and (2) the periods when she was at risk of being recruited. When combined with information about the sampling probability (which is under our control) and the response rate (most of which is not), such information can be used to calculate the assumed sampling probabilities. When summed, the inverses of these rates yield estimates of the prostitute population's size.

Structured interviews will last between 60 and 90 minutes. In addition, a blood sample will be taken after appropriate counseling of each subject. Each subject will be paid $50 for participating in the study. Data will be collected anonymously, and subjects can get the results of their blood tests by collecting them in person at a specified date and time or by visiting an established testing and counseling center. Tests will be performed for HIV-1 antibodies, hepatitis B surface antibodies, human t-cell leukemia virus-1 and 2 antibodies, and syphilis. Follow-up counseling and results, if desired, will be available through the testing and counseling center or through referrals to social service agencies.

Although the procedures for data and blood sample collection will be the same for all sample strata, contact procedures will differ. For street prostitutes and sex industry workers, field staff, equipped with a van, will go out in teams. The teams will include a driver who doubles as a security guard (unarmed) and interviewers who are trained to draw blood samples and provide pretest counseling. The van will be used to store lab supplies and cash for respondent payments. It will provide a clean, well-lighted place for taking blood samples and for interviewing. If necessary, it will also provide a refuge and means of mobility in case of trouble.

Most prostitutes will be contacted on the job. If necessary, interviews will take place at a mutually convenient time outside working hours, to minimize their loss of income. However, many of the subjects will be interviewed during working hours; for them, the size of the payment may be important.

Because the data will be collected anonymously, the same person could be interviewed more than once. The field staff will be looking out for repeaters and will ask each prospective subject if she has been interviewed before. However, without some means of identifying the respondents or a staff large enough to cover an entire area in one 24-hour period, the possibility of some double counting cannot be eliminated.

Specific Problems Related to Sampling

Unit of Analysis

We plan to use a variety of units of analysis, including prostitutes, prostitute-client encounters, and person hours spent in prostitution. The choice of unit will depend on the purpose of the analysis. For example, in characterizing the prostitute population, the prostitute will be the unit of analysis. However, considerable variation in prostitutes' levels of activity is expected, with many prostitutes engaged in prostitution only on a part-time basis. To characterize the prostitute work force, work force participation (person hours worked) will be used as the unit of analysis. To analyze the risk of HIV transmission, we will need to use the individual prostitute-client encounter as the unit of analysis.

For each unit chosen for reporting particular analyses, it is important that we be able to relate this unit to the unit of sampling, so that the sample can be weighted inversely proportional to the sampling probability. For the sample of street prostitutes, the sampling unit is approximated by person-hours spent on the street; a full-time prostitute has a higher probability of being included in the sample than a parttime prostitute. With this sampling approach, analyses that are based on person-hours as the unit of analysis are straightforward. For analyses that focus on prostitutes as the unit of analysis, a sample obtained cross-sectionally overrepresents full-time prostitutes; therefore, the sampled prostitutes must be weighted inversely by their level of work, defined by person-hours. For instance, a full-time prostitute in the sample should receive half the weight of a half-time prostitute in the sample, because the former is twice as likely to be sampled. To put it another way, if the characteristics of the population of all women engaging in prostitution are being estimated, the sample must be weighted to correct for the known overrepresentation of some types of prostitutes and underrepresentation of others. For analyses that are based on acts as the unit of analyses, the sample must be weighted by the encounter rates (number of encounters per unit of time worked).

For sex industry workers, the sampling unit will be the shops and the prostitutes who work at these shops. The list sample or area probability sample identifies the shops to be included in the sample; both shop-level and individual-level interviews will be conducted. To analyze using person-hours or encounters as the unit of analysis, the weights of the sample must be adjusted.

Nonresponse Bias

Reporting bias is a serious problem in human sexual research. In this study, many sampled prostitutes will be encountered who refuse to be interviewed, as well as respondents who are selectively cooperative; for example, who agree to be interviewed but decline to provide a blood sample. If the nonresponse is nonrandom, that is, if the respondents and the refusals differ in their HIV infection rate, the data would be affected by nonresponse bias: estimates based on the respondents would differ from what would have been obtained from the refusals.

Nonresponse bias is difficult to deal with in any survey research. We will mainly focus on evaluating the potential for having serious nonresponse, and hope that most of our major analyses do not have such a serious problem. If some of the analyses fail the test, the results will have to be qualified. However, the potential for nonresponse bias in this study should be substantially lower and the ability to evaluate that potential substantially higher than has been the case in prior studies based on volunteers or jailed prostitutes.

In many survey studies, the population studied is one that can be fairly well characterized by available data that are independent of the survey. In that case, to assess the potential for a serious nonresponse bias, it is possible to compare the characteristics of survey respondents with what is known about the population.

However, very little is known about the prostitute population that can be used for this type of analysis. Observable characteristics such as race, type of dress, approximate age, type of location, and so forth, of those who refuse to be interviewed can be collected. In addition, how respondents with similar observable characteristics to those who refuse compare with other types of respondents with respect to their reported behavior can be examined. This provides some information on the extent of response bias that is associated with these observable characteristics, but obviously provides none about any bias that is uncorrelated with these characteristics. This approach was used in a recent RAND study of the homeless, and will be made use of here as well.

Another approach is to use information on the difficulty of completing an interview with respondents as a way of judging possible differences between respondents and nonrespondents. Some interviews are more difficult to obtain than others; for example, they require more interviewer persistence or persuasion. Reluctant and difficult-to-reach respondents may offer clues to the characteristics of nonrespondents, who are even more reluctant or difficult to reach. If measures are taken of the difficulty of obtaining each interview, it is then possible to examine, within the sample of completed interviews, the relationship between completion difficulty and respondent characteristics on the one hand and responses to key items on the other. This provides one basis for assessing possible biases introduced by nonresponse. Although hardly a definitive solution to the problem, this is feasible and worth doing.

Still another approach is to offer additional incentive payments to a random subsample of refusals, and then to compare the responses of those initially refusing with those of other respondents to gain some idea about the distinctiveness of nonrespondents. This was considered as a possible strategy, but rejected as infeasible in a field study of this population, where an active grapevine can be expected to quickly broadcast news of any differential incentives.

For these reasons, the nonresponse bias will be dealt with (1) by taking all feasible measures to minimize the extent of nonresponse, (2) gathering as much information as possible on the characteristics of nonrespondents, (3) measuring interview completion difficulty for respondents, and (4) analyzing data gathered in (2) and (3) to assess how nonrespondents might differ from respondents in their characteristics and behavior.

DISCUSSION PAPER

Samples for Studies Related to AIDS

William D. Kalsbeek

Introduction

The common theme in these papers is that they all in one way or another deal with the problem of sampling in surveys that are intended to advance our understanding of the acquired immunodeficiency syndrome (AIDS). But there is where the similarity ends. The variety of sampling strategies displayed in these studies arises not just from the creativity exhibited by the investigators but from the varying settings in which these studies were done. By seeing this diversity in study settings, we can appreciate just how differently the sampling problems in AIDS research manifest themselves. The diversity of problems and solutions in these five papers also serves to remind us that just as no universal cure for the common cold or for cancer exists because of the multiple forms they assume, there is no universally applicable solution to sampling problems in AIDS surveys.

Categorizing Surveys on AIDS

Generally speaking, sampling in AIDS research boils down to the matter of sampling to examine a rare trait, where possessing that trait often is undesirable and therefore highly stigmatized by the majority of society. This broad portrayal of the problem, while inadequate in fully capturing all aspects of its diversity, does enable us to see that finding solutions requires techniques that can overcome both the rarity and stigma of AIDS.

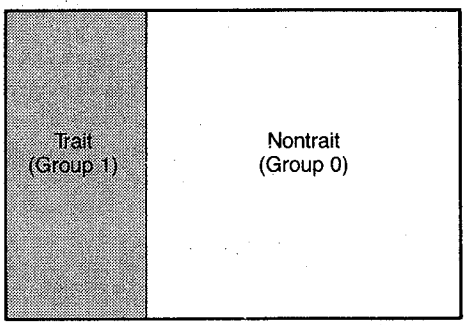
Examining the variety of sampling problems in AIDS research requires that we begin by categorizing the many different settings in which the problems appear. Figure 1 represents a population of N members for whom the AIDS survey is targeted. The shaded area is for that segment of the target population with some AIDS-related trait (Group 1) and the unshaded area corresponds to those without the trait (Group 0). We let N_1 and N_0 denote the number of population members with and without the trait, respectively, and use the symbol $P_1 = N_1/N$ to represent the proportion of the population possessing the trait.

Using these terms as the basis for discussion, we can now examine how design settings in AIDS research differ. Their variation seems to be largely explained by the following:

- the scientific aims of the study;
- the definition of the AIDS-related trait being studied in the population; and
- the definition of the target population.

Each of these considerations has its own set of implications on the design setting that will be examined briefly.

Figure 1. Key elements in AIDS-related sampling from a target population



Number of population members:	N_1	N_0
Proportion of the population in the group:	P_1	$1-P_1$

William D. Kalsbeek is with the School of Public Health, University of North Carolina, Chapel Hill, North Carolina.

The aims of AIDS-related surveys seem to be one or both of the following two types. One major aim of a study might be to estimate the prevalence rate (P_1) of the trait in the population from a sample of the target population. This was a major goal for the papers by Capell and Schiller and by Freeman and associates. The Capell and Schiller study, and other statewide telephone interview surveys and a recent supplement to the National Health Interview Survey, were done to estimate rates of the knowledge and risk associated with AIDS in the general adult population. Freeman and associates, on the other hand, estimate rates of the risk of infection, acquisition, and transmission from a general population sample of males. When the object is to estimate relatively small values of P_1 from relatively small samples, the sampling process might involve an oversampling of areas of known concentrations of Group 1. Such was the case for both of these studies.

A second aim in AIDS-related surveys might be to study a sample of those in the population with some trait by either sampling from a constructed list of Group 1 members or by screening a sample of the target population for the trait. In some instances the trait being investigated includes the entire group of persons at some level of risk to acquiring AIDS, whereas in others, Group 1 may be limited to a subset of persons at this level of risk (for example, intravenous drug users). Sampling from a Group 1 listing was the aim of the papers by Fleishman and others, by Bradford and Keeter, and by Berry and colleagues, with persons receiving treatment for AIDS in nine centers as the object of the study by Fleishman, and others; male prostitutes in the Bradford and Keeter study; and female prostitutes in the survey described by Berry and colleagues. In each of these studies, sample selection was done from specially constructed Group 1 lists.

Surveys on AIDS can also be distinguished by the definition of the AIDS-related trait being studied. Generally speaking, traits differ according to the level of risk associated with getting AIDS. Lifestyle and sexual behavior may establish various degrees of risk for acquiring the AIDS virus, as was the case for all but the paper by Fleishman and others. The study by Freeman and associates goes so far as to classify respondents as being at high risk by their answers to a number of behavioral items. The trait for the next higher level of risk is acquisition of the AIDS human immunodeficiency virus (HIV). This trait is not considered in any of the studies presented here, but is used in the National Seroprevalence Household Survey currently being pretested for the Centers for Disease Control. The ultimate level of risk in defining the trait is the development of clinical symptoms for AIDS, as was the case for the definition of Group 1 in the study of patients in treatment by Fleishman and others.

The third basis for distinguishing surveys on AIDS is the definition of the target population. Here the range of classification goes from local (for example, cities, individual counties, small groups of contiguous counties) to statewide to national. The primary impact of geographic scope on an AIDS survey is on its overall complexity and ultimately the cost of both the sampling and data gathering activities of the survey. This is especially true for surveys where list sampling is done for certain elusive high-risk traits, such as for male and female prostitutes as described in the papers by Bradford and Keeter and by Berry and colleagues. Although surveys on AIDS have been done at all levels, most have been done for target populations defined at the state and local levels, including all but the study reported by Fleishman and others which covered nine metropolitan areas in the United States.

Problems and Implications

Having classified the settings within which samples must be drawn for surveys on AIDS, specific design problems must be faced. Two underlying problems are implied in the earlier general statement on the difficulties of sampling in this context. One has to do with the smallness of P_1 for most AIDS-related traits and the relatively large sample sizes that are needed as a result. The second problem of sampling to estimate small rates or to sample rare population attributes is well-known in the statistical literature. Some of the earlier treatment of the subject was given by Kish (1965) and Birnbaum and Sirken (1965) and has been followed by a substantial literature which has been extensively reviewed in two relatively recent papers, one by Kalton and Anderson (1986) and the other by Sudman and Kalton (1986).

At the heart of the matter is the exceptional cost required to achieve adequate representation of Group 1 in a sample to achieve acceptable levels of precision for survey estimates. A key sampling issue then becomes to find ways to reduce the cost per discovered sample member with the trait, while retaining the integrity of the sample in terms of the mean squared error of estimates.

The other underlying problem in sampling for AIDS-related surveys has to do with the strong stigma attached to many of the lifestyles and behaviors that are known to place individuals at high risk for acquiring the AIDS virus. Those possessing these characteristics include gay and bisexual men, intravenous drug users, and both male and female prostitutes. All are relatively rare and inaccessible for research purposes. The papers by Bradford and Keeter and by Berry and colleagues, each dealing with the problem of gaining access to prostitutes, point to the difficulties inherent in studying these segments of America's social underground. Thus, in addition to the significant challenges one faces in obtaining useful substantive data from these groups, there remains the problem of identification for choosing a reasonable sample.

Because much of the behavior that places them at high risk for AIDS is also illegal or judged to be immoral or both by most of society, this elusive segment of American society is especially reluctant to be enumerated for any reason, much less to have their lifestyles exposed through penetrating questioning. For these reasons, obtaining complete and current lists of these segments is extraordinarily difficult and costly. For both sampling and data gathering, ways must be found to approach these people without conveying a threat that will cause them to retreat at the sight of the enumerator.

Sampling Techniques for AIDS-Related Surveys

Having identified some of the common design settings and discussed the sampling problems inherent to those settings, let us now consider how one might deal with these problems in surveys on AIDS. The techniques to be reviewed are plausible for these design settings but have not all been used. Moreover, the frequency of use for the ones that have been found useful is not the same among methods.

Nonprobability Sampling

Most of the surveys on AIDS have been confined geographically to local areas and the majority of these have focused on samples of high-risk groups. Inasmuch as the usual manner of probability sampling from well-constructed lists has often been too difficult and costly even at the local level, nonprobability samples of convenience have been used.

Included in this classification of methods are those samples that are chosen by requesting volunteers from a specific group, such as college students (Simkins & Eberhage, 1984), or by approaching respondents in parks and other public places (Temoshok & associates, 1987). Also included are those designs where an isolated but convenient group is identified through a list or some other device and then enumerated completely (for example, hospital workers from personnel lists as in O'Donnell and associates, 1987). Bradford and Keeter's report on the study of Richmond's male prostitutes is another example of this approach to sampling.

Although widely used, these nonprobability samples of convenience carry with them a mixed bag of strengths and weaknesses. The principal attraction is their relatively low selection costs, mainly because existing lists are used or, in some instances, are not needed. On the other hand, the analyst of these data must be careful to point out (and they usually do) that inferences beyond these narrowly defined groups is not possible. Unfortunately, despite these disclaimers conclusions beyond the confines of the sampling process are undoubtedly drawn from this class of convenience samples. The limitations of statistical inference beyond the sample itself constitutes the major disadvantage of this class of convenience samples. Related to this shortcoming is the potential damage of misappropriated inference by drawing conclusions regarding the general population from these samples.

General Population Screening with Disproportionate Sampling

In this class of strategies the object is to sample the target population and, by application of eligibility criteria for the AIDS-related trait, obtain a sample that contains a relatively higher proportion of Group 1 members than the proportion (P_1) of such numbers in the target population. This oversampling of Group 1 members is usually done through disproportionate sampling of strata identified at some point during sample selection. The goal of sampling in this kind of design may either be to estimate P_1 with greater precision or to produce a suitably large sample of Group 1 members to interview.

Any mode of data collection can be used for the screening process; however most common in this country because of its relative efficacy and cost is screening by random digit telephone sampling. This was the approach used in the papers by Capell and Schiller, for the adult population of California and by Freeman and associates for adult males in Los Angeles County. Oversampling of high-risk groups in each of these studies was accomplished by applying larger sampling rates to those areas that are likely to have higher percentages of these groups.

Other options are possible, however, in doing this oversampling. One is to modify the Waksberg-Mitofsky random digit sampling design so that clusters of telephone numbers are chosen with probabilities proportional to the number of residential telephone numbers that include Group 1 members (as opposed to proportional to the total number of residential telephone numbers). An alternative for certain traits is to limit screening to individuals within households by taking all Group 1 members and either none or a small percentage of Group 0 members in selected households.

Although often more costly than the telephone approach, some kinds of screening are best done by personal interviews. In these studies the criteria for establishing membership in Group 1 can only be done through direct personal contact, to do such things as examine or draw blood from members of the screening sample. The seroprevalence study on AIDS, whose pretest findings are reported later in this volume, is an example of this type of screening study. Screening by mail is a third possibility, although because of its traditionally low return rate in this country it is not usually a viable candidate for sampling in AIDS surveys. This would be especially true when screening criteria require information of a sensitive nature (for example, sexual behavior or drug usage).

Unfortunately, one often finds that oversampling in those studies which employ disproportionate sampling of clusters is only moderately successful in increasing the representation of Group 1. This is because the relative increase in the Group 1 sample size is directly tied to how segregated the trait is in the target population. More specifically, oversampling will work best when the sampling strata, to which the higher sampling rates are applied, consist entirely of Group 1 members. Anything less than almost complete saturation of the trait in the disproportionately sampled strata will be markedly less successful in increasing representation of Group 1 in the sample. This is because with the Group 1 members coming in higher numbers from these strata, there will be more Group 0 members coming from those same strata.

Nonoverlap List Sampling

Two classes of probability sampling for rare traits call for direct sampling from at least partial lists of Group 1 members. Lists of this type for surveys on AIDS may be available or can be developed from social groups, religious organizations, health care providers, educational institutions, and certain targeted publications. These

lists often cover different segments of a particular AIDS-related trait, and the ultimate unit of sampling in these designs is usually the individual.

In one of the two classes of designs the sampling frame is taken exclusively from an existing series of related but nonoverlapping sources. Inasmuch as the frames are presumed to be available and the lack of overlap among sources simplifies the selection process, samples falling into this class might be considered a type of randomized convenience samples to go with their nonrandomized counterpart mentioned earlier. For example, a part of the sample for the study by Fleishman and associates reported here, used a combined list of patients from nine urban treatment facilities to assess the health care received by persons with AIDS. Other examples of nonoverlap list sampling have drawn upon such things as lists of students (Price & associates, 1985) and health care workers (O'Donnell & associates, 1987).

In this class of designs the final list either adequately covers the segment of the population with the trait, or necessity demands the justification that what is available is also acceptable. Herein lies the rub of this approach because, with its advantage of avoiding large frame construction costs seldom will the frame's coverage be adequate. Fortunately in some cases it is feasible to improve the available list through the process of "snowballing," in which members on the initial frame are contacted to identify additional Group 1 members, the new members identify yet other members, and so on, until saturation by some set of criteria is achieved.

Overlap List Sampling

In this second class of list sampling procedures the sources are combined to improve coverage of Group 1, but unlike the prior class there exists a certain amount of overlap among sources. The phrase "sampling from multiple frames" is often used to describe the selection process in this class of designs. For example, the frame for the sample of female prostitutes in Los Angeles as described by Berry and associates was developed from lists obtained for three partially overlapping subsets of this population group: street prostitutes, sex industry workers, and call girls. Another possible application of this approach would be for sampling gay and bisexual men in an area where the frame is constructed from subscriber lists for local gay newspapers, from a list of gay bars, and from a list of gay social, religious, and advocacy organizations in the area.

The hope in combining lists from multiple sources is that for the price of dealing with any between-group overlap the combined lists, will more closely match the intended Group 1 membership. In the study of female prostitutes this match is likely to be quite good, but for sampling homosexual men the suggested sources would in all likelihood jointly represent only those self-acknowledged gay and bisexual males who are younger and more active socially.

The matter of sample coverage aside, another key issue in sampling from multiple frames is how to handle the overlap. Several options are available. One is to eliminate it before selection, although this is often impractical. Another option is to uniquely link to one of the lists each target population member on the combined lists and to implement this so-called "unique counting rule" during sample selection. The intent of this option is to give only one opportunity for selection to those population members who appear on more than one list. While this rule remedies the problem statistically, it may not be the most efficient operationally because some selections are rejected from the sample if they appear on multiple lists and were chosen from one to which they were not uniquely linked. The third option is to select the sample from the unaltered combined lists and then to compensate for the multiple selection opportunities in the estimation process. Although this process can theoretically produce unbiased estimates, the variation in selection probabilities contributes to increased variances in these same estimates.

Network Sampling

One additional design strategy that has been successfully used in sampling for other rare traits, but has not yet been tried in AIDS research, is called network or multiplicity sampling. Unlike sampling from multiple frames where the multiple linkage of population members is treated as an impediment, network sampling exploits this multiplicity by allowing members of Group 1 to enter the sample through more than one sampling unit. The set of sampling units through which a population member might be chosen is some well-defined and operationally feasible social network such as brothers and sisters, residents of the same building, or members of the same church or synagogue. While screening respondents' networks to identify persons with the rare trait can increase the take of Group 1 in the sample, the down side of this approach is that network sizes must be determinable to calculate selection probabilities for the sample, and the variation in these probabilities due to variation in network sizes will once again increase the variance of survey estimates. Informants must also be willing to identify members in the network with the rare trait.

Dealing with the Stigma of AIDS

Some remedies to the impact of the stigma of AIDS on sampling and related activities have been tried and hold promise. One is to use "insiders" to perform sampling and data-gathering activities. For example, the study of male prostitutes reported in the papers by Bradford and Keeter successfully established an entry phase to data gathering in which the interviewer became visible in the area where interviewing was to be done, to reduce the suspicion that would later come in conducting interviews. Another remedy is to solicit endorsement from organizations whose support of the study could help to positively influence participation by the reluctant among those being listed or later chosen in the sample.

Conclusions

Because it combines the difficult issues of sampling both rare and elusive populations, which in and of themselves continue to provide fertile ground for methods

research, the problems of sampling for AIDS surveys are numerous, substantial, and therefore far from being solved. However, the even greater problems that could befall society if an AIDS epidemic were to reach full potential demand that we diligently seek solutions to these sampling issues. Toward this end perhaps some of the areas where further research might prove useful are the following:

- Assessing the effects of the stigma of AIDS on the process of frame construction and participation in AIDS surveys.
- Determining the best type of data gatherers for sampling tasks (for example, "insiders" with no survey experience or experienced survey interviewers with no inside connections to persons with the trait).
- Investigating the feasibility of network sampling for some rare traits (for example, intravenous drug users).
- Studying the utility of using endorsement to improve response rates in sampling for certain traits.
- Investigating the utility of inference from "convenience samples," thus providing some guidance as to how they might be most useful (for example, for selecting members of focus groups to identify data needs in AIDS research).

References

Kalton, G., & Anderson, D. W. (1986). Sampling rare populations. *The Journal of the Royal Statistical Society,* 149 (1), 65-82.

Kish, L. (1965). *Survey sampling.* New York: Wiley & Sons.

O'Donnell L., O'Donnell, C. R., Pleck, J. H., & associates. (1987). Psychosocial responses of hospital workers to the acquired immunodeficiency syndrome (AIDS). *Journal of Applied Social Psychology,* 17, 269-285.

Price, J. H., Desmond, S., & Kukulka, G. (1985). High school students' perceptions and misperceptions of AIDS. *Journal of School Health,* 55, 107-109.

Simkins, L., & Eberhage, M. G. (1984). Attitudes towards AIDS, herpes II, and toxic shock syndrome. *Psychological Reports,* 55, 779-786.

Sudman, S., & Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology,* 12, 401-429.

Temoshok, L., Sweet, D. M., & Zich, J. (1987). A three city comparison of the public's knowledge and attitudes about AIDS. *Psychology and Health,* 1, 43-60.

DISCUSSION PAPER

Samples for Studies Related to Acquired Immunodeficiency Syndrome

Daniel G. Horvitz

These papers have focused on sample designs for acquired immunodeficiency syndrome (AIDS)-related studies. Capell and Schiller and Freeman and associates examine sampling aspects of telephone surveys designed to measure, in defined populations, the distribution of risk factors or behaviors related to human immunodeficiency virus (HIV) infection. Fleishman and colleagues discuss sampling issues in a study to evaluate a program aimed at providing health and social services to persons with AIDS. Bradford and Keeter and Berry and coworkers focus on sample designs for studies of prostitutes to gather information needed to structure more effective programs aimed at preventing the spread of HIV.

While the sampling problems discussed in these papers are quite interesting, they are better considered in a total survey design context, and some of these comments will reflect that orientation. Efficient survey design requires that attention be given to nonsampling errors as well as those due to sampling. Thus, it may be more cost effective to devote some of the resources available to either control or reduce nonsampling errors or both. The challenge to the survey designer is to use the resources available in ways that will minimize the total survey error.

Designing the Survey

Survey designs for population-based studies of AIDS-related risk behaviors must address issues and challenges that are not present in health surveys in general. First, the population segments of interest represent only a small proportion of the total population. Second, AIDS has been identified as primarily a disease of homosexuals and intravenous (IV) drug users. Those segments of the population are particularly sensitive to the consequences of revealing their lifestyles for fear of subsequent discrimination by employers, landlords, insurance companies, and others. In the absence of hard data on the likelihood of participation in risk-behavior surveys by gays, IV drug users and others at risk of HIV infection, one must assume that estimates derived from such surveys have a clear potential for significant bias due to nonresponse by these key at-risk groups.

Third, the general public considers sexual behavior, the primary topic of AIDS-related surveys, to be a very private matter, and hence may not respond to such surveys at the higher-than-average rate usually realized in health surveys. Fourth, even if sample persons agree to participate in an AIDS-related survey, they may not answer each and every question candidly. Again, in the absence of hard data on the validity of responses to detailed questions about sexual behaviors that put a person at risk for HIV infection, one must also assume a clear potential for a significant response bias in the survey estimates based solely on the responses to these sensitive questions.

Can these issues and challenges be addressed? It depends on whether we can design AIDS-related surveys that bring the response and nonresponse biases under control. This might be accomplished by creating survey conditions that achieve high levels of participation by at-risk sample persons and hence reduce these biases to an acceptable level. If that is not possible, then we need to develop survey designs that provide sufficient information about the response and nonresponse biases to enable adjustment of the survey estimates. In today's sensitive environment, we just do not know whether we can design and carry out AIDS-related surveys that will meet acceptable survey standards.

Response Rates

I am not aware of any study which provides data on the response rate of persons at risk for HIV in a population based survey of sexual behaviors associated with

Daniel G. Horvitz is with Research Triangle Institute, Research Triangle Park, North Carolina.

Table 1. Example with response rates unrelated to attribute

Attribute category	Expect in sample	Response rate (%)	Expected no. of respondents
Yes	50	0.68	34
No	950	.68	646
All	1000	.68	680

NOTE: The expected survey estimate for the attribute using the response data only is 34/680 = 0.05, the attribute's expected value.

HIV infection, let alone the bias due to nonresponse in the estimate of a specific at-risk behavior. Nor am I aware of any data on the differential between the response rate realized for sample persons not engaging in that behavior.

We should be aware that the response rate, when attempting to measure a relatively infrequent behavior, can be rather uninformative about the bias due to nonresponse. The example in Table 1 shows that even with a relatively low response rate (68 percent), the estimate using the response data only is unbiased when the propensity to respond is not related to the attribute of interest. On the other hand, Table 2 shows that even with a relatively good overall response rate (88 percent), but with a large differential in the response rates for those with and without the relatively infrequent but sensitive attribute, the response data underestimates the proportion with that attribute by 43.2 percent.

The point of these preliminary remarks is that it is not appropriate to focus exclusively on sampling error and to ignore the nonresponse and measurement error components in the design of AIDS-related surveys. Some portion of the survey budget should be allocated to the measurement and control of the nonsampling errors in the data.

Telephone Surveys

The paper by Capell and Schiller focuses on improving the efficiency of the California statewide telephone survey of HIV risk behavior. Data from a previous survey on persons engaging in high risk behaviors were used to classify Zip codes across the state into high, medium, and low risk strata. The high and medium risk strata were oversampled with some success, in that interviews were completed with higher numbers of persons at risk than expected from random sample. The ideas and approaches in the paper are interesting, and the following suggestions come to mind:

1. Consider using public health and census statistics at the Zip code level to further refine the stratification. For example, counts of persons with AIDS, counts of persons treated for IV drug usage, counts of persons tested for HIV infection in public clinics, counts of HIV infection among child-bearing women, proportion of never married males aged 35 to 54.
2. Consider computing the optimum allocation of the sample to the strata (using AIDS incidence rates, for example) rather than choosing the sampling rates arbitrarily.
3. Consider also assessing the efficiency of the sampling design by comparing the estimated variance of the survey estimate of the total persons in a given risk group in the state with the estimated variance achievable with a simple random sample.

Several nonsampling error concerns not addressed in the paper also come to mind. The paper makes no mention of the telephone survey coverage bias. Nor is the reader informed of any details about the 32 percent of the original sample that chose not to participate in the survey and the potential for nonresponse bias in the estimates. No mention is made of the potential for some proportion of the survey respondents to have denied engaging in high risk behaviors and the consequences of such denials.

Capell and Schiller point out that men reporting sex with other men were harder to reach in the survey. The reader would have been helped if the authors had included a formal statistical test and rejection of the hypothesis that the proportion of gay and bisexual males among all males interviewed after, say, five unsuccessful attempts was no greater than the proportion realized during the first five attempts.

Up to 20 attempts were made with each sample telephone number. The authors make a case for persistence in seeking to complete interviews and report the extent to which the State's estimate of gay and bisexual men would be underestimated by restricting the numbers of attempts. Without wanting to belabor the point, the underestimation of that group could be 100 percent or more due to nonresponse bias or risk denial.

In assessing the value of repeated callbacks, the cost of completing an interview should also be considered. It may be more cost effective in terms of identifying at-risk persons to increase the sample size and make fewer attempts per sample case. A weighting procedure to account for the not-at-home cases would be required, of course.

The survey also included questions on close friends who engage in high risk activities. This is a network sampling approach that clearly deserves more thorough testing. Estimates derived from these data should be weighted (inversely by the number of close friends) to account for the potential for multiplicity; that is, multiple reporting of persons with a given risk status.

Table 2. Example with response rates related to attribute

Attribute category	Expect in sample	Response rate (%)	Expected no. of respondents
Yes	50	0.50	25
No	950	.90	855
All	1000	.88	880

NOTE: The expected survey estimate for the attribute using the response data only is 25/880 = 0.0284. The relative bias in this estimate is (0.0284−0.05)/0.05 = −0.432 or −43.2 percent.

Population Screening

Freeman and associates examine the need for population-based data on AIDS-related risk behavior at the community level, and the issues to be addressed in designing a survey to provide the needed data. They gain some very interesting and valuable first-hand experience by conducting a computer-assisted telephone interview (CATI) survey of HIV risk levels among Los Angeles County's 18- to 60-year-old male population. A very high oversampling rate (14 to 1) was used in the 18 census tracts with historically high prevalence rates for AIDS.

The realized response rate of approximately 40 percent raises serious questions about the validity of the survey estimates. The authors recognize the need to improve participation, but they conclude, prematurely, that "low rates will remain endemic to studies on AIDS-related behavior." For example, the response rate in the recent pilot survey conducted in Allegheny County, Pennsylvania, for the National Household Seroprevalence Survey (NHSS) was 81 percent. In the face of the low response rate, the authors also expressed their willingness "to settle for 'good enough' estimates." However, significant progress in the development of techniques to enhance participation in face-to-face surveys has already been made in the NHSS. In view of the limited numbers of risk behavior surveys to date, we have just begun to scratch the surface of their quality potential.

Emphasis on the need for designs that control nonsampling errors as well as sampling errors is not intended to imply that only very costly NHSS-type surveys are acceptable. Low-cost surveys that produce less accurate estimates can be "good enough" provided their estimates are not subject to large biases of unknown magnitude. Telephone surveys might very well be good enough for estimating HIV risk behavior distributions, but there is not enough evidence to feel comfortable about them as of now.

Freeman and colleagues do not make clear whether the algorithms used to classify sample persons by risk take into account attenuating factors such as monogamous relationships of the use of condoms. Epidemiologists, of course, may view these behaviors as having a negligible effect on the survey estimates of the proportions of homosexuals at risk.

Sampling and Accessing Issues

Fleishman and associates provide a useful discussion of the problems encountered in attempting to sample and interview people with AIDS in a given community. The authors conclude that it is "extremely difficult to achieve a probability sample" of this group. Although that conclusion may be accurate, it is not entirely discouraging. Inasmuch as more than several items of data were available on all clients or patients selected into the sample, meaningful weighting-class adjustment of the study estimates to account for nonresponse seems to be a definite possibility.

Because the samples were selected initially from a community-based organization (CBO) list and subsequently from clinic patient lists, dual frame sampling and estimation theory might prove helpful in developing combined survey estimates. Attention needs to be given to determining the overlapping as well as the nonoverlapping portions of the two sampling frames.

In Table 2, these authors show that almost half of the CBO sample cases remained in limbo because case managers had not solicited their consent to participate in the survey. Rather than complete the survey with no further information about this group, it is suggested that the CBOs involved be asked to make a special effort to complete the consent process for a random subsample. Since this special subsample need not be large for any CBO (only a pooled analysis is contemplated), the requisite cooperation from the case managers might be realized. Interviews with all those who give consent should also be attempted. The data for this sample then become quite useful in that they represent the 1,300 or so "no disposition" cases.

This study of persons with AIDS requested each respondent, following completion of the risk behavior questionnaire, to sign two consent forms, the first giving permission to be recontacted within a year for a second interview and the second giving permission to examine the respondent's medical records. This was viewed as a valuable and successful procedure by the authors. A question arises, however, as to whether this is an ethical procedure in human subjects research. Should not the initial consent process include all components of the research process in which the sample person will be asked to participate? Otherwise, can we claim that the sample person's initial decision to participate is a fully informed decision? Answers to these questions are needed, particularly for those engaged in AIDS-related surveys.

Studies of Male Prostitutes

The Bradford and Keeter paper provides valuable information on outreach procedures that proved to be very effective in gathering data from almost all the males working as street prostitutes in two specific areas of Richmond, Virginia, at some time during the course of the field effort. The two areas "were selected for intensive coverage, because they were the locations where most activity takes place." Four other areas were identified as sites where male prostitutes work, but they were not covered in the study because observation of these areas before initiation of data collection "revealed very little activity."

Although it is doubtful, in this instance, that coverage of those eligible for the survey suffered significantly, the other four areas probably should have been covered on a time-sampling basis. This is, each of the areas where male street prostitute activity had been observed is selected on a probability basis to be covered by the survey team for a sample of nights covering the data collection period. The areas should be sampled with probability proportional to level of activity so that the areas with

more activity are selected to be covered in more of the nights than the areas with less activity. With this design, the data collected over the course of the study can be used to generate estimates for the entire set of areas, rather than just for the two areas with more intensive activity just before the survey. It also guards against any shift in intensity of activity from one area to another.

Study of Female Prostitutes

The final paper by Berry and others provides a comprehensive discussion of plans for a probability sample of the Los Angeles prostitute population. The most challenging problem in the proposed survey is to be able to compute for each prostitute selected into the sample her exact probability of being included in the sample. The inclusion probabilities, as the authors recognize, are key to all subsequent analyses, because expansion of the sample data to the population of all Los Angeles prostitutes will be accomplished for each measure of interest by weighting that measure for each prostitute in the sample by the inverse of her inclusion probability, and summing these weighted measures over the sample.

The authors propose to develop several sampling frames based on customer solicitation methods. These frames may overlap, with some prostitutes included in more than one of them. Prostitute work habits vary both by time and location, which adds complexity due to multiplicity to the sampling process. Berry and her associates are aware of the multiplicity problem but do not provide a specific solution other than to imply that they will adjust the sampling probabilities to address it. In the face of these overlapping frame and multiplicity complications, the following suggestions are made.

Assuming that all the requisite frame data have been gathered on the city blocks where street prostitutes solicit and on the massage parlors, clubs, and other work locations of sex industry workers, these two frames should be combined into a single list of locations. Each location is assigned a specified number of sampling units or clusters of prostitutes, if you will, based on the advance information about the number who work at that location. The cluster size should approximate the number of interviews expected to be completed in a specified data collection daily work segment, such 4-hour or 6-hour periods. Given a carefully defined and constructed set of sampling units each tied uniquely to a specific location, this frame should be expanded to include, for each location, a set of potential data collection daily time segments, covering a 1-week period, that make the most sense for that location. In effect, the sampling frame now has a set of sampling units defined over time and space.

These sampling units can be stratified by type of location and possibly time of day, depending on the number of data collection teams. A set of time and location-defined sampling units would be selected at random without replacement from each stratum, distributed over the days allocated to data collection. With this procedure, locations will be selected with probability proportional to the number of clusters assigned to them in the frame. Also, a given multicluster location could be sampled for coverage during more than one time segment in a given day.

The data collection teams must have procedures for listing (and subsampling, if necessary) the prostitutes working at each location during the time segment selected for coverage. The subsampling rate would be determined by the number of sampling units (clusters) assigned to the location initially.

As noted above, the potential for multiple counting of prostitutes is high because the times and locations they work vary from day to day. To adjust for multiple counting the total number of sampling units (defined over time and space) with which each prostitute in the sample is linked, that is, her "multiplicity" must be determined. This problem can be resolved by asking each prostitute in the sample to report for each of the daily data collection time segments the number of different locations they usually work during that time slot over a period of, say, a week (or a month, since it is important that the number reflect the number of different locations that might be worked during the course of the data collection.) The estimation weight used for each sample prostitute (that is, the inverse of inclusion probability for the sampling unit that brought her into the sample) must be divided by her multiplicity when generating estimates for all prostitutes in Los Angeles.

One additional point. Since some, if not all, of the prostitutes in the population will be linked to more than one sampling unit, they can be selected into the sample more than once. While they need not be interviewed each time they are selected (beyond the first), their data need to be included each time they are selected into the sample.

Berry and co-workers point out that with their proposed sampling plan for the sample of street prostitutes, full-time prostitutes will have a higher probability of being included than part-time prostitutes. They propose therefore to weight each sample prostitute inversely by their level of work, in analyses in which the prostitute is the unit of analysis. This is not necessary in the sampling plan proposed above since the differentials between full-time and part-time prostitutes are taken care of through the use of time and location-defined sampling units and multiplicity measures. Clearly, any analyses of the prostitute work force should use person-hours worked as the unit of analysis.

The proposed sampling plan also includes a strategy for substituting for prostitutes or workplaces that refuse to participate. This strategy is not an acceptable solution to the nonresponse problem. A better solution is to develop a set of weighting classes based on the characteristics available for both respondents and nonrespondents, and to adjust the analysis weights of the respondents in each weighting class by the inverse of the response rate for the weighting class.

If all efforts to convert a random subsample of refusals take place after all other field work is completed, it may be feasible to offer a higher monetary incentive. If is doubtful, of course, that street prostitutes who refuse initially can be located easily for a postsurvey conversion effort. A postsurvey refusal conversion effort should be

possible with sex industry prostitutes and with call girls, however. The incentive offered should be large enough (probably at least twice the average hourly earnings rate) to cause the prospective respondent to give serious thought to participating.

Total survey design is discussed in some depth in the proceedings (NCHSR, 1977) of the first conference on health survey research methods. The ideas developed therein certainly deserve serious consideration in the design of AIDS-related surveys.

Reference

National Center for Health Services Research. (1977). *Advances in Health Survey Research Methods* (DHEW Publication No. HRA 77-3154). NCHSR Research Proceedings Series. Washington, DC: U.S. Government Printing Office.

Designing a Household Survey to Estimate HIV Prevalence: An Interim Report on the Feasibility Study of the National Household Seroprevalence Survey

Michael F. Weeks, Daniel G. Horvitz, Peter L. Hurley, and Robert A. Wright

Introduction

Since the first cases of acquired immunodeficiency syndrome (AIDS) were reported in mid-1981, the number of diagnosed AIDS cases in the United States has surpassed 90,000. However, because individuals who have been infected with the human immunodeficiency virus (HIV) often do not develop AIDS for years after initial infection, the full scope of this epidemic is unknown. Current estimates indicate that at least 365,000 individuals will have developed AIDS by 1992. If it continues to spread unchecked, the epidemic could place unprecedented demands on the U. S. health care system, with the possibility of affecting fundamental aspects of our society. To develop prevention and treatment strategies for a health crisis of this magnitude, health officials must have accurate estimates of current levels of HIV infection in the general population.

The Centers for Disease Control (CDC) has initiated a family of studies that is designed to monitor HIV infection among certain subgroups of the general population. These studies include patients attending selected hospitals, sexually transmitted disease clinics, drug abuse clinics, and family planning clinics. Collectively these studies test over 1 million persons annually for HIV. While they produce useful data on the incidence of HIV infection and possibly trends in infection over time, the CDC studies include only self-selected rather than randomly selected individuals. As such, they cannot produce valid estimates of the number of people currently infected in the general U. S. population.

Michael F. Weeks and Daniel G. Horvitz are with Research Triangle Institute, Research Triangle Park, North Carolina. Peter L. Hurley and Robert A. Wright are with the National Center for Health Statistics, Hyattsville, Maryland.

The authors acknowledge and express their appreciation to the project staff at Research Triangle Institute and the National Center for Health Statistics, who have done the design work reflected in this paper.

Because the rate of HIV infection in the general population remains unknown, there is a real need to measure the prevalence of HIV infection in a random sample of the U. S. population. Such a study, if feasible, could provide the information necessary to plan effective treatment and prevention strategies and to target intervention efforts and resources to the groups and regions of the country that are hardest hit. The National Household Seroprevalence Survey (NHSS) is in response to this need.

The primary objective of the National Household Seroprevalence Survey is to estimate the prevalence of HIV infection in the U. S. noninstitutionalized civilian population aged 18 to 54 years. The study is being sponsored by the National Center for Health Statistics (NCHS), an agency of CDC. The project is being conducted in two phases. The first phase is a feasibility study that involves testing field procedures and methodologies in a pilot study, and a pretest. If the results of Phase 1 indicate that a national survey is both feasible and capable of producing new and useful data on the AIDS epidemic, Phase 2 will consist of a national survey of approximately 50,000 household respondents.

The Research Triangle Institute (RTI) was selected by NCHS to conduct the National Household Seroprevalence Survey, effective May 1, 1988. The initial pilot study was conducted in Allegheny County (Pittsburgh), PA, in January 1989 and involved in-home blood specimen collections and questionnaire administrations from an area frame household sample of 263 respondents. The pretest is scheduled to be conducted in a different site, probably Dallas County, TX, in the summer of 1989 and will involve approximately 1,600 respondents. A decision regarding the national survey will be made after Research Triangle Institute submits its pretest report, which is due in December 1989.

This paper describes the key design issues confronting the National Household Seroprevalence Survey, the implementation and results of the pilot study, and the plans for the pretest. A concluding section summarizes the current status of the feasibility phase of the project.

Design Issues

The National Household Seroprevalence Survey poses a number of problematic design issues that are largely attributable to the survey's objective of estimating HIV prevalence in the household population. Some of the more unique aspects of the survey include the sensitivity of the subject matter, low prevalence rate of HIV infection, disproportionate prevalence rate among segments of the population, concern about anonymity and privacy protection, the need to collect a blood sample and risk behavior information, and the large potential for both response and nonresponse bias. Although other surveys of sensitive topics may face some of these problems, the number and complexity of the design issues confronted by the National Household Seroprevalence Survey are unique in the experience of the Research Triangle Institute and NCHS staff charged with resolving them. The principal design issues confronting the National Household Seroprevalence Survey are described in this section.

Public Relations

A distinguishing feature of the National Household Seroprevalence Survey is the scope of the public relations (PR) effort required to maintain effective liaison with the media and other persons and groups interested in or concerned about the survey. While most national surveys involve some level of PR effort, there is usually limited public interest in the survey, and the PR component requires a relatively modest commitment of project resources. For the National Household Seroprevalence Survey, however, this is decidedly not the case. On the contrary, intense interest in the survey can be expected from a number of quarters, and it will be necessary to develop and implement effective PR strategies to be responsive to these groups.

Unlike most surveys, the media can be expected to provide extensive coverage of the planning and implementation of the National Household Seroprevalence Survey. AIDS is news, and a federally sponsored household survey designed to estimate the prevalence of the AIDS virus is big news. As the project staff is learning in the feasibility study, a significant PR effort is required to cope with such intense media interest. Since most survey researchers are inexperienced in this field, it is necessary to include PR professionals as an integral part of the project staff.

Liaison must also be established and maintained with other interest groups, including representatives of groups at risk for HIV infection, community leaders in areas where the sample is concentrated, national and local public health officials, and the survey research scientific community. To be successful, the National Household Seroprevalence Survey must involve these interest groups in a meaningful way and gain at least their tacit support, if not their endorsement, of the survey. Given the high level of public interest and the controversial nature of the survey, it is safe to assume that opponents of the project will receive their share of media coverage, and negative media coverage can be fatal to a survey as controversial as the National Household Seroprevalence Survey.

In the feasibility study an attempt has been made to involve interest groups by establishing and working closely with a number of advisory groups, including a national technical advisory panel, a national policy advisory panel, and local community advisory panels in field test sites. It seems clear that a continuation of some version of the advisory panel approach will be a necessary component of the national survey.

Sample Design

The National Household Seroprevalence Survey will be designed to yield 50,000 participants aged 18 to 54 who are in the civilian U. S. household population. A deeply stratified, multistage probability sample of households is planned. The first-stage sample of primary sampling units (PSUs) will be counties stratified on the basis of the most recent number of reported AIDS cases, a statistic that is available from Centers for Disease Control.

The second-stage sampling units will be small-area segments (such as census blocks) that will be stratified within primary sampling units using both public health and census data. The statistics chosen to create the second-stage strata will be those that reflect local area concentrations of persons with characteristics or behaviors that put them potentially at risk for HIV infection.

Some indication of the availability of suitable variables on which to base the second-stage stratification is provided, at least for metropolitan counties, by the findings in Dallas County, TX, the probable site of the National Household Seroprevalence Survey pretest. In Dallas the following information will be used for stratification purposes: counts at the census block level of at-risk persons who have visited the Dallas County Health Department's sexually transmitted disease clinic; counts of AIDS cases at the census tract level; counts of persons by ZIP code who have been tested for HIV in a public clinic; counts of reportable disease cases, such as syphilis and hepatitis B, at the block level; counts of persons treated for illicit drug usage at the block level; and the proportion of never-married males age 25 to 34 in 1980, also at the block level.

Users of public clinics in Dallas County are disproportionately black and Hispanic. To reflect a balanced racial mix, the high- and low-risk strata will be constructed so that they will have the same racial and ethnic composition as the county as a whole.

Anonymity and Privacy Protection

It has become increasingly clear that the National Household Seroprevalence Survey will not be feasible unless the survey design offers strict anonymity. Survey participants will be asked to provide a blood sample and to complete a self-administered questionnaire that asks for information about HIV risk behaviors. For the survey to be successful it is essential that study procedures safeguard the identity of survey participants and preclude the linkage of respondents' names and addresses with their blood test results or questionnaire responses.

Moreover, these procedures must be so creditable that they offer eligible sample persons the perception as well as the fact of strict anonymity.

Anonymity is necessary to overcome fears and concerns about potential consequences of participation in the National Household Seroprevalence Survey, particularly among those in the at-risk population. Gays, intravenous (IV) drug users, and others at risk for HIV are very sensitive to the potential for discrimination (for example, by employers, landlords, insurance companies) should their risk behavior status become known.

Privacy procedures currently in use and planned for the national survey include:

- No address information or other files or materials that could possibly identify a sample housing unit or sample area is retained at the central office after fieldwork begins.
- No names of sample persons or members of their household are ever recorded.
- Address information for study participants is destroyed in the field; it is never sent to the central office.
- The questionnaire is self-administered in private, sealed with tamperproof tape, and then sealed in a return envelope which is mailed immediately to Research Triangle Institute.
- Different ID numbers are used for the blood sample and questionnaire; the link is maintained in a secure computer dictionary at the central office.
- No personally identifiable information is recorded on the blood specimen or questionnaire.
- Written privacy procedures are developed for each aspect of the survey.
- An internal Privacy Officer is an integral part of the project staff and is responsible for monitoring compliance with the prescribed privacy procedures.
- All project staff sign an agreement that details the specific privacy procedures that apply to them and commits them to adhere to these procedures.
- An independent Privacy Committee monitors compliance with the privacy procedures. The Committee members make unannounced visits to the RTI central office, field data collection sites, and the offices of subcontractors.

Collection of Blood Samples

Collection of a blood sample from survey participants is a necessary and problematic requirement of the National Household Seroprevalence Survey. The key issues here are where the blood collection activity takes place, who collects the blood, the method of blood collection, and the logistics of shipping blood samples to a laboratory for analysis.

A basic premise of the approach to the blood collection requirement is that respondents should be able to give the blood sample in their homes. To require respondents to travel to some central site outside of the home to provide the blood sample would adversely affect the response rate as well as pose some difficult logistical problems. On the other hand, provision should be made to accommodate the occasional respondent who wants to leave the home to give the blood sample.

With regard to who will collect the blood samples, four alternatives have been evaluated:

1. Employ survey interviewers and train them to be phlebotomists.
2. Employ phlebotomists and train them to perform the survey interviewing tasks.
3. Employ both interviewers and phlebotomists, but have the interviewer work alone during the initial household contact. Under this approach, the interviewer would be responsible for screening households, identifying eligible sample persons, and securing their participation. The interviewer would then have to make an appointment for a follow-up visit by both the interviewer and phlebotomist to collect the blood sample and complete the questionnaire administration.
4. Employ both interviewers and phlebotomists and have them work together as two-person survey teams, eliminating the need for a return visit to collect the blood sample (although the respondent would still have the option of making an appointment for a later time and different place, if desired).

The first two options have been eliminated as impractical, for a variety of reasons, and Option 4 is preferred over Option 3. While the team approach is somewhat more expensive, intuition and experience to date suggest that it is important to be able to complete data collection on a single visit and that the need for a follow-up visit to complete data collection would probably have a detrimental effect on the response rate.

Respondents will be offered the traditional venipuncture method of blood collection. If a person objects to this method, the phlebotomist will offer a fingerstick and microtainer method as an alternative. Respondents will not be given an initial choice of the two methods because venipuncture offers several advantages over the fingerstick method. It is easier, quicker, less painful, and, based on pilot study results, has a higher success rate in terms of producing an adequate quantity of blood.

Collecting blood samples in a household survey adds some complexity to field operations, involving special packaging, refrigeration, and prompt shipment to a laboratory for analysis. It is also desirable to have a field-reporting system in place that is capable of tracking the shipment and receipt of blood specimens. Although the associated field procedures are numerous and detailed, experience has shown the process to be feasible.

Collection of Risk Behavior Data

After the blood sample has been taken, each respondent will be asked to complete a self-administered Sample Person Questionnaire (SPQ). The primary function of the Sample Person Questionnaire is to provide data that will enable us to classify respondents by level of risk for HIV. Assuming valid responses, the relationship of HIV status to risk behavior could then be studied. The SPQ level-of-risk data could also be used to help adjust the HIV prevalence estimates for nonresponse, provided

the questionnaire is completed by a valid subsample of those who refuse to provide a blood sample.

In addition to demographic data and questions about whether the respondent has ever had specific sexually transmitted diseases, the Sample Person Questionnaire covers recent and past IV drug use and recent and past sexual behavior. Specifically, respondents are asked to report whether they used needles to inject drugs and the frequency with which needles were shared in the 12 months before the interview and since January 1978. The sexual behavior questions ask the number of different sexual partners by sex of partner, frequency of use of condoms by sex of partner, and frequency of receptive anal intercourse in the 12 months before the interview and since January 1978. Questions about sex with IV drug users, bisexual men, and prostitutes since January 1978 are also asked.

Despite the strict anonymity provided to sample persons, some might still be suspicious that their SPQ data could be linked back to them. To reduce such concerns and to further encourage candid answers to three key sensitive drug and sexual behavior questions, these questions will be asked indirectly, using the "item count" technique. This technique is similar to randomized response in that it provides information about the sensitive behaviors of the sample persons as a group, but does not provide any information about which of the sample persons individually have engaged in those behaviors.

Briefly, the item count technique imbeds the sensitive behavior question in a set of, say, four behavior questions and asks a random half of the respondents to report how many of the total behaviors they have experienced, but not which ones. The other half of the respondents are shown the same set of questions minus the sensitive question and asked to report how many of these behaviors they have experienced. The difference in the mean number of behaviors reported by the two half-samples provides an estimate of the proportion of respondents who have experienced the sensitive behavior.

A Spanish version of the Sample Person Questionnaire will be used in the pretest and national survey. Two questionnaire administration procedures are being considered for nonreaders: (1) having the interviewer read the questions to the respondent, and (2) having the respondent use headphones to listen to an audio version of the questionnaire. In either case the respondents will be expected to mark the appropriate response categories in their questionnaires without assistance from the interviewer. Both methods are being evaluated in simulated interviews with nonreaders.

Blood Testing

The blood samples collected by the field staff will be sent via an overnight delivery service to a central laboratory for analysis. The protocol for testing blood samples for HIV antibodies is straightforward. An initial enzyme-linked immunoassay (ELISA) test is performed on all blood samples. If the test is negative, the analysis result is reported as negative. If the test is positive, two additional ELISA tests are performed. If both are negative, the analysis result is reported as negative. If one or both of the additional tests are positive, a Western blot assay is performed and the result of that test is reported as the analysis result.

External quality assurance procedures will include unannounced onsite inspections by RTI and NCHS staff and by members of the Privacy Committee, and use of blinded performance evaluation samples. Use of a single laboratory rather than multiple locations simplifies quality assurance procedures and eliminates the potential for interlaboratory variation. All tests must be performed using Food and Drug Administration (FDA)-approved materials and procedures. It is possible to have the testing process result in an indeterminate outcome. Consideration is being given to using additional laboratory procedures developed by the U. S. Army to minimize the number of final indeterminates.

Reporting of Blood Test Results

Whether to report blood test results to survey participants who want to know their HIV status is a difficult issue that has received considerable attention by National Household Seroprevalence Survey project staff and our advisors. If blood test results are reported, it is clear that the reporting should be done by a trained HIV counselor, regardless of the test result (positive, negative, or indeterminate). On the other hand, such a reporting system could be difficult to implement and still preserve the anonymity of the survey participant. Thus, the tension here is between the desire to report blood test results to those who want to know and the necessity of providing anonymity and privacy protection to survey participants.

Initially, an elaborate scenario was developed that would make test results available while still preserving anonymity. Under this approach, the participant is given a secret number in a sealed envelope at the time the blood sample is taken. The secret number is known only to the participant and is linked to a different number on the blood sample through a secure computer dictionary. To obtain the blood test result, the participant must call one of the local counselors on a list provided, give the counselor their secret number, and make an appointment to meet with the counselor. Before the meeting with the participant, the counselor would call Research Triangle Institute to obtain the blood test result corresponding to the secret number.

In carrying out this process, participants would never need to reveal their names, addresses, or telephone numbers to the counselor, nor would the survey staff at the central office need to know any personally identifiable information to provide the test result that matches the secret number. However, it became clear when testing this approach in focus groups that it was too complicated to be readily understood and that it lacked the perception of anonymity, if not the fact. The focus groups were also concerned about the need to visit a counselor and the associated risk of a breach of anonymity. Given this feedback, it was decided to abandon the counselor-based reporting system.

Now it has been concluded that the best solution to this problem is not to offer test results to survey participants, explaining that there is no way to do this and still

provide adequate safeguards of their anonymity. However, arrangements will be made with a local HIV testing and counseling center and survey participants who want to know their HIV status will be referred to the local agency for a second test and appropriate counseling. In many localities anonymous HIV blood testing and counseling is available free of charge. In locations where there is a charge for this service, a voucher system will be developed so that the cost of this service will be direct-billed to the project.

Attainment of a High Response Rate

Attainment of a high response rate is important to enhance the credibility of the National Household Seroprevalence Survey in the public mind. The response rate serves as a measure of public acceptance of the survey and, as such, will be of considerable interest to both supporters and opponents of the survey. A high response rate also helps reduce the risk of nonresponse bias, although as noted in the following section, one could achieve a very high response rate in the National Household Seroprevalence Survey and still have a serious nonresponse bias if at-risk persons are overrepresented among the nonrespondents.

Procedures designed to maximize the response rate include:

- public relations efforts with the media and other interest groups;
- use of a toll-free Project Hotline to answer any questions sample persons may have or to verify the identity of field survey personnel;
- the mailing of a household lead letter and use of a sample person letter signed by the director of the Centers for Disease Control and the Surgeon General;
- use of a 7-minute videotape, with opening and closing remarks by the Surgeon General, played by the interviewer on a portable videotape player;
- field procedures that allow sample persons to participate on an anonymous basis;
- the capacity to collect blood by a medical professional in the sample person's home;
- use of an interviewer and phlebotomist team approach, so that blood collection can take place immediately, without the need to schedule an appointment for a follow-up visit;
- payment of a $50 cash incentive to participants;
- follow-up contacts with refusals by a specially trained field staff; and
- a special follow-up effort with a sample of nonrespondents that might involve additional incentives to obtain a blood sample or at least the Sample Person Questionnaire.

Nonresponse Bias

If sample persons belonging to a high-risk group, such as gays or IV drug users, participate in the National Household Seroprevalence Survey at a significantly lower rate than do low-risk persons, then the prevalence estimate derived from the sample data could seriously underestimate the true prevalence (Table 1). Since finding a person with HIV-positive status is a rare event, even if 90 or 95 percent of a sample of eligible persons participate, the bias due to nonresponse would still be large if those refusing to participate include a high proportion of the at-risk persons in the sample. On the contrary, if persons of different risk levels participate at similar rates, even if the rates are relatively low, then the nonresponse bias problem is negligible (Table 2).

It is possible to adjust the HIV prevalence estimate for differential nonresponse between those at risk and those not at risk in a sample, but this presumes knowledge of the risk classification for respondents and for at least a sample of the nonrespondents.

Response Bias

In view of the sensitivity of the questions asked in the Sample Person Questionnaire, some high-risk respondents who provide a blood sample might deny the behaviors that put them at risk. While this would not bias the overall estimate of HIV prevalence, it would lead to an overestimate of HIV prevalence among low-risk persons. Without knowledge of the levels of misclassification in the responses to the risk behavior questionnaire, it would be decidedly inappropriate to attempt to estimate HIV prevalence by risk category. Despite the presence of response errors in self-reported risk behaviors, the SPQ data might still be used effectively to reduce the nonresponse bias in the HIV prevalence estimate.

Quality Assessment

A number of techniques for assessing the quality of the NHSS data and the effects of nonresponse and re-

Table 1. Effect of dissimilar participation rates by risk level on the validity of an HIV prevalence estimate

Risk level	HIV prevalence rate	Sample size	Participation rate (%)	Number participating	Expected HIV+ in sample
High	200/1000	50	30	15	3.0
Medium	50/1000	50	40	20	1.0
Low	1/1000	900	96	864	0.9
All	13.4/1000	1000	90	899	4.9

NOTE: HIV prevalence estimated from this sample is: 4.9/899 = 0.0055 or 5.5/1000. The relative bias in this estimate is: (0.0055 − 0.0134)/0.0134 = −0.589 or −58.9%

Table 2. Effect of similar participation rates by risk level on the validity of an HIV prevalence estimate

Risk level	HIV prevalence rate	Sample size	Participation rate (%)	Number participating	Expected HIV+ in sample
High	200/1000	50	63	31.5	6.3
Medium	50/1000	50	67	33.5	1.7
Low	1/1000	900	65	585.0	0.6
All	13.4/1000	1000	65	650.0	8.6

NOTE: The HIV prevalence estimated from this sample is: 8.6/650 = 0.0132 or 13.2/1000. The relative bias in this estimate is: (0.0132 − 0.0134)/0.0134 = −0.015 or −1.5%.

sponse bias are available. These include: (1) comparison of blood test results with reported risk behaviors, (2) comparison of NHSS statistics with data from other sources, (3) stratification of second-stage sampling units, (4) special follow-up procedures with a sample of nonrespondents, (5) the item count method of asking questions indirectly, and (6) statistical verification.

A comparison of respondents' HIV blood test results with their self-reported risk behaviors will provide a measure of the validity of the SPQ data, since persons who test positive for HIV should report at least one risk behavior. This technique has limited value, however, because it provides no information about the validity of the risk behavior data provided by respondents who are not HIV positive.

A standard methodology for assessing the quality of survey data is to compare statistics from the survey with data from other sources. Although reliable statistical data on the prevalence of HIV risk behaviors in the general population are sorely limited, there is an opportunity to compare NHSS data with extant hepatitis B data. Since a person's blood status with respect to hepatitis B is a surrogate measure for certain HIV risk behaviors, the NHSS blood samples could be tested for hepatitis B as well as for HIV antibodies and our results compared with national data on hepatitis B prevalence available from the National Health and Nutrition Examination Survey.

Stratification of the second-stage sampling units on the basis of HIV risk level will also serve as a quality assessment measure. By comparing high- and low-risk strata, some measure of the effects of risk level on participation is obtained. It should be recognized, however, that this is a marginal test. It depends on an effective classification of the area sampling frame into high- and low-risk strata, which is problematic since the households in the areas assigned to the high-risk strata will still have many more low-risk eligible persons than high-risk eligible persons.

Another quality assessment measure is a special follow-up study of a sample of nonrespondents. After the normal fieldwork has been completed, a sample of the nonrespondents could be drawn and special procedures used in an effort to secure some level of participation from them. The sample could be divided into two random subsamples. For one half-sample an elite subset of the field staff would make a final follow-up contact and attempt to secure full participation (blood sample and Sample Person Questionnaire), offering an enhanced incentive. The data collected from this subsample would be used to derive a prediction equation for HIV infection among nonrespondents as a function of risk level reported in the Sample Person Questionnaire.

The other half-sample of nonrespondents would be offered the standard incentive for full participation but asked to complete the questionnaire only; they would not be asked to also provide a blood sample. The HIV status of this subsample will be estimated using the prediction equation developed from the other half-sample, thereby providing a basis for estimating and adjusting for the bias due to nonresponse to the National Household Seroprevalence Survey.

It should be recognized that while the Nonrespondent Follow-up Study would address the issue of differential participation by risk groups, its success depends on securing a creditable participation rate (at least 70 percent) for each subsample.

The item count method for asking sensitive HIV risk behavior questions, described above, will also serve as a quality assessment measure since it has the potential of providing estimates of the response bias in the direct question responses to key risk behavior questions in the Sample Person Questionnaire. It is, however, a relatively new and untested technique.

Finally, a statistical verification procedure can be used to assess the quality of the NHSS data. Statistical verification requires the existence at the primary sampling unit level of public health files, preferably computerized, that have information about the risk behavior of some of the individuals in the National Household Seroprevalence Survey sample. The public health data can be used to assess the validity of the risk behavior questionnaire data provided by at-risk respondents as well as the survey participation rate of at-risk sample persons.

Quality assessment of the questionnaire data can be implemented by matching a data file containing only an encoded census block number and demographic data from the questionnaire (birthdate, race, sex, and marital status) against the same set of data items in the public health files available for the primary sampling unit. The degree of agreement between the risk behavior reported in the National Household Seroprevalence Survey and the risk behavior recorded in the public health files for matched cases will indicate the quality of the question-

naire data reported by NHSS respondents at risk for HIV.

The same matching procedure can be used for eligible sample persons who refuse to participate. The household screening form has demographic data (age, race, sex, and marital status) for each adult member of the household, and this information for NHSS nonrespondents can be used in the matching process. The matched cases will provide the data needed to estimate the HIV risk status of NHSS nonrespondents. These estimates can be compared with the same estimates derived for the matched cases among the NHSS respondents, yielding a measure of the level of differential nonresponse to the survey for persons at risk and persons not at risk for HIV.

It should be noted that no names or addresses need be used in either matching process in the statistical verification procedure. The anonymity and privacy protection afforded NHSS sample members can be fully respected and maintained at all times.

The primary problem with statistical verification in the context of HIV infection is that the number of matched cases will generally be small, on the order of 15 to 25 per 1,000 sample persons. This limits its utility as a quality assessment method for the pretest, but may not do so for the national survey.

The most efficient and most common method of measuring the quality of survey data is direct assessment. This technique uses existing files, samples, registries, or other lists of individuals who have a characteristic of interest; for example, a hospitalization within the past 12 months. To assess the quality of data collected about the characteristic of interest, a sample is selected from the corresponding list. The survey instrument is then administered to this sample and the data reported in the survey compared with the data in the file that was used to generate the sample.

With regard to the National Household Seroprevalence Survey, public health files contain data on behaviors that put a person at risk for HIV infection and these could serve as an effective sampling frame for direct assessment. This method is not acceptable for the National Household Seroprevalence Survey, however, because it involves the release to the survey contractor of address information of at-risk persons without their prior consent. This is viewed by many as unethical, even if the at-risk addresses were mixed with control addresses so as to blind the survey contractor to the risk status of any specific address. In addition, the agency providing the information would know which individuals were in the direct assessment sample, which would constitute a breach of anonymity. Finally, sample persons would need to be informed how they were selected for the NHSS survey. Since persons who are at risk for HIV are very sensitive to the potential for discrimination, it is not likely that they would agree to participate in the survey after learning how they were selected.

Monitoring of the Fieldwork

Given the unique and demanding design features of the National Household Seroprevalence Survey, it was clear from the inception of the project that an extraordinary level of supervision and control over the fieldwork would be required to conduct the survey successfully. Judging by the various options available with regard to field communications, there is convincing evidence that conventional methods of communication among central office staff, the offsite field supervisors, and the interviewer staff were too slow and unreliable to provide the information flow necessary to manage the project effectively. The only system that would provide the necessary level of communications was one that established computer-based links among all three parties.

Accordingly, Research Triangle Institute has developed a fully automated communications system and is in the process of testing and enhancing this system during Phase 1 of the NHSS. In this system each interviewer and offsite field supervisor is equipped with a microcomputer that is linked via a telecommunications hookup with a host computer at Research Triangle Institute. At the conclusion of each day's fieldwork, the interviewers key into their computer case-specific codes indicating the results of their field efforts that day. These data are automatically transmitted to the RTI host computer between midnight and 5:00 a.m. the next morning. Between 5:30 a.m. and 7:00 a.m., the field supervisors' computers automatically call the RTI host computer and retrieve the field production data for their interviewers. At any time after 7:00 a.m. the field supervisors' computers can produce a variety of reports showing the status of each interviewer's workload, based on production data keyed through the previous evening. Meanwhile, the RTI host system processes the new field data and incorporates it into the central project control system database. By 12:00 noon each day, computer-generated reports are available for project management showing the overall status of the fieldwork as of the previous evening.

This automated communications system also provides for the exchange of computer mail messages among the central office, supervisors, and interviewers. In addition, the system will be expanded to include time and expense data as well as production information. This will facilitate fiscal management of the survey as well as expedite the payment of the interviewer staff.

Pilot Study

Background

The pilot study was conducted in Allegheny County, PA, in January 1989 in cooperation with the Allegheny County Health Department and its Community Advisory Committee. It had two overall objectives: (1) to evaluate a set of methods and procedures for possible use in a national survey, and (2) to determine whether persons selected for the survey would participate.

Washington, DC, was selected as the initial site for the pilot study. However, concerns about the study on the part of District health officials and certain other community leaders triggered a wave of generally negative publicity that would have made it difficult to obtain an unbiased measure of the feasibility of such a study.

Consequently, it was decided to cancel the District pilot and select another site. After considering a number of alternative sites, Allegheny County was selected as a fairly representative community that would probably provide a good initial measure of the public's willingness to participate in such a survey.

Methods

The pilot study was designed as an area probability household survey. All households in Allegheny County had an equal chance to be selected for the survey. Two-person survey teams consisting of a field interviewer and a phlebotomist visited each household that was selected for the survey. The interviewer determined, by completing a screening interview with a household resident, whether the household contained any age-eligible persons. Age eligibility for the pilot study was defined as being 18 to 54 years old at the time of the household interview.

If there were age-eligible residents, the interviewer randomly selected one of these persons to be asked to participate in the study. The study was explained in detail to the sample person through the use of an explanatory letter, a special videotape presentation, and an informed consent form.

If the person agreed to participate, the phlebotomist collected a small blood sample, using either a venipuncture or fingerstick technique. The person was then asked to complete, in private, a 10-minute Sample Person Questionnaire. This questionnaire requested basic demographic information as well as information about factors that may put the individual at risk for HIV infection. The questionnaire was completely self-administered. After completing it the respondent sealed the questionnaire with tamperproof tape and put it into a return envelope for mailing back to Research Triangle Institute. Persons who provided both a blood sample and a completed questionnaire received a $50 incentive payment.

The pilot study was designed to allow survey respondents to participate on an anonymous basis. No names were ever recorded, address information was destroyed in the field after completion of data collection, and no link was possible between a survey participant and their blood test result or questionnaire responses. After fieldwork was complete, a separate Nonrespondent Follow-up Study was conducted in an attempt to better understand reasons for nonresponse.

Data Collection Results

The pilot study sample included a total of 473 occupied sample housing units. Of these, 450 were successfully screened, for a screening response rate of 95 percent.

A total of 308 eligible sample persons were identified in the screened housing units. Of these, 263 participated in the pilot study by giving a blood sample and completing a Sample Person Questionnaire, for a sample person response rate of 85 percent. Of the 45 sample persons who did not participate, 36 refused, 5 could not be contacted, and 4 could not be interviewed due to a physical or mental impairment.

The overall response rate was 81 percent, computed as the product of the screening and sample person response rates. Project staff found this to be an encouraging result, given the initial concern about whether sample persons would participate in such a survey.

Item response rates were also computed for the 22 questions in the Sample Person Questionnaire. All of the items had at least a 92 percent response rate and all but three had a response rate of 98 percent or higher. The latter group included the questions that addressed high risk behavior.

Six percent of questionnaire respondents reported at least one of five selected HIV risk behaviors. While this cannot be viewed as a valid estimate of HIV risk in Allegheny County, it does indicate that at least some persons with HIV risk behaviors decided to participate in the survey.

Plans for the Pretest

The overall objectives of the pretest will be to (1) further refine survey methods and procedures for possible use in a national survey, (2) obtain a second, more precise, measure of the willingness of persons to participate in the survey, and (3) implement methods designed to assess the quality of the data.

Discussions are currently in process with the Dallas County Health Department and its community advisory panel concerning the possible conduct of the pretest in Dallas County during July to September 1989. Changes in the pilot study design being considered for the pretest, subject to approval by the health department and the advisory panel, include:

- An increase in the sample size to approximately 1,600 respondents. This will increase the number of at-risk persons in the sample and provide more precise estimates of response rates.
- Construction of sampling strata on the basis of expected HIV risk level, using information that is correlated with the geographic distribution of persons at risk for HIV infection. A comparison of response rates among the risk strata will provide some indication of the willingness of at-risk persons to participate in the survey.
- Use of a higher sampling rate in the strata with an expected elevated prevalence of at-risk persons, to include more at-risk persons in the sample.
- Development of a short "refusal conversion" videotape, for use in recontacts with persons who initially refuse to participate. If necessary, the video can be left with the person to watch at his or her leisure.
- Use of the venipuncture blood collection method during the initial presentation to the sample person. The fingerstick method will be offered only if the sample person objects to the venipuncture method.
- Expansion of the Sample Person Questionnaire to include additional questions about risk behaviors. This will provide a better measure of risk status for survey participants.
- Use of the item count method of asking sensitive questions indirectly, to minimize response bias.

- Redesign of the Nonrespondent Follow-up Study. Instead of asking pretest refusals about their reasons for refusal, as was done in the pilot, special procedures will be used to attempt to secure their participation. Half of the refusal group will be asked to provide both the blood sample and the questionnaire, while the other half will be asked to complete the questionnaire only. The Nonrespondent Follow-up Study will help increase the overall response rate as well as provide information on the relationship between nonresponse on the one hand and HIV prevalence and risk behaviors on the other.
- Testing the blood samples for hepatitis B as well as for HIV antibodies. A comparison of the prevalence of hepatitis B in the sample with known hepatitis B prevalence rates for the local area will provide a measure of the willingness of at-risk persons to participate in the survey.
- After completion of the pretest, use of the statistical verification procedure described above to anonymously match the survey data file with data files of at-risk persons maintained by outside sources, such as a public health clinic. The matches will be analyzed to provide statistical estimates of the response rate for at-risk persons and the accuracy of their self-reported risk behavior data.

Conclusion

The feasibility study of the National Household Seroprevalence Survey is now at the half-way point, with the pilot study completed and the pretest still ahead. The pilot study had two objectives: (1) to evaluate a set of survey methods and procedures for possible use in a national survey, and (2) to assess the willingness of a representative community to participate in the survey. With regard to the first objective, the assessment is that the survey methods and procedures used in the pilot proved to be generally feasible, subject to further refinement in the pretest. Key lessons learned to date include:

- The importance of public relations with the media and other groups interested in the survey.
- The necessity and feasibility of offering survey participants a set of field procedures that give the perception as well as the fact of anonymity.
- The impracticality of reporting blood test results to participants.
- The feasibility of the interviewer and phlebotomist survey team approach.
- The feasibility and apparent effectiveness of equipping interviewers with videotape players and showing participants a videotape in their homes.
- The feasibility of the procedures for packaging, shipping, and analyzing the blood samples.
- The feasibility and utility of the computer-based field communications system.

With regard to the second pilot study objective, the 81 percent overall response rate was encouraging and exceeded the expectations of most project staff and advisors. The high item response rates for the self-administered risk behavior questions were also encouraging. Finally, the fact that 6 percent of respondents reported at least one selected HIV risk behavior suggests that at least some persons at risk for HIV will decide to participate in the survey.

The pretest will provide an opportunity to (1) further refine survey methods and procedures, (2) obtain more precise data on the willingness of persons of different HIV risk levels to participate in the survey, and (3) assess, albeit to a limited extent, the quality of the data and the effects of response and nonresponse bias. If these three goals are achieved, a reasonably informed assessment as to the feasibility of a national household seroprevalence survey should be achieved.

Samples for Studies Related to AIDS

Richard Warnecke, Recorder, and Johnny Blair, Chair

The discussion focused on problems of coverage (or definition of the sampling frame) and nonresponse. Each is discussed separately below (also see Groves, in this volume).

Coverage. Participants generally agreed that the issue of what constitutes an appropriate frame for studies related to AIDS depends largely on the research question. If the objective is to characterize the population at risk either by its behavior or by seroprevalence, then carefully drawn probability samples with known statistical properties are required. Such samples must allow estimation of rates in the general population and enable estimation of possible bias. Participants cited many national studies that exemplify this type of survey. However, the studies described in this session's papers and in the succeeding floor discussion typically had less ambitious objectives and focused more on public health issues for local populations. That data be "good enough" was a frequent refrain. Much of the discussion, therefore, centered on efficient sample frames that, while addressing the immediate needs of the investigator or sponsor, have limited generalizability beyond the subjects selected for the study. One clear implication of this approach is that reports of the results of such studies must pay careful attention to the limitations of generalization.

Further discussion clearly reflected problems common to all studies of AIDS that try to define an appropriate sampling frame. Several important characteristics of the targeted high-risk populations were identified:

- It is largely self-identified.
- It is stigmatized by the larger society.
- It may be subject to sanction if publicly identified.
- Some elements of the population engage in illegal acts.
- It is a rare population.
- It is a population in which those highly visible members in identifiable geographic clusters may differ significantly from less visible, more integrated members.
- It is not a single group but includes a subgroup that is at particularly high risk of being infected with HIV, a subgroup that has been infected, and a group that is at risk of transmitting HIV.

Because of these characteristics that define targeted populations, obtaining an appropriate sampling frame, even for getting data that are "good enough" raises complex issues. It was pointed out that the most fundamental question may be whom to include in the frame. The target population is not identifiable by some uniform demographic or other characteristic and cannot be located through a single source such as geographic location. Identifying and obtaining access to separate subgroups defined by level of risk adds additional complexities. These issues were considered in the papers presented and in the floor discussion that followed. Various approaches were taken to sampling different target groups and no one satisfactory approach exists. The studies described by Capell and Freeman have attempted to identify those at risk for HIV or those who might transmit the virus. To locate the subjects for the sample, these studies use traditional RDD telephone sampling methods with geographic stratification. Two issues related to this approach to sampling were discussed at length. The first of these was the role of convenience samples. This has been the approach taken by the pilot study for the seroprevalence survey being conducted by Research Triangle Institute (RTI) and has been found useful if the target group is homosexual or bisexual men who are likely to congregate or reside in certain areas. In the discussion it was noted that even in such areas some of these men are likely to be very discrete about their lifestyle or may not even acknowledge themselves as part of these groups even if their sexual practices include high-risk behaviors. Convenience samples may not identify such individuals and, further, these individuals may not report the behaviors of interest in a standard RDD sample.

Several research issues were discussed as they relate to strategies for sampling frames of gay and bisexual men.

1. The extent of telephone coverage among this popu-

Richard Warnecke and Johnny Blair are with the Survey Research Laboratory, University of Illinois.

lation needs to be specified. It was noted that some large national surveys may provide these data. A more general issue may be the need to identify the appropriate ways of locating respondents who are not linked to traditional households.

2. Using more qualitative methods, such as those being developed at the Cognitive Laboratory at NCHS, researchers need to determine the extent to which members of this population will identify themselves in response to screening questions and to develop appropriate formats for such questions.
3. The efficiency of mixed frame samples that combine lists with RDD approaches must also be assessed.

Problems in locating representative samples of AIDS patients and male and female prostitutes were also discussed (see Fleishman, Bradford, and Berry, respectively, in this volume). AIDS patients need to be identified and interviewed so that the health services delivery issues related to caring for them can be established. This population is usually sampled from service provider lists. Selection into a sample may be heavily influenced by gatekeepers. Thus, developing samples of this population requires both access to information (frames) and cooperation of intermediaries. Resulting problems are akin to those involved in studying other protected populations such as the chronically ill, the elderly, and children. Because biases may result from incomplete information or access, researchers must find means to reach the population directly and to secure enough cooperation to estimate biases that result from institutional list samples.

In the remaining papers, male and female prostitutes were also discussed in terms of sampling issues separate from those discussed above. These issues relate to identifying the populations and devising methods by which they can be sampled efficiently. Two salient population characteristics were discussed extensively: first, prostitutes tend to be geographically mobile and to have intermittent careers; second, female prostitutes in particular are not always accessible or identifiable because gatekeepers may present a formidable barrier in some situations. As a result, sampling methods such as some variation of capture-recapture techniques were suggested for consideration.

It was also pointed out that because these may be highly mobile groups, sampling techniques that do not use the household as the unit of enumeration will probably be of high priority in future research. Berry and Bradford have attempted to use multiple informants who are working with prostitutes as enumerators. However, not much is known to date about the effects of these approaches on coverage and the sampling biases in the resulting frames.

Nonresponse. The second major theme of the floor discussion concerned noncoverage, especially bias resulting from refusals and lack of access to potential respondents. Participants generally agreed that the problems of nonresponse relate partly to the characteristics of the population discussed above. Nonresponse is likely to be correlated with the attributes of interest, resulting in severe biases due to nonresponse not typical in other surveys. Horvitz pointed out in his discussion paper that when nonresponse is correlated with the attribute of interest and the attribute is rare in the population, the resulting bias is likely to be substantial. Thus, a major problem is to obtain good cooperation from the respondents. As with the coverage issues, the problems of cooperation are likely to vary depending on the subgroup within the target population.

There was consensus that in those studies (for examples see Capell and Freeman, in this volume) which focus on general population surveys, three related issues seem most critical: tolerance for the subject matter; invasion of the respondent's privacy and informed consent; and the interviewer-respondent relationship, specifically the need to convince the respondent to respond to the behavioral items in question. Because these issues are concerned with the way in which the interview and related subject matter are introduced to the respondent and his cooperation requested, the researcher's first priority must be to gain support for such studies within the target population. In part, refusals are to be expected because of the societal disapprobation attached to the behaviors in question. Thus, making the population aware of the study and its purposes might be useful, but it was pointed out that this approach has had varying effectiveness.

A related point discussed was that the effectiveness of recruiting subjects from the target population in general surveys using RDD may have much to do with the interviewer's approach to the respondent and how well the interviewer is able to convince the respondent of the legitimacy of the interview. Bradford discussed the importance of using "insiders" as interviewers, particularly when accessing populations that are engaged in illegal activities such as prostitution or intravenous drug use. It was suggested that data from studies using informants as interviewers be examined further. Two such studies were mentioned: one by Ellison and Boles at Georgia State University on male prostitutes and another by Wiebel at the University of Illinois School of Public Health. Interviewer training of the respondent as part of the interviewing process was also suggested.

Also discussed was the effect of requirements for informed consent on cooperation and on interviewer-respondent rapport. Institutional review board (IRB) requirements for consent at the beginning of the interview appear to have created problems with respondent recruitment in at least one research setting in which the investigators were required to tell respondents at the outset that the interview would contain detailed questions about their sexual practices. There was general consensus that such demands might have been excessive and that a consent process conducted in stages might have achieved the same purpose and not induced respondents to refuse initially. The approach recommended was one in which as the questions reached the point where detailed inquiry into potentially sensitive behaviors was required, the interviewer would inform respondents about the content and advise them of their right to refuse to proceed. Such an approach would at least provide some data on these respondents. Moreover, it was argued that respondents become increasingly informed as an interview progresses and hence can make

better decisions after they have experience with the interview.

The issue of IRB constraints is really a special case of the more general question about how respondents are to be recruited for such surveys. Because of the sensitivity of the subject matter and the nature of the questions, respondent recruitment and screening techniques may benefit from preliminary studies employing cognitive research techniques. It was suggested that because these studies must meet local IRB requirements, they be done before final clearance was requested and, in fact, could add to the legitimacy of the methodology presented for approval.

Also discussed as a clearance issue was the role of gatekeepers. As noted above, the gatekeeper is an important factor both in obtaining the lists that may be used to develop a frame and in accessing particular selected respondents. However, it was pointed out that a gatekeeper, by definition, creates a two-stage process in obtaining access to the respondent and will inevitably reduce access and introduce bias. The varying roles of the gatekeeper as described by Fleishman, Bradford, and Berry were discussed. It was pointed out that the gatekeeper could be a barrier to access (Fleishman) or could facilitate it (Bradford). Berry mentioned the need to contact and work with gatekeepers but reported no interaction with them. The community surveys described by Freeman and Capell also addressed the gatekeeping function. The former was sponsored by a local gay group that actually received the funding and subcontracted the research. In the latter study, gatekeepers were not used but were recommended to improve future response rates.

Several unresolved questions about nonresponse were suggested as topics for controlled experiments:

1. How should these respondents be approached? What form should the screening and recruitment in the survey take? What can researchers learn from studies of these items in cognitive research laboratories?
2. Who is the best type of interviewer for these studies? When does an "insider" used as an interviewer improve cooperation? What are the costs in terms of possible bias? Are there dangers of invasion of privacy in such situations?
3. How accurately do potential respondents report their status? Is the size of this population being underestimated because of misreporting? How might the magnitude of such underestimation be measured?
4. When is a gatekeeper effective in obtaining access to these populations? What biases can be anticipated from using gatekeepers? How are these individuals best approached?

Synthesis

Sample designs for AIDS-related studies discussed here have generally fallen into two broad categories: first, samples of the general population, which may or may not include oversampling particular demographic or high-risk behavior groups; and, second, samples limited to special populations of particularly high-risk or HIV-infected persons. Therefore the sample design issues for these general classes of studies are summarized separately, since their methodological problems and effective sampling procedures are quite different.

General population samples. To the extent that surveys of the general population aim only to assess knowledge, attitudes, and behaviors of the population at large, the design problems are mainly those of questionnaire design, measurement, and respondent cooperation. The sample designs do not differ from those of other surveys of the general population. Many AIDS studies, however, attempt to oversample particular subdomains of the total population. These target subdomains may be defined simply by demographic characteristics—such as blacks or Hispanics—or may be defined by behavior characteristics—such as homosexual or bisexual males—or even by some multifaceted set of high-risk behaviors. It is in this latter area that problems particular to AIDS surveys arise.

Methods for oversampling demographic groups generally take two forms: the screening of a general population sample large enough to locate desired numbers of the demographic group; or, if the target group clusters geographically, disproportionate stratification or cluster design procedures to improve the efficiency of locating target population members.

These methods are well known and not particular to AIDS studies. It is the situation that occurs when the target group is rare or hidden that is of most interest. If the behavior characteristic must be combined with a demographic one (such as homosexual blacks or Hispanics), the difficulty increases rapidly. In these situations, neither screening nor stratification has been effective.

For cost reasons, most efforts to oversample target groups such as homosexual males have been conducted by telephone using standard RDD procedures. Even if one assumes that an acceptable operational definition of the target group has been developed, the problems of screening on such a highly sensitive characteristic are enormous. If, alternatively, the target members are identified later in the interview, then one faces the problem of including large numbers of nontarget population members in the sample, with the attendant effect on cost. Even using telephone samples (and putting aside questionnaire issues), the cost of locating such populations by simple screening procedures is often prohibitive.

The issue of coverage is also related to screening for samples of many of these populations. Today most subdomains of the U.S. population are well covered in telephone frames. Still, to the extent that target population members are found in nontelephone households—whether disproportionately or not—frame coverage bias is increased. Although this bias may be small in absolute terms, when considered relative to the size of an already rare population, the effects on design efficiency may not be trivial. Little is known about the true coverage of these populations in telephone frames. Only to the extent that target population members have characteristics—mainly, being young and in low-income households—that are known to be related to nontelephone ownership can the undercoverage be estimated.

Disproportionate stratification has been used in some studies but has not produced the major increases in efficiency realized for other populations. This seems due primarily to the fact that the target populations are not heavily clustered; although sometimes a contributing factor is the lack of appropriate data for defining strata and determining sample allocations.

The amount of clustering expected depends on the particular population. In some areas, for example, there is heavy clustering of homosexual males in certain neighborhoods; but in many others, such clustering does not exist at all or not to an extent that a design can benefit from it. For other groups, the situation varies.

Surprisingly, given the urgency of the AIDS crisis, some possible design approaches have yet to be pilot tested very extensively. One approach to the problem of coverage might be face-to-face administration or a dual frame design in which both telephone and face-to-face methods are used.

Another method that has been rarely tested for its applicability to AIDS surveys is use of network samples. Although it has been effective in sampling other rare and elusive populations, this method may present problems of informant knowledge of target behaviors or conditions. The identification of useful counting rules may also be problematic. Nevertheless, network sampling deserves some careful testing because of its possible usefulness for a few subdomains.

Capture-recapture methods have proven effective for counting some elusive or mobile populations. This approach might be useful for sample surveys designed primarily to count a target group such as intravenous drug users. There are three large problems with this method, however. First, the capture-recapture design requires independence between the two stages; this may not easily be achieved with populations that are engaged in illegal activities. Second, the ability to determine whether or not a person is in the subdomain or not may be very difficult if that person has reason to conceal membership. Third, the usual problem of correctly matching persons found in the capture with those found in the recapture is likely to exist when the method is applied to AIDS studies.

In sum, no major innovations in sample design or modifications of traditional procedures have produced efficient designs for locating target subdomains at low cost. One must also recognize that the problems of respondent self-identification are difficult to separate from those of sample design. Even if powerful methods were developed for increasing the presence in a sample of target population households, there would still be few gains—or even confirmation of the efficacy of the sampling procedures—if respondents denied membership in the domain.

Special high-risk or infected populations. The target populations of AIDS studies are generally defined by their health condition (such as HIV-positive or AIDS- or ARC-diagnosed) or frequent engagement in illegal or proscribed behavior (such as intravenous drug use or prostitution). Such groups are so rare that the notion of locating any reasonable numbers of them through a general population frame would make sense only in very unusual situations. Most often samples are selected from special list frames.

The two general categories of frames are hospital, clinic, or other organization lists of patients, clients, or members; and frames constructed by the researcher specifically for enumerating and sampling the target group members. To date, the use of both frame types in a single study has not been common.

With any constructed list—such as a list of places that prostitutes frequent—the researcher is concerned with coverage and list accuracy. The hospital, clinic, or other organization list frame, however, brings with it other problems that influence the effectiveness of the sampling procedures. In addition to the standard frame problems, there are additonal issues of negotiating with institutions or individual gatekeepers to reach the target group. Furthermore, in most of these types of studies, the institutions are not selected by probability methods, which limits the generalizability of findings to those places included. It should be noted, however, that there is no inherent reason, beyond cost and time constraints, why careful samples of institutions could not be a first stage of selection.

The key methodological issues raised in such studies begin with population definition. The list may identify both target and nontarget members in addition to some persons whose membership may not be clear from the list itself. Presence on the list cannot be allowed to define the study population. Ideally, the researcher should be able to work with an organization's officials to obtain auxillary information necessary for classification. Such officials may also affect access to patients or clients, thus introducing unknown biases into the resulting sample. These problems are more administrative than methodological, but their outcome may affect the results.

In some studies of high-risk populations no list may exist. The field staff simply attempts to interview all members of the target population that can be identified and contacted in a particular geographic area during a specified period. Although this approach may be acceptable for limited local or pilot goals, there are at least two severe problems. First, it is difficult to know how complete the coverage really is or to assess undercoverage. Second, it is difficult to know how the presence of the research team might affect the visibility of persons engaged in illegal behavior.

Given the increasing concerns for confidentiality, difficulties of access to target population members through special organization lists will probably continue to be more of a political and administrative problem than a technical one.

Two additional considerations for AIDS sample designs may also be noted:

1. Because sample design issues for AIDS studies do not differ radically from other designs for special populations, some traditional practices may simply need reiteration. Careful discussion of the limitations of particular studies remains important, especially concerning possible sources of nonsampling errors. The use of standard estimates of bias and mean square errors, as well as sampling variances, to evaluate designs should be encouraged. The proper role of nonprobability samples

for pilot tests and other limited uses should be given systematic thought. And, although concerns about undercoverage and accrual of sufficient cases for analysis are important, they should not overshadow other considerations of sound probability designs. The potential for misuse of data from these surveys should be kept in mind.

2. In addition, improvements in sample design will not be effective unless they are accompanied by the increased willingness of special population groups to cooperate in such studies. To this end, careful thought needs to be given to how self-identification with behaviors related to AIDS risk affects cooperation by high-risk groups. Toward this end, cognitive investigations may be used to address how best to deal with these effects. Because the role of the interviewer is crucial, systematic field experimentation is also needed to explore which interview situations are best served by interviewers who are current or past members of the target group, and when is it best that the data gatherer have no identification with the survey population.

FEATURE PAPER

Life Course and Network Considerations in the Design of the Survey of Health and Sexual Behavior

Edward O. Laumann, John H. Gagnon, and Robert T. Michael

Introduction

The goal of the Survey of Health and Sexual Behavior is to describe the distribution of sexual practices in the general population and the way this distribution changes in response to changes in social context. Two aspects of sexual conduct, therefore, should be integral to the analytic framework. One is the systematic variation in sexual activity in response to lifecycle changes and changes in the social and cultural environment. The other is the systematic variation in sexual pair formation—the fact that sexual contact is not a process of random mating and that the characteristics of the pair, rather than the individual, determine the content and execution of the sexual transaction. Accordingly, the analytic framework for the development of the questionnaire has been driven by the life-course perspective and social network theory. Indeed, use of a network approach is essential if the data collected are to be linked to the epidemiologic modeling of sexually transmitted diseases (STDs), including acquired immunodeficiency syndrome (AIDS).

The Life-Course Perspective and Sociosexual Development

Patterns of sexual conduct are age- and status-graded (Reigel & Meachum, 1976a, 1976b; Clausen, 1972) and not only a function of physical growth and decline. Biological changes associated with age clearly have a fundamental role in shaping human sexual development, particularly as they relate to reproduction. Within the broad outlines of reproductive change, however, there is a great deal of variation in the timing and sequence of events in an individual's sexual life history. This variation reflects the impact of social and normative constraints on the expression of sexuality, and leads us to use the term sexual conduct, as well as sexual behavior (Gagnon & Simon, 1973).

Several stages in the life course have been identified as relevant in the study of sexuality: Prepuberty, early and late adolescence, and a sequence of stages in adulthood defined by marital status (single, married, and post-marital). In the early stages of the life course, prepuberty and early adolescence, life-course effects are highly correlated with age or cohort effects. Generally, knowing whether someone is 5 years old, 11 years old, or 16 years old is likely to provide a reasonable guide to the type of sexual experience they have had. However, although age and biological development play a large role early in the life course, even in this period significant variation remains. In addition, the kinds of events that occur during these formative years and particularly their timing relative to the prevailing social norms (for example, late puberty or early coitus), are likely to have an effect on later patterns of adult sexual life (often as a function of outlier effects).

During adolescence and particularly after high school the link between age and stage in the life course weakens, and the effects of age on sexual conduct are less consistent. An 18 year old just out of high school may go to college, join the military, or enter the labor force, and may choose to marry or stay single. These choices will affect the type of sexual conduct he or she engages in as well as the pool of potential sexual partners. After age 20, knowing whether someone is 25 or 35, or even 55, probably tells one less about her or his sexual activities than knowing whether she or he is married, single, or divorced. In short, with postadolescent sexuality, variation in the timing and sequencing of events predominates, and age grading of sexual conduct gives way to a largely age-independent process of social grading.

To capture the descriptive and dynamic aspects of postadolescent sexual variation, the principles around

Edward O. Laumann is with the Division of Social Sciences, University of Chicago. John H. Gagnon is with the Department of Psychology and Sociology, State University of New York, Stony Brook. Robert T. Michael is with the National Opinion Research Center, University of Chicago.

The research was supported in part by PHS-NICHD Contract No. N01-HD-8-2907.

which adult sexuality is organized must be identified and used to structure the questionnaire. The major determinants of long-term variation in adult sexual conduct are likely to be the presence or absence of a primary sexual partner, married, cohabiting, or otherwise, and the birth of children. Other life-course events, such as changes in employment status, major health problems, and stressful incidents, will also contribute to the patterning of sexual behavior, both in the short and long term.

The presence or absence of a primary sexual partner is probably the most important of these variables. Given contemporary patterns of mate seeking, cohabitation, marriage, and couple breakup, it is likely that more than two thirds of all adults have at least one change of primary partner over their lifetime. Thus, for many people, the stages of search, commitment, and breakup tend to be repeated a number of times over the adult life course. Sexual conduct in these periods of transition will differ from that in periods of stability, both in terms of the meaning it has for the subject and in the extent of the risks it is likely to entail. Those who cycle through the stages of exclusive pair formation bring additional experience to bear with each iteration of the sequence.

With respect to the risks associated with sexual behavior, there are probably specific windows of vulnerability in the life course, particularly during adolescence and young adulthood. The process of acquiring new forms of conduct (for example, drinking, driving, drugs, sex) in relatively untutored circumstances seems to increase rates of error-laden actions (for issues related to premarital pregnancy, see Furstenburg and associates, 1982). Other status transition periods, such as military service, college attendance, or divorce, are likely to be associated with a temporary rise in the number of sexual partners or changes in sexual practices or preferences. Even during stable periods, the business trip or vacation may temporarily suspend a routine sexual script and provide an opportunity for erotic exploration.

The life-course approach to sexual development cannot be reduced either to chronological age-grading or to a simple linear sequence of events. Instead, variations in sexual and fertility behavior are embedded in social contexts that affect the timing and sequence of events in sexual development (Hogan, 1981). Taking a life-course perspective does not mean accepting a belief in universal patterns of human development or even a modified Eriksonian view of life stages and transitions (Erikson, 1963). Rather, it is a recognition that sexuality, as well as other social activities that shape sexuality, are timed by normative and structural forces in the society. The staging of the life course thus depends on social developments rather than on the unfolding of organism (Neugarten & Datan, 1973; Nardi, 1973; Uhlenberg, 1978). In the United States, both interpersonal scripts for sexual conduct (the who, what, when, where, and why of conduct) and sexual networks change across the life course. As a result, both sequences of and opportunities for sexual interaction can be linked to stages in the life course.

The life-course approach also suggests that event history methods are an appropriate analytic tool. Event history methods have become a standard demographic technique in data collection and analysis (the reader is referred to any recent issue of the journal *Demography*). To use these techniques, it is necessary to develop a model of the relevant events that demarcate the stages in the sexual life course. Clearly, marital, cohabitation, and fertility events will form the basic framework for the model. Additional events reflecting economic change (unemployment, job change, travel) may also be relevant.

While the life course is one of the important ways of understanding or at least accounting for sexual conduct, it needs to be cross-cut by the strata in the society that delimit the networks of social interaction and provide opportunities for sexual conduct. Some of these stratification systems begin early in life and allocate individuals to different lines of development. Some of these are based on biological deficits, which, when responded to by the society, create specially segregated clusters in the society, such as the deaf, the physically handicapped, the retarded. Others rest on ascriptive characteristics such as race or religion or the socioeconomic statuses of parents. Clearly some of these statuses are relatively easy to change, whereas others are not (Featherman, 1980). Additional stratifications appear later in life. College attendance by some and going to work after high school by others often consolidate prior but more permeable boundaries between young persons of different social classes in high school. New networks form that exclude former potential sexual partners and offer openings for new ones.

The Social Structuring of Sexual Networks

Sexual conduct, more than many other kinds of human behavior, is a matching process that is characterized by the pair of actors involved and not by the single individual. From this point of view, the unit of analysis is the sexual transaction and not the respondent. Obtaining information only on the respondent's sexual behavior and characteristics, and not on the characteristics of their partners, provides an incomplete picture of this transaction. Such one-sided information would reduce the products of a survey such as this to counting the number of people who engage in certain activities. Two things would be lost as a result. The first would be the ability to explore the social and psychological bases of sexual conduct, a loss that would critically reduce the value of these data for public health intervention. The second would be the ability to use these data to model the spread of sexually transmitted diseases (including AIDS). Epidemiologic models of disease spread are based on contact rates—-the number of transactions that occur between members of different groups (May & Anderson, 1987). Thus the characteristics of both parties in a sexual transaction are needed for these models.

The sexual transaction and the parties to this transaction are at the analytic center of the National Survey of Health and Sexual Behavior. The strategy is tied to conventional sampling procedures, but theoretically

committed to identifying the groups and contexts within which sexual conduct, contraceptive behavior, and STD or human immunodeficiency virus (HIV) transmission occur. This strategy is reflected in the use of a systematic accounting frame for describing the respondent's network of sexual contacts. Each respondent chosen for this survey will have such a network. The network may consist of a single partner (for example, a spouse), or of both a primary partner and one or more secondary partners. The network could also be empty, that is, the respondent might have no current sexual partners. Detailed information will be collected on the respondent's sexual partners in each of three time frames (last event, last year, last 5 years). The network strategy as it is used here will enable us to investigate the social bases of sexual conduct and the distribution of risk.

Sexual conduct does not take the form of random mating but of highly selective pairings. The sexual opportunities available to the respondent and the type of partners deemed appropriate will vary systematically from one social group to the next. The social composition of sexual networks will therefore vary systematically as well. Heterosexuals, for example, will target a different group of contacts than will homosexuals, adolescents a different group than those in their midforties, married persons a different group than the divorced, and so on. These differences in the social composition of sexual networks have important consequences both at the societal and the individual level.

At the societal level, the nonrandom element in the choice of sexual partners has direct implications for the epidemiologic modeling of disease transmission. Large groups in society tend to have limited contact with one another, associating instead with others of like type, be it defined by race, age, religion, class, or sexual preference. We know that this departure of social structure from complete mixture will slow the diffusion process, and that it may potentially prevent completely the spread of a disease to certain groups (Coleman, 1986). Aggregating information in the sexual network to estimate the contact rates between and within groups (Laumann, 1966, 1973; Laumann & Pappi, 1976; Fararo & Skvoretz, 1987), we can simulate the trajectories of diffusion across groups. Variation in the extent of contact, the number of groups, relative group size, the proportion infected, and the existence of indirect paths would affect the shape and spread of these trajectories. At this level, the minimal set of partner characteristics needed for epidemiologic projections include race, age, gender, sexual preference, and marital status.

This network accounting strategy has the potential to provide a very detailed mapping of the social groups who are at risk of sexually transmitted disease (including HIV), whether due to their direct transactions with other high risk groups or to their indirect vulnerability through a third, bridge party. In addition, using multivariate analysis techniques (Schiffman & associates, 1981), it will be possible to cluster according to social and demographic characteristics, which would help to identify along which social axes (for example, race, socioeconomic status, age) public health interventions are most needed.

At the individual level, the social context of a sexual transaction is strongly related to the presence or absence of risky behavior, whether it be risk of disease, pregnancy, or victimization. The activities engaged in by a couple who share a long-term commitment in the context of domestic routines are likely to differ from a nonrepeat incident between strangers in a hotel. Thus, data on a respondent's inventory of sexual behaviors, which are needed to establish the level of objective risk, must be collected in a partner-specific format.

Current epidemiologic evidence suggests that the efficiency of disease transmission varies by sexual technique. It is thus necessary to collect partner-specific information on at least three techniques: the frequency of anal intercourse (distinguishing between receptive and insertive), oral sex, and vaginal intercourse for each partner in each time frame. Partner-specific data on condom use in the last year also needs to be collected, with disease and contraceptive intentions clearly distinguished. Because there may also be risks associated with a wider range of sexual behaviors, and with concomitant nonsexual behaviors such as douching or the recreational use of drugs, information will be obtained on the full range of sexual practices employed by the respondents in their last events within the last year for up to three partners.

These individual-level effects also feed back to the larger societal picture. Simple analyses of the incidence of these activities will provide a picture of the distribution of risk across the population. Analysis of volume (or frequency) of the activity will give some perspective on the proportion of all sexual behavior that is a risk, variations in group exposure rates, and the contributions of specific groups to the total pool of risky activity. With respect to disease, therefore, the incidence, regularity, and prevalence of sexual behaviors provide the basic physical parameters of transmission, whereas the contexts and social bonds give the social parameters of diffusion. Network analysis provides a conceptual framework in which these physical and social components find a natural synthesis.

Network analysis has undergone a rapid process of maturation and sophistication over the past 20 years (Berkowitz 1982; Marsden & Laumann, 1984; Burt, 1982, for general introductions). It has developed an impressive array of analytic techniques, clarified technical issues, and accumulated a rich tradition of empirical research on diverse substantive themes. In each of these areas, network analysis has demonstrated its flexibility and capacity to generate significant insight into social phenomena at both the microlevels and macrolevels of analysis (Coleman, 1986). With respect to the research discussed here, the application of a network approach will provide both a superior theoretical framework and a natural bridge between the survey data and epidemiologic models for the projection of disease transmission and intervention efficacy.

What the network strategy assumes is that the measurement error introduced by the respondents' description of the characteristics of their sexual partners will be relatively small. This is not an assumption made lightly or without verification. There is some evidence that re-

spondent reporting is fairly accurate, both in regard to sexual activity and in regard to social characteristics. In a recent study by Coates and others (1988), the primary partners of seropositive gay men were located, and both the index case and the partner were administered a questionnaire on the type and frequency of sexual behaviors that they had engaged in as a pair. There was very high agreement between the two reports: an average correlation of about 0.72, and only one correlation below 0.50. In earlier studies of the accuracy of reported social characteristics of partners in friendship networks, Laumann (1973:27-39) found fairly high correlations between the respondent's descriptions and the partner's actual characteristics. Not surprisingly, the degree of accuracy is highest for simple demographic characteristics, such as age (about 86 percent agreement within 2 years; sex and race were not included in this study), and lower for nonascriptive characteristics, such as political party preference (about 53 percent). Factors that might be expected to be important in friendship formation were reported very accurately as well; for example, religion (about 85 percent when Protestant denomination is left unspecified). There were no systematic biases observed in the errors that were made, with the exception of political party preference, which the respondents were more likely to report as similar to their own. The results of this analysis are encouraging but not conclusive, because it is not known whether the patterns observed for friendship networks will also hold for sexual networks.

The approach described above implicitly treats sexual development as a dynamic sequence of events (Laumann & Knoke, 1987:21-35; Ensinger, 1987). It assumes that a single sexual event is embedded in the context of antecedent, concurrent, and impending events. To understand sexual conduct, it takes into account how actors perceive and respond to an opportunity structure that is created by a temporal sequence of events. At the sociocultural level such events include public health campaigns concerning sex-related health risks, the availability of various contraceptives and treatment for sexually transmitted diseases, and shifts in patterns of family formation. At the level of the individual there are lifecycle transitions involving marital status, employment status, and geographic mobility. Events at both these levels may lead to changes in the patterns of sexual behavior, from restrictive to permissive patterns or vice versa. These changes in behavior in turn affect contraceptive use and the potential spread of sexually transmitted diseases.

Life Course, Networks, and Questionnaire Design

Before discussing some of the major considerations of questionnaire design for a national study of sexual behavior, it seems necessary to put the history of the questionnaire design into perspective. Originally 1 year was to be spent developing a set of design alternatives for questionnaire and sampling strategies for a national study of health and sexual behavior. The central feature of the questionnaire design effort was commitment to a modular approach within a specific set of sampling alternatives (Table 1). A few months into the process, however, the course of the research changed, and development of materials for a large-scale pretest of a particular instrument as well as methods for the collection of data was begun. Therefore, instead of developing drafts of all modules, both core and supplemental, of the proposed questionnaire, a single pretest instrument was designed that incorporated material from both core and supplemental modules. The goal was to develop an instrument that would include as many as was feasible of the major design features of the questionnaire to be employed in the full-scale study. While one of the reasons for embarking on a pretest was to provide preliminary data for policy purposes, it was also intended to provide as much information as possible on various options that were being considered for the full-scale study. Rather than try to create a version of an instrument that could be used in the major study, the pretest instrument was designed to be rich in the elements—questions and approaches—that needed testing. Questions that have been used before and are relatively well understood were purposely not included.

Table 1. Modules being considered by the NHSB survey

Core modules proposed	Modules represented in pretest questionnaire
Demographic information	some
Sexual practices and partners	most
Contraceptive practices	little
STDs and medical history	most*
Health knowledge, attitudes, and practices	little
Supplemental modules	
Retrospective sex history	some
Sex attitudes and beliefs	little
Social and political attitudes	none
Economic cofactors	none
Substance use	most
HIV and STD testing	none

* The most in this case refers primarily to information about STD knowledge, attitudes, and experience; far less information was included about the medical or psychological history of the respondent. Two additional modules, one on gender attitudes, beliefs, and practices and a second on the interpretative frames that individuals use to make sense of their sexual experiences, are also under development but not included in the pretest.

This redirection of effort has had two major implications. First, some elements originally envisaged do not yet exist, including parts of the questionnaire that are to be used as supplemental modules. Second, since several alternative design features are to be assessed in the pretest, it is not possible to report on particular options on these points until the pretest data have been collected and analyzed.

Several major interrelated issues exist in the general design of the National Health and Sexual Behavior (NHSB) questionnaire: (1) the importance of narrative flow in the questionnaire; (2) the division of the substantive material to be covered into core and supplemental modules; (3) the temporal framework for asking questions about partners and specific sexual, contracep-

tive, and STD-related practices; and (4) the method to be employed for asking sensitive questions.

Narrative Flow

One of the most important considerations in developing a questionnaire, especially one of the complexity and sensitivity of the NHSB survey, is that it have meaning and coherence for the person being interviewed. A survey interview is a specialized type of conversation, one in which the respondent is asked to reveal a great deal of information about himself or herself. Most people like to talk about themselves, although they rarely, if ever, get a chance to talk honestly about their sexual experiences. The chance to do so in a supportive, nonjudgmental environment can be a positive and satisfying emotional experience. To ensure that this is the case, special care must be taken in all aspects of the interviewing process, but especially in writing and structuring the questionnaire itself.

First and probably most important, the construction of the questionnaire must ensure a logical narrative flow to the interview. The respondent must be allowed to tell his or her story in a reasonable and natural way, in spite of the unusual, nonquotidian fact of being interviewed. To this end, the purpose of the interview must be clearly communicated at the outset. This is best done by being honest about the motivation of the study, about its scientific value, and its relation to the national health emergency precipitated by the AIDS epidemic, in short, about the crucial nature of the data being collected. After this introduction, the questioning must proceed in a reasonable and logical manner. The inquiry should proceed from less sensitive to more sensitive topics; this allows development of rapport between interviewer and respondent before the more difficult subjects are broached. And topics need to build on each other.

Both of these features are illustrated in the pretest questionnaire, which begins with (1) demographic information and moves to (2) major relationships (cohabitation and marriages), and then to (3) fertility. Only after this background is established does (4) enumeration of sexual partners in the last year begin. From partner enumeration the questionnaire moves to (5) the most recent sexual encounter(s) within the last year for up to three different partners. Sexual practices as well as actions to prevent pregnancy or disease or both are discussed in the context of the specific events. Only then does the instrument seek summary information on partner data from (6) the last year and on (7) the last 5 years. The importance of using major life-course events, and in particular major sexual relationships, marriages, and cohabitations, to organize the discussion of other sexual partners and practices has already been discussed. The questionnaire then concludes with questions on (8) early childhood and adolescence (including experiences with sexual violence), (9) general sexual attitudes and knowledge, (10) health, drug use, and STD-related attitudes and history (including AIDS knowledge, attitudes and experience), and sensitive information on (11) family and personal income. Any questionnaire about sexual behavior must maintain a similar type of coherence and structure. This has been crucial to the design of the pretest instrument and needs to be carried over to the construction of multiple versions of the questionnaire if a longitudinal version of a national survey were undertaken.

Questionnaire Modules

One fundamental design decision in the development of the survey instruments for the NHSB survey was to divide the substantive components into modules. This was necessitated by the remarkable breadth and depth of the material to be covered. Much survey experience suggests that the length of a research interview should ideally average no more than 1.5 hours because of respondent and interviewer fatigue and the likelihood of competing time commitments and distractions. The amount of information and the level of detail that is of interest in this study could easily require 4 or 5 hours of interview time. For example, an original draft of the questionnaire for the pretest, which was intended to cover only the sex core and materials most in need of pretesting from the supplemental modules, averaged about 2.5 hours. Since the proposed sampling design calls for a very large basic sample broken up into replicates with a variable number (zero, one, or two) of reinterviews, fundamental information can be secured from each respondent (some only one time, some several times), while information on certain topics will be sufficient if obtained from only one large cross-sectional replicate. Therefore, it makes sense in some cases to break the questions to be asked into separate sections of the interview schedule (an actual module); other questions on the same topics will be integrated at appropriate points in the main body of the interview schedule. For example, a supplemental module on substance abuse that will contain items on the life history of alcohol and drug use may well stand alone, whereas other items will be integrated into each of the sections of the instrument that refer to appropriate time and partner frames. Thus, some alcohol and drug-use questions will be asked in relation to particular sexual events or as activities with particular partners in the last year.

The integration of questions from the various modules into a single integrated whole along these lines is especially important in maintaining coherent narrative structure. This kind of integration enhances recall and improves the accuracy of respondents' reporting; at the same time, it reduces the possibility of irritating respondents and thus reduces the risk that they will refuse to answer or even break off the interview. This narrative structure is in part developed through the employment of discrete time frames, discussed below.

At this stage the final pretest questionnaire has fully taken shape. Achieving the final form of the instrument for the NHSB survey will require additional time for development and the results of the pretest. Table 1 presents the modules being considered and indicates those that are represented (as a whole or in part) in the pretest questionnaire.

The pretest questionnaire contains items from both the core and supplemental modules that were originally proposed. It draws material from the first four of the five core modules. The last core module, on health knowledge, attitudes, and practices, is still in the early stages of development, and the items on health included in the pretest questionnaire are mainly related to sexually transmitted diseases and AIDS. The pretest questionnaire draws on two of the supplementary modules. The 5-year section and a section on early childhood and adolescent experiences cover some of the areas that would be included in a retrospective sex history module.

Some more alcohol and drug use questioning would probably be needed for a complete module on substance use, although this module may be nearly complete. The modules on sex and political attitudes and on economic cofactors have been roughly sketched out and some items have been written. Only a few of these items have been included in the pretest questionnaire. The final supplemental module, on HIV and STD testing, was originally conceived as possibly including actual testing (that is, collecting blood samples), but with the fielding of a national household seroprevalence survey this has not been pursued. Instead, this module should contain questions about whether respondents have been tested for STDs and HIV and what effect, if any, this has had on their lives. Some questions along these lines were developed but later dropped from the pretest, mainly owing to time considerations.

Timeframes

One of the major problems in constructing an interview focusing primarily on sexual behavior is to create a framework that will enhance respondents' ability to report on their sexual experiences. It is important to take into account that a large proportion of sexual behavior takes place within relatively long-term relationships and that such behavior usually becomes routine. Highly routinized behavior is generally less memorable than the unique and unusual. For the relatively small number of respondents who have a large number of partners, reporting may be problematic because of the amount and heterogeneity of information to be covered. We therefore chose to develop a scheme for collecting data on sexual behavior that would allow capture of the greatest part of the variation in both partners and sexual events in such a way as to minimize the memory burden on respondents. This translated into decisions about a set of timeframes for asking questions about partners and events.

The questionnaire as developed for the pretest is constructed around three major timeframes: 2 to 4 weeks, 1 year, and 5 years. Each timeframe has a specific purpose; each is particularly well suited to answer certain kinds of questions. Although the experience of the pretest is not yet available to fine tune and assess this approach, it is increasingly clear that the basic idea of collecting data for a variety of time periods is essential. Thus, although the pretest may suggest that the lengths of the timeframes selected should be modified (for example, the goals of the 2-week period might better be accomplished using 1 week or a full month), there is little question of the analytic and theoretical advantages of getting information on multiple time windows.

Two- to Four-Week Period. The shortest timeframe about which questions are to be asked is the past month. The major purpose of this timeframe is to gather information on the amount of sexual activity for all respondents within a period that is recent enough to be easily remembered. This section includes items about sexual activity with partners as well as masturbation. The actual reporting window is a 2-week period, but that window is determined by the last time a respondent had sex with a partner in the last month, if at all. The dates of all sexual encounters in the 2 weeks that include the last event with a partner are then to be sought.

The Last Year. The primary reporting period for partners and sexual events is the last year. Data from the 1988 General Social Survey indicate that 80 percent of the adult population had only one partner or none in the last year. Most people, then, should have no trouble enumerating their partners for this period. It is true that 20 percent reported more than 1 partner in the last year, but very few people reported many partners (0.7% had more than 10 partners). Thus, the last year seems to be the most reasonable time period in which to discuss particular partners and particular sexual encounters.

In the pretest version of the questionnaire, the sexual partners and behavior section begins with the enumeration of the sexual partners of the last year. Respondents will be asked for the first name, initials, or a pseudonym for each partner. The subsequent questions about the partners, both their social characteristics and the nature of their social and sexual relationships with the respondent, are tied to specific individuals. This is central to the network approach motivating this study. Rather than aggregate counts of the number of partners who fit into various categories (for example, sex, race, age, social status) data will be gathered at the individual level. This will make it easier for the respondent to answer the questions and it will increase immeasurably the richness of the data available for analysis.

Full information on the last sexual event is to be gathered for up to three partners in the last year. This includes quite detailed demographic and social status data, as well as sexual, contraceptive, and STD prevention practices for each dyad. The event-specific information will provide still more richness to the data being collected. It will be possible to investigate variations in sexual, contraceptive, and STD preventive practices, not only within the dyad but also within the specific event. Again, the specificity of the information sought should improve accuracy of recall and reporting as well as allowing for another distinct level of inquiry. For example, such factors as drug and alcohol use in sexual encounters may effect what sexual acts are engaged in, whether condoms are used, and so forth. The cost of collecting this level of detail is valuable interview time, and this has limited the number of partners about which such information can be gathered. In addition to these event-

specific data, additional though more limited demographic and sexual data will be collected for up to another 12 partners from the last year. Recall in these cases will be improved by linking sexual behaviors to specific partners in a specific time period. Summary information on contraceptive and STD prevention practices are also collected by partner for the year-long period.

Five-Year Period. Sexual and related practices are likely to vary over time, and much of our interest is in the variation due to partners. Thus, this section is critical. It is also the place in the interview where memory burden is potentially the greatest. The original version of the pretest questionnaire asked questions covering a 10-year period, but when the length of that version had to be cut by a third, the 10-year section became a 5-year section. This section of the interview is also one of the most complicated for the interviewer to administer because it is structured around major sexual relationships (defined as marriages or sexual relationships involving cohabitation); a type of organization that makes it much easier for the respondent to answer. Using the major relationships as markers, the respondent will be able to report on the more casual sexual relationships that coincide with or intervene between his or her major relationships. (Of course, for many respondents there will be no other or only a few relationships to recall.)

The Operational Definition of Sex

Among the additional innovations in the pretest version of the questionnaire the definition of sex that is used is paramount. The most common meanings associated with the phrases "having sex" or "sex partner" are orgasm and vaginal intercourse. Not wanting to rely merely on the respondents' own variable definitions, it had to be specified what is meant by "sex." Instead of either intercourse, intromission, or orgasm, a broader definition of sexual activity was chosen; "any mutually voluntary activity with another person that involved physical contact and sexual excitement or arousal (that is, feeling really turned on) even if intercourse did not occur."

The purpose of this definition is to cast a wide net at the outset, to include a larger number of sexual partners and acts. This definition should not be too broad, but if it is found to be so, leading to the inclusion of too many people and events where the sexual activity is merely arousal, it can be narrowed. The advantage of using it is increasing the likelihood that the survey will capture data on acts and partners that might be excluded by the narrower definitions. For example, much homosexual activity, male and female, does not involve intromission. If orgasm had been chosen as the criterion for sex, genital activity might well have been excluded including intercourse for people or even partners who do not have orgasms as a component of a specific or all forms of sexual activity. Since the instrument will collect information on the actual sexual activity involved, analysts will be able to differentiate the various levels of sexual involvement.

Eliciting Sensitive Information

In many surveys that involve asking very personal and sensitive questions, alternatives to direct questioning, such as self-administered forms or randomized responses, are used. These were considered along with alternatives such as using a "walkman" tape recorder with a recording of the question to be answered. But these alternatives make it impossible to collect network and event data of any complexity. Self-administered forms in particular necessitate a very simple line of questioning with practically no skip patterns. In the course of developing the instrument, however, it was found that the specific sexual behavior questions, which are probably the most sensitive of all, might be too complicated to ask verbally because the list of responses was long and the alternatives numerous.

In response to these conflicting requirements, cards listing people responses were developed and shown or handed to the respondents. These hand cards have numbers beside the verbal descriptions, and it is the numbers that the respondent says to the interviewer. The interviewer, who does not have the text of the answers on the interview schedule itself, just records the number, and this is made clear to the respondent. This procedure avoids making respondents use sexually explicit language that might be embarrassing or unnatural for them (for example, terms such as oral or anal sex), and it has proven to be more comfortable for respondents in pilot interviews. There is a noticeable release of tension on the part of respondents when they see how this works, and in a number of cases they have explicitly stated in debriefing after the interview that they found this technique made the interview easier.

The utility of this device is limited for respondents who have reading problems. However, ways to get around this are available. Respondents can be shown and read a list of definitions that include phrases from the hand cards. Respondents who merely have difficulty reading but who are not completely illiterate should be helped by this. Furthermore, hand cards are frequently used in face-to-face survey interviews, and interviewers are instructed in how to help respondents who exhibit any difficulty in reading the cards.

Conclusion

The goal of this paper is to indicate the ways in which the theoretical and analytic perspectives of the life course and social networks are used to shape the design of a survey instrument for gathering data on health and sexual behavior. Both the substantive areas of interest, health and sexuality, need to be understood as organized by the structure of social relationships rather than as a result of the characteristics of a detached individual. As a result of these theoretical concerns, a questionnaire has been designed that is narratively coherent, focused on particular life events and transitions, and that centers on information about socially structured sexual partnerships. In this design an effort has been made to address

the problems of appropriate time frames for recall, the sensitivity of the information that we wish to gather, and the full variety of sexual conduct that is evidenced in the general population.

References

Berkowitz, S. (1982). *An introduction to structural analysis: The network approach to social research.* Toronto: Butterworth.

Burt, R. (1982). *Towards a structural theory of action: Networks of social structure, perception, and action.* New York: Academic Press.

Clausen, J. (1972). Life course of individuals. In M. W. Riley, M. Johnson, & A. Fones (Eds.). *Aging and society,* (vol. 3). New York: Russell Sage.

Coates, T. A., Calzavara, L. M., Soskolne, C. L., & associate. (1988). Validity of sexual histories in a prospective study of male sexual contacts of men with AIDS or an AIDS-related condition. *American Journal of Epidemiology,* 128, 719-728.

Coleman, J. S. (1986). Micro-foundations and macro-social theory. In S. Lindberg, J. S. Coleman, & S. Novak. (Eds.). *Approaches to social theory.* New York: Russell Sage Foundation.

Ensinger, M. (1987). Adolescent sexual behavior as it relates to other transition behaviors in youth. In S. Hofferth & S. Hayes (Eds.). *Risking the future* (vol. 2). Washington, D.C.: National Academy of Sciences Press.

Erikson, E. H. (1963). *Childhood and society* (2nd ed.). New York: W. W. Norton.

Fararo, T., and Skvoretz, J. (1986). E-state structuralism. *American Sociological Review,* 51, 591-602.

Featherman, D. (1980). Schooling and occupational careers: Constancy and change in worldly success. In O. G. Brim & J. Kagan (Eds.). *Constancy and change in human development.* Cambridge, MA: Harvard University Press.

Furstenburg, F. F., Jr., Lincoln, M., & Menken, J. (Eds.). (1982). *Teenage sexuality, pregnancy and childbearing.* Philadelphia: University of Pennsylvania Press.

Gagnon, J. H., & Simon, W. (1973). *Sexual conduct: The social sources of human sexuality.* Chicago: Aldine.

Hogan, D. (1981). *Transition and social change: The early lives of American men.* New York: Academic Press.

Laumann, E. O. (1966). *Prestige and association in an urban community.* New York: Bobbs Merrill.

Laumann, E. O. (1973). *Bonds of pluralism: Form and substance of urban social networks.* New York: Wiley & Sons.

Laumann, E. O., & Knoke, D. (1987). *The organization state: Social choice in national policy domains.* Madison: University of Wisconsin Press.

Laumann, E. O., & Pappi, F. U. (1976). *Networks of collective action.* New York: Academic Press.

Marsden P., & Laumann, E. O. (1984). Mathematical ideas in social structural analysis. *Journal of Mathematical Sociology,* 10, 271-294.

May, R. M., & Anderson, R. M. (1987). Transmission dynamics of HIV infection. *Nature,* 326, 137-142.

Nardi, A. (1973). Person perception research and perception of the life course. In P. Baltes & K. Shaie (Eds.). *Life span developmental psychology: Personality and socialization.* New York: Academic Press.

Neugarten, B. L., & Datan, N. (1973). Sociological perspectives on the life course. In P. Baltes & K. Shaie (Eds.). *Life span developmental psychology: Personality and socialization.* New York: Academic Press.

Reigel, E. F., & Meachum, J. A. (1976a). *The developing individual in a changing world. Historical and cultural issues* (vol. 1). Chicago: Aldine.

Reigel, E. F., & Meachum, J. A. (1976b). *The developing individual in a changing world. Social and environmental issues* (vol. 2). Chicago: Aldine.

Schiffman, S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods and applications.* New York: Academic Press.

Uhlenburg, P. (1978). Changing configurations of the life course. In T. Haraven (Ed.). *Transitions.* New York: Academic Press.

FEATURE PAPER

Establishing the Comfort Zone: Developing Interviewer Competence and Confidence in a Survey on a Sensitive Topic

Barbara Campbell, Pat Phillips, Rebecca Zahavi, Ellen Williams, and Sally Murphy

The Issues

Surveys on sensitive topics, drug use, bankruptcy, homelessness, and the like, pose a variety of challenges to the survey researcher. Debate in the literature has focused on the threat such surveys pose to respondents. Researchers have assessed this threat, examined the ways in which it can be exacerbated or reduced through such means as question wording and question context, and investigated its effects on survey responses and nonresponse (Sudman & Bradburn, 1982) and the repercussions of respondent threat for the reliability and validity of data.

Another perspective that warrants scrutiny is the threat such surveys pose for interviewers. Interviewer threat can include discomfort with questionnaire content, lack of confidence that the data can be collected, and actual fear of physical harm from the respondents or others in the respondents' environment.

If interviewers succumb to such perceived threats they may introduce bias into the data, a bias that can take several forms. If interviewers lack confidence they can invite refusals; if interviewers are fearful, they can fail to followup nonresponse vigorously. If interviewers are uncomfortable with the questions, they can give signals that will reduce the respondent's willingness to give full answers. Clearly, the risk to data quality from interviewer threat is related to and may contribute to respondent threat.

Barbara Campbell, Pat Phillips, Rebecca Zahavi, Ellen Williams, and Sally Murphy are with the National Opinion Research Center, University of Chicago.

The authors acknowledge the work of their colleague Woody Carter who contributed materially to the direction of the comfort-zone training materials.

The training approach and materials described here were developed for the National Study of Health and Sexual Behavior under Contract No. N01-HD-8-2907 with the National Institute of Child Health and Human Development. Preparation of this paper was supported by the National Institutes of Health and NORC.

Survey managers address these threats to interviewers in three major phases of the field management process: recruiting, training, and supervision. In these phases survey managers routinely (1) ensure the selection of the right staff for a given project, (2) instill them with the skills and assurance needed to perform their work, and (3) monitor performance in practice. However, all of these are especially critical when interviewers and respondents are dealing with sensitive subject matter. This discussion focuses primarily on the particular challenges faced by the National Study of Health and Sexual Behavior (NHSB) in one of these phases—interviewer training— and describes the program that has been developed to instill in the interviewing staff the competence and confidence necessary to conduct a successful survey on the subject of sexual behavior.

The National Study of Health and Sexual Behavior has confronted an unusually difficult set of training issues. These issues include helping the interviewers achieve ease with the subject matter of the survey in general, with the vocabulary associated with the subject matter, and with the specifics of behavior as delineated in the survey instrument, all of which are generally considered to be private and usually not the topics of social discourse. In addition to putting the interviewer at ease, training also needs to focus on how the interviewer can put the respondent at ease with the general content and with the specifics of the interview schedule.

Added to these issues of the NHSB content is an extremely complex questionnaire, which includes rostering of different sexual partners in relation to different events and different timeframes and necessitates recording in several associated instruments. Thus, the administrative burden for interviewers is also higher than normal.

This paper deals with the perceived threat to interviewers and the effects of such perceptions on interviewer behavior and data quality. The subject of real threat, such as a possibility of physical harm, though not covered in this paper, is dealt with in the training ma-

terials for the study. Contacting on the respondent's territory always contains an element of risk. Moreover, the subject matter of the NHSB survey instruments could prove provocative to some respondents. Therefore, trainers must make interviewers aware of real risk and provide them with strategies for controlling the interview situation, without frightening them away from the study or setting them up to expect a negative experience. In fact, skill and comfort in dealing with perceived threat may reduce the possibility that a situation could escalate into one of real risk.

The Personnel and the Process

The recruiting or staffing plan for the National Survey of Health and Sexual Behavior calls for the use of social science interviewers to administer the interview schedule to respondents. Some sex researchers feel that only clinicians have the appropriate training to do such interviewing. It is a tenet of the survey research profession that lay interviewers can be trained to do almost anything within the realm of social science data collecting, and in fact their lack of in-depth knowledge of the topics under study is an advantage. Subject matter experts, including graduate students, may be too invested in their subjects to be objective data collectors (Bradburn & associates, 1979). Witness the success of training survey interviewers to handle the Diagnostic Interview Schedule (DIS) and to administer psychological and aptitude tests to children in respondents' homes on the National Longitudinal Survey of Labor Market Experience, Youth Cohort (NLS/Y). In fact, few clinicians have had adequate training in how to ask questions about sexual behavior, and many clinicians could benefit from the training program that has been devised for the NHSB interviewers.

Selecting trainers for the interviewers is as important as selecting the interviewing staff. Again, some would assert that clinicians with strong backgrounds in sexual counseling might make the best trainers of interviewers for a survey on sexual behavior. The National Opinion Research Center (NORC) project staff chose instead to follow the logic of its interviewer staffing decision, that is, to use staff expert in administering survey instruments. The training plan therefore calls for NORC's most senior and experienced field staff, its field managers, to serve as leaders of the training sessions. Field managers are the direct supervisors of the interviewing staff. Experienced field managers can convey to the pragmatically oriented interviewers a confidence that the survey will work in the real world, in a way that academicians cannot match. In addition, they are far closer to the instincts and concerns of interviewers and the likely reactions of respondents than are either academicians or the central office project staff. Their expertise in survey research will be supplemented in the training sessions by consultants in substantive areas germane to the interview, including sex research and therapy, sociology, and epidemiology.

With these fundamental staffing decisions made, senior field managers who were designated to manage the data collection during the field period were brought into the process at a very early stage as task leaders for formulating the training plan and developing the materials. The field managers began their involvement by taking part in the pilot testing of the instrument. Informal testing of the instrument on medical personnel and members of general and special populations (for example, the handicapped, gay men) had been started by project staff. Having the field managers join this effort allowed them to interact with the research team members and survey staff who were developing the instrument, and to provide much practical advice and to influence directly the refinement of the survey instruments. During this pilot testing phase the training team members also read the initial proposal and had direct contact with the principal investigators in debriefing situations, thus gaining a deeper appreciation of the theoretical models that underlie the study.

The Training Plan

After the training team members became adept in the use of the questionnaire and helped to develop some of the related materials, the team tackled the issue of the training plan. The National Opinion Research Center's general approach to training interviewers rests on theories developed in the field of adult education (Winfield, 1986; Hartley, 1985; Anderson, 1969; Hsia, 1971; Merrill, 1971; Mager & Clark, 1963; Mager, 1972). These theories have, for the most part, gained industry-wide acceptance and stress elements such as involving the learner actively in the instructional process. After considering the complexity of the instruments together with the sensitivity of the topic, the team decided that a combination of home study materials and 4 days of in-person training sessions would be necessary to equip the interviewers for the NHSB fieldwork.

Assuming that the trainees as a group would be unaccustomed to explicit discussion of sexual behavior with strangers, the team decided that the training would need to be structured to bring about an adjustment of interviewer attitudes, reducing naivete where it existed and conditioning the interviewers to feel at ease with sexually explicit materials. Since such a complex thawing and stretching of attitudes does not happen overnight, the team decided that the special training dealing with the sensitivity of the topic should be interwoven throughout the training components, rather than tackled in a single, short session. Thus, sections of the training materials designed to address this issue were slated for the home study materials, which interviewers would receive several days in advance of their training sessions; for several sessions at intervals during the 4-day, in-person briefing sessions; and for the Interviewer Manual, which is to be used as a reference throughout the field period.

Initially the discussion about what to call the special training components revolved around terms like "sensitivity training" and "desensitivity training." But it was soon decided that the vagueness of the term sensitivity training could by itself induce anxieties, while desensitivity training had a manipulative cast suggestive of de-

programming from cult membership. The team believed it would be preferable to find something that would accentuate the positive, and came up with "the comfort zone." The idea behind this term is that everyone has his or her own comfort zone around the words and concepts of sexuality, some quite narrow, some quite broad. The purpose of the comfort zone training is to take each individual, first the supervisors and then the interviewers, no matter the size of his or her comfort zone, and enlarge that zone to encompass many more concepts and words. This approach is a very positive one, with the individual in control, purposely broadening his or her base. The name models positive behavior and stresses the flexibility of boundaries.

Contents of the Comfort Zone Materials

The contents of the home study materials, as developed by the team members, include the following:

- Reviewing the introduction in the Interviewer's Manual to learn the purposes and importance of the study.
- Reading excerpts from Chapter 2 of *Sexual Behavior in the Human Male* (Kinsey, 1948), which presents timeless insights into the issues of establishing rapport with the respondent, maintaining confidentiality, and the utility for the interviewer of expanding his or her own background knowledge. It demonstrates for the interviewer that such interviewing can be done and in fact was done some 40 years ago.
- Completing a two-page broadmindedness self-assessment, adapted from Yaffé and Fenwick (1988), to aid the interviewer in charting the dimensions of his or her own comfort zone and the directions in which it might need to be enlarged.
- Reading the glossary from the book *The Complete Guide to Safe Sex* (McIlvenna, 1987) which deals with sexual terminology, both clinical and slang, to prepare the interviewer to hear and understand any language the respondent may use in response to the questions posed in the interview schedule. Although the questionnaire itself includes only clinical terminology that is defined for the respondents, it is important that the interviewer's comfort zone be stretched beyond the boundaries of the questionnaire so she or he is fully prepared for other eventualities in the course of data collection activities.
- Answering a seven-page multiple choice exercise, developed for the project and reviewed by the principal investigators, to allow the interviewer to test his or her knowledge of AIDS and sexually transmitted diseases (STDs) and to test familiarity with sexual terminology.
- Reviewing the questionnaire to gain a basic familiarity with the survey instruments.

The related sections of the in-person interviewer training sessions include the following:

- A presentation by a senior member of the research team, such as one of the principal investigators or a representative from one of the funding agencies, to describe the background and importance of the study and the rationale for the various sections in the questionnaire. Such a speaker verifies the importance of the study to the interviewers and serves to engender enthusiasm for the research goals of the project.
- "Breaking the Language Barrier," an interactive session designed to give interviewers practice in saying aloud words regarding sexuality that they would not normally say in conversations or interviews, some of which they will need to read to the respondent as part of the interview schedule. This is also intended to stretch the interviewers' comfort with explicit language to the point where the clinical language of the questionnaire seems tame by comparison.
- "The Environment of the Comfort Zone," an interactive session focusing on how the interviewer could expand or narrow the comfort zone of the respondent through his or her use of appropriate or inappropriate dress, language, attitudes, and behavior. This session includes role playing by interviewers in which they are to propose how they would handle a variety of hypothetical scenarios.

An important feature of these last two sessions is the use of humor. Some people have a need to laugh, giggle, or make double entendre remarks when dealing with this subject matter. Laughter also reduces tension, so such outlets have been provided as part of the training plan.

- "What If" questions and answers, a session designed to alleviate interviewers' concerns by allowing them to ask questions that have occurred to them during the home study or during the previous sessions. Interviewers need not save all questions for this session; it will be left to the judgment of the individual trainers which questions to handle immediately during the sessions and which to save for the "What If" session.

These sessions will be interspersed with other training activities. Chief among these will be sessions devoted to mastery of questionnaire administration. This activity, although not specifically designed to broaden the interviewers' comfort zone, is expected to do so in two ways. First, practice alone will enhance the interviewer's feelings of ease with the materials. Practice interviews involving role playing will prepare interviewers to interact with respondents whose lives illustrate a variety of sexual patterns. Second, mastery of the very complex document will enhance confidence. (In fact, the complexity of the document demands such concentration that interviewers engaged in administering it correctly will have little time for the psychological involvement necessary to become embarrassed. This was the experience of the field managers who did pilot interviews.)

Expectations for Success

The training is expected to be successful for several reasons already discussed; chiefly, the quality of the training materials and the skills of the session leaders. Added to this is the expected level of interviewer motivation. In preliminary recruiting efforts, field managers have found that many interviewers are highly motivated to work on this study even before they have had the energizing experience of personal training. First, they believe in the cause, that is, the social utility of the data

for the funding agencies to combat significant health problems. Second, many of the best interviewers love a challenge, and doing a study as challenging as this one, a study that some people say cannot be done, is a powerful motivation for many.

This does not mean that every interviewer who self-selects for the study and who is thought suitable by his or her field manager will be able to handle the content and the complexities of the interview schedule. Nevertheless, high positive motivation combined with carefully crafted materials and adept trainers constitute a powerful force for success.

The success of the training plan will be evaluated by a number of methods: Debriefing the trainers and some selected interviewers after the training sessions; monitoring interviewer performance during the field period; and finally, having the interviewers fill out a questionnaire at the conclusion of the field period that will collect their views on the adequacy of their training as well as other reflections on their data collection experience.

Summary of Recommendations

In sum, for training lay interviewers on sensitive topics, the following are recommended:

- Use experienced field managers, or the most experienced senior field staff, as trainers; use experts as consultants.
- Involve the field managers early in developing survey materials; do not present them with a fait accompli in instruments or training materials.
- Use a name with positive connotations for the training sessions dealing with the sensitivity of the material.
- Have the special materials and sessions interspersed throughout the training program, so interviewers can digest the materials and expand their attitudes over time.
- Use interactive sessions that include humor and that stretch the boundaries of the interviewers beyond those of the survey instruments.
- Include practice in questionnaire administration sufficient to allow the interviewers to consolidate their comfort with the subject matter and the survey instruments.

References

Abelson, R. (1981). Psychological status of the script concept. *American Psychologist,* 36, 715-729.

Anderson, J. A. (1969). Single-channel and multi-channel messages. *Audio-Visual Communication Review,* 17(4), 428-434.

Ash, P., & Abramson, E. (1952). The effect of anonymity on attitude questionnaire response. *Journal of Abnormal and Social Psychology,* (47), 722-723.

Barton, S. L. (1958). Asking the embarrassing question. *Public Opinion Quarterly,* 22, 67-68.

Belbin, E., & Belbin, R. M. (1972). *Problems in adult retraining.* London: Heinemann.

Bower, B., Black J., & Turner, T. (1979). Scripts in text comprehension and memory. *Cognitive Psychology,* 11, 177-220.

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

Crowne, D., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence.* New York: Wiley.

Erdelyi, M., & Kleinbard, J. (1976). Has Ebbinghaus decayed over time? The growth of recall (Hypermnesia) over days. *Journal of Experimental Psychology: Human Learning and Memory,* 4, 275-278.

Gagnon, J. (1977). *Human sexualities.* Glenview, IL: Scott, Foresman and Company.

Hartley, J. (1985). *Designing instructional text* (2d ed.). London: Kogan Page.

Hsia, H.J. (1971). The information processing capacity of modality and channel performance. *Audio-Communication Review,* 19(1), 51-75.

Kinsey, A., Pomeroy, W. R., Martin, C. E., & associates. (1948). *Sexual behavior in the human male.* London: W.B. Saunders.

Linton, M. (1982). Transformations of memory in everyday life. In U. Neisser (Ed.). *Memory observed.* San Francisco: U.H. Freeman and Company.

Loftus, E. (1982). Memory and its distortions. In A. G. Kraut (Ed.). *G. Stanley Hall Lectures.* Washington, DC: American Psychological Association.

Mager, R. F. (1972). On the sequencing of instructional content. In I. K. Davies, & J. Hartley (Eds.). *Contributions to an educational technology.* London: Butterworths.

Mager, R. F., & Clark, C. (1963). Explorations in student-controlled instruction. *Psychological Reports,* 13, 71-76.

Mandler, J., & Johnson, N. (1977). Remembrance of things passed: Story structure and recall. *Cognitive Psychology,* 9, 111-151.

McIlvenna, T. (Ed.) (1987). *Complete guide to safe sex.* Beverly Hills, CA: Specific Press.

Merrill, M. D. (1971). Paradigms for psychomoter instruction. In M.D. Merrill (Ed.). *Instructional design: Readings.* Englewood Cliffs, NJ: Prentice-Hall.

Miller, G. (1979). Images and models, similes and metaphors. In A. Ortony (Ed.). *Metaphor and thought.* Cambridge, MA: Harvard University Press.

Miller, G., & Johnson-Laird, P. (1976). *Language and perception.* Cambridge, MA: Harvard University Press.

Orne, M. (1962). On the psychology of psychological experiment: with particular reference to demand characteristics and their implications. *American Psychologist,* 17, 776-783.

Rosenthal, R., & Jacobson, L. (1966). Teachers expectancies determinants of pupils' I.Q. gains. *Psychological Reports,* 19, 115-118.

Rumelhart, D. (1975). Notes on a schema for stories. In D. Bobrow, & A. Collins, (Eds.) *Representation and understanding: Studies in cognitive science.* New York: Academic Press.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding.* Hillsdale. NJ: Lawrence Erlbaum Associates.

Smith, H.C. (1966). *Sensitivity to people.* New York: McGraw-Hill.

Smith, P.B. (1974). The skills of interaction. In P.C. Dodwell (Ed.). *New horizons in psychology (Vol. 2.)* London: Penguin.

Sudman, S. (1967). *Reducing the cost of surveys.* Chicago: Aldine.

Sudman, S., & Bradburn, N. (1982). *Asking questions, A practical guide to questionnaire design.* San Francisco: Jossey Bass.

Winfield, I. (1986). *Learning to teach practical skills.* New York, New York: Nichols.

Yaffé, M., & Fenwick, E. (1988). *Sexual happiness, a practical approach.* New York: Henry Holt and Company.

FEATURE PAPER

Methodological Experiments in the National Survey of Health and Sexual Behavior

Virginia S. Cain

Introduction

The lack of current data on the sexual behavior of the American population has become glaringly apparent with the emergence of the AIDS epidemic. The data available are either outdated (Kinsey, 1948; Kinsey, 1954) or from relatively small-scale studies of special groups within the population such as college students, gay men, or prostitutes (Quinn & others, 1988; Winklestein & associates, 1988). A recent publication provides the most recent national data (1970) on adult sexual behavior (Fay & associates, 1989), yet these data were not collected to address the questions raised by the AIDS epidemic, and the nonresponse rates on some of the items are of concern. The National Survey of Health and Sexual Behavior (NHSB) will attempt to fill in many of the gaps by surveying adults 18 years of age and older about their sexual behaviors and partner characteristics. The current project, conducted by National Opinion Research Center (NORC) with principal investigators Edward O. Laumann, Robert T. Michael, and John H. Gagnon, is a 2,300 case pretest of a questionnaire designed to collect data on an array of sexual behaviors in the general population, age 18–54, in the United States. Table 1 shows the total sample size.

Because sexual behavior is most often regarded as a private matter, the difficulties in obtaining reliable and valid data are of concern. The behaviors that the respondents are asked to report on have the potential for being embarrassing, subject to social disapproval, and illegal. The types of data needed to respond to the issues of the acquired immunodeficiency syndrome (AIDS) epidemic include extramarital sexual interactions, sex with prostitutes, and same-gender sexual contacts. To obtain a complete picture of the sexual lives of the population, it is necessary to gather data on low-risk activities such as masturbation, which may be viewed as even more private than sexual activities that involve a partner (Bradburn & associates, 1978). Many of the problems that exist when any survey is conducted are exacerbated when dealing with a topic that may be viewed as threatening. The subject of sexual behavior may lead to higher refusal rates for participation in the interview. Possibilities exist for overreporting of some behaviors and underreporting of others and more item refusals than typically found in surveys dealing with less sensitive topics.

Despite the many obstacles to collecting accurate data, the need for such information is compelling. The task of the researcher is to find the most effective way of gathering accurate data within the constraints of survey research. Toward this end, the NHSB survey has

Table 1. Interview mode by respondent characteristics*

Subgroups	Interview mode	
	Personal	Telephone
Youth (18–19 years old)	194	18
Hispanic	235	31
Black		
Male	281	18
Female	273	20
All other	1,463	296
Total	2,446	383

* Based on a 75 percent response rate.

Virginia S. Cain is with the Demographic and Behavioral Sciences Branch, Center for Population Research, National Institute of Child Health and Human Development, Bethesda, Maryland.

Table 2. Sample sizes by experimental group and respondent stratum

Respondent group	Interview location/interviewer match
White males	At home, total 382
	127 white male interviewer
	127 white female interviewer
	128 field manager's (FM) discretion
	Local site, total 399
	133 white male interviewer
	133 white female interviewer
	133 FM's discretion
White females	At home, total 401
	all at FM's discretion
	Local site, total 430
	all at FM's discretion
Black males	At home, total 161
	64 black female interviewer
	64 white female interviewer
	33 black male interviewer
	Local site, total 129
	52 black female interviewer
	52 white female interviewer
	25 FM's discretion
Black females	At home, total 143
	71 black female interviewer
	72 white female interviewer
	Local site, total 140
	70 black female interviewer
	70 white female interviewer
Hispanics	At home, total 123
	123 all at FM's discretion
	Local site, total 138
	138 all at FM's discretion

undertaken plans for a pretest of its questionnaire which would include several methodological experiments designed to test the impact of (1) face-to-face interviews versus self-administered questionnaire (SAQ); (2) telephone versus face-to-face interviews; (3) gender and race matching of interviewer and respondent; (4) location of interview, that is, interviewing in the home versus a more neutral site such as an office; and (5) open-ended reinterviews. Sample sizes by experimental group and respondent stratum are presented in Table 2.

Face-to-Face versus Self-Administered Questionnaire

The data in the NHSB survey are collected primarily through a face-to-face, interviewer-administered questionnaire. The type of information needed in AIDS research, particularly AIDS modeling, requires the collection of extremely complicated data. For example, detailed information is collected on the characteristics of the last three sexual partners in the past year and the activities in the last sexual encounter with that person. Less detail is obtained for other partners in the past year and still less on activities in the past 5 years. The skip patterns needed to gather such data are clearly beyond the skills of even the most educated respondent.

A self-administered questionnaire would seem to provide significant advantages in the study of sexual behavior. It is an economical means of data collection that avoids the bias that may be introduced by an interviewer. It affords the respondent the privacy to answer questions that may elicit socially undesirable responses (Knudsen & associates, 1967; Sorensen 1972).

Nonetheless the self-administered questionnaire has serious limitations regardless of the topic. When used in the general population, the questionnaire must be designed for those respondents with rudimentary reading skills and even then will be too advanced for some of the respondents (Jensen & associates, 1987). It is not well- suited to gathering complex data with more than the most basic skip patterns. In these cases it is likely to yield high levels of missing data. Variation in interpretations of items in self-administered questionnaires can introduce bias into the data. A good interviewer can create an atmosphere conducive to open, honest reporting (Kinsey, 1953) whereas a self-administered questionnaire may allow a respondent to more easily refuse to answer a question that is threatening (Johnson & Delamater, 1976).

After the face-to-face interview in the NHSB survey, the respondent will be asked to fill in a short self-administered questionnaire that is placed in an envelope and sealed before returning it to the interviewer. The introduction to the self-administered questionnaire informs the respondent that some of the questions may be repetitious but that "Many people think it may be easier to write the answers to some questions on paper instead of saying them to another person." This allows for the comparison, in several areas, of data collected by the interviewer with that on the self-administered questionnaire. These areas include sex in exchange for money, same-gender sexual contact, and use of intravenous drugs. Data will be analyzed to assess whether the privacy afforded by the self-administered questionnaire produces increased reporting of behaviors that are generally socially disapproved.

Face-to-Face versus Telephone Interviewing

Several of the advantages and disadvantages of face-to-face interviewing have already been discussed. Clear advantages are for the interviewer to be able to develop an atmosphere conducive to reporting sensitive information and the opportunity to gather complex data. The face-to-face interview also permits the interviewer to provide photo identification and other documents allowing the respondent to more easily check the credentials of the interviewer. However, the face-to-face mode of collection is a relatively expensive means of gathering data. Travel time to the respondent's home can be extensive, and completion of the interview may entail several trips.

The telephone interview can provide a reasonable alternative. Like the face-to-face interview, the questionnaire used in a telephone interview can be complex with many complicated skip patterns. It may offer the possibility of substantial cost savings. Interviewers working from telephone banks can conduct many interviews in a single day, and no travel costs are involved. Furthermore, the relative anonymity of a phone interview when compared to a face-to-face interview may result in respondents' increased ease in reporting threatening or embarrassing behavior.

However, the telephone interview is not without problems. The first is that the sample drawn for a telephone survey can only be drawn from those in the population who have telephones. This would result in undercoverage of the minority and lower income populations, populations of particular interest in AIDS-related research. Second, the telephone interview does not lend itself to a lengthy questionnaire. This impacts not only on the amount of data that can be collected but possibly the quality. Results from cognitive psychological studies of memory (Means, 1989) suggest that significant improvement in memory of events, particularly recurring events, occurs when a context for the event is created during the interview. In the shorter telephone interview, the creation of the context must occur at the expense of other substantive areas. A further concern in collecting data on sensitive topics is the opportunity for the interviewer to develop rapport with the respondent before beginning the sensitive questions. The type of questions required for a survey collecting data on AIDS risk behavior includes details of the respondents' sexual encounters. It is not clear that the interviewer can develop sufficient rapport over the telephone to gather high-quality data on sensitive topics from the respondent.

After balancing the various advantages and disadvantages of each of the modes of data collection, the decision was made to conduct the NHSB survey through face-to-face interviews. The in-person interview should provide the highest quality data with the least amount of underreporting of sensitive data. However, since there are many advantages to telephone interviews, the NHSB survey will administer a subset of the items from the larger survey to approximately 300 telephone respondents. Unfortunately, the experimental design will not permit the decomposition of the effect into that due to shorter interview form or mode of data collection.

Gender and Race Matching of Respondent and Interviewer

Face-to-face interviews have the potential of introducing bias into the data owing to the necessary interaction between the interviewer and the respondent. The presentation of self by the respondent may be influenced by characteristics of the interviewer (Bradburn & associates, 1979). Although this is always an issue in either face-to-face or telephone interviewing, the sensitive nature of the topic of sexual behavior heightens the concern. The literature is mixed in its view of the importance of matching the gender and race of the respondent with that of the interviewer. Researchers coming from a clinical perspective recommend matching on gender, race, and age of interviewer and respondent. Other studies have found little impact of sex of interviewer on the responses of either men or women about their sexual behaviors (Darrow & associates, 1986; Johnson & Delamater, 1976). The effect of an interviewer's race on a respondent's answers was greatest on questions regarding racial issues such as race relations (Bradburn & Sudman, 1979).

While the importance of interviewer-respondent rapport is recognized, there is some concern that too much rapport can develop and lead respondents to report behavior that will increase their status in the eyes of the interviewer (Siegel & Baumann, 1986). This has led to the suggestion that interviewers and respondents be dissimilar on one or two key characteristics to avoid too much rapport (Williams, 1968).

As one of the methodological experiments included in the pretest of the NHSB survey, some respondents will be matched to interviewers of the same gender and race. For the most part, interviews will be conducted by the standard national field interviewing staff, predominantly white and female. No special matching will be done for the white female respondents. Among white male respondents, some will be interviewed by white males and some by white females. Black female respondents will be interviewed by either black female or white female interviewers. The number of black males in the NHSB pretest does not permit a test of the four gender and race possibilities for the interviewer characteristics. Interviewers of black males will be either white females, black females, or black males.

The matches proposed for the NHSB pretest were in part driven by the characteristics of the national pool of interviewers. The costs of the survey can be reduced if it is possible to hire interviewers from this pool, which is predominantly white and female. It is less costly to use experienced interviewers who will be extensively trained on the sensitive topics in the present survey than to recruit and sufficiently train inexperienced interviewers. The cost of the gender and race matching will be balanced against the quality of the data gathered. When possible, data will be examined for the impact of interviewers' characteristics on both the underreporting and overreporting of some behaviors. For example, young male respondents may overreport sexual activities to young male interviewers.

Location of Interview

The collection of high quality data requires an atmosphere in which the respondent feels comfortable in admitting to socially undesirable behavior. In addition to characteristics of the interviewer, the characteristics of the physical setting in which the interview takes place may impact on the quality of the data collected. Generally it is preferable to conduct an interview in private to ensure that others in the vicinity do not distract the respondent or influence the responses. The highly personal nature of the data collected in the NHSB survey

make it critical that privacy be maintained. However, concerns about privacy may be only one of the ways in which the physical location of the interview impacts on the data collected. In a study of gay men, Darrow and others (1986) hypothesized that the men would be more candid in reporting sexual behaviors and drug use when interviewed in a familiar setting. Results of the analysis did not show any differences that could be attributed to place of interview.

In the NHSB survey, the expected direction of the effect of location of interview on data collected is opposite to that hypothesized by Darrow and colleagues (1986). It has been suggested that respondents may be less likely to report extramarital sexual contacts or earlier sexual relationships while being interviewed in their own home. A more neutral setting such as a room in a clinic or community center may be more conducive than the respondent's home to the accurate reporting of such behavior.

The NHSB survey will test the effect of interview location by randomly assigning respondents to one of two interview situations: in home or local site. Sites will be chosen to provide neutrality, privacy, convenience, and security.

Open-Ended Reinterviews

The question of validity often arises with respect to survey data, and it is of particular concern when respondents are asked to report on sensitive and threatening topics. Research has shown that the amount of reporting of sensitive behaviors, particularly those behaviors that can be quantified, can be increased by allowing the question to be constructed with an open-ended question format (Bradburn & Sudman, 1979). However, the financial and logistical constraints of a large-scale survey make it infeasible to employ this type of question construction for more than a small subset of the questions.

One of the methodological experiments included in the NHSB pretest is an open-ended reinterview of approximately 135 respondents to determine how much of the data is lost or distorted by the use of standard closed-ended survey questions, and which of the questions are most subject to influence by question format. One hundred of the respondents will be selected on the basis of demographic characteristics for reinterview before their original interview. Approximately 35 respondents will be selected for reinterview on the basis of certain behavioral characteristics determined at the first interview or seemingly inconsistent information. The reinterviews will be highly clustered geographically and conducted within a month of the original interview to minimize problems of recall and changes in context between the interviews.

Several members of the research team will carefully review the original interview conducted with a respondent selected for open-ended reinterview. The reinterview will be specifically designed for the individual respondent based on previous responses. Certain topics, such as behaviors that put people at risk of HIV infection or other potentially sensitive topics, will be included in all of the reinterviews. Other topics will be included because they resulted in ambiguous or inconsistent data in the first interview. The interviews will be conducted by members of the research team or specially trained interviewing personnel.

The effects of the previously described experiments will be assessed through a variety of means. First, general issues of cooperation will be addressed. The interviewer will provide a subjective assessment of how cooperative the respondent was. More objective assessments can be made of the number of initial refusals, the number of contacts required to get a complete interview, and the number of interviews that were terminated prematurely. Following completion of the interview, questionnaires will be examined for item nonresponse and the level of reporting of private behavior.

Extensive efforts will be made to obtain a completed questionnaire from all respondents. If necessary, the methodological experiment will be abandoned rather than lose a case to a refusal. At the point when a field manager would go beyond routine procedures to complete a case, an interim evaluation of the case will be made.

Summary

The onset of the AIDS epidemic has made salient the need to understand human sexual behavior. As one of the main routes of transmission of HIV is through sexual behavior, there is a great need for data on detailed specific behaviors and partner characteristics. The purpose of the pretest of the National Survey of Health and Sexual Behavior is twofold. First, it will provide a test of a questionnaire designed to collect sexual behavior from the adult population in the United States in sufficient detail for use in AIDS education and modeling efforts. Second, it will test the effectiveness of a number of different methodologies designed to elicit accurate, reliable data on very sensitive topics.

References

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design.* Washington, DC: Jossey Bass.

Bradburn, N. M., Sudman, S., Blair, E., & associate. (1978). Question threat and response bias. *Public Opinion Quarterly*, 42, 221-234.

Darrow, W. W., Jaffe, H. W., Thomas, P. A., & associates. (1986). Sex of interviewer, place of interview, and responses of homosexual men to sensitive questions. *Archives of Sexual Behavior,* 15(1), 79-88.

Fay, R. E., Turner, C. F., Klassen, A. D., & associate. (1989). Prevalence and patterns of same-gender sexual contact among men. *Science,* 243, 338-348.

Jensen, B. J., Witcher, D. B., & Upton, L. R. (1987). Readability assessment of questionnaires frequently used in sexual and marital therapy. *Journal of Sex and Marital Therapy,* 13, 137-141.

Johnson, W. T., & Delamater, J. P. (1976). Response effects in sex surveys. *Public Opinion Quarterly,* 40, 1965-181.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male.* Philadelphia: Saunders.

Kinsey, A.C., Pomeroy, W. B., Martin, C. E., & associate. (1954). *Sexual behavior in the human female.* Philadelphia: Saunders.

Knudsen, D. D., Pope, H., & Irish, D. P. (1976). Response differences to questions on sexual standards: An interview-questionnaire comparison. *Public Opinion Quarterly,* 31, 290-297.

Means, B. M. (1989, January). Memory issues in survey reports: A cognitive interview technique for reducing response errors. Paper presented at the winter meeting of the American Statistical Association, San Francisco, California.

Quinn, T. C., Glasser, D., Cannon, R. O., & associates. (1988). Human immunodeficiency virus among patients attending clinics for sexually transmitted diseases. *The New England Journal of Medicine,* 318(4), 197-203.

Siegel, K., Baumann, L. J., & Feldman, D. (1986). Methodological issues in AIDS-related research. In D. Feldman & T. M. Johnson (Eds.). *The social dimensions of AIDS* (pp. 15-39). New York: Praeger.

Sorensen, R. C. (1972). *Adolescent sexuality in contemporary America.* New York: World Publishing.

Williams, J. A., Jr. (1968). Interviewer role performance: A further note on bias in the information interview. *Public Opinion Quarterly,* 32, 287-297.

Winkelstein, W. Jr., Wiley, J. A., Padian, N.S., & associates. (1988). The San Francisco Men's Health Study: Continued decline in HIV sero-conversion rates among homosexual/bisexual men. *American Journal of Public Health,* 78(11), 1472-1474.

FEATURE PAPER

Comparison of Results of Personal Interview and Telephone Surveys of Behavior Related to Risk of AIDS: Advantages of Telephone Techniques

David V. McQueen

Introduction

A major study at the Research Unit in Health and Behavioural Change (RUHBC) at the University of Edinburgh involves face-to-face and computer-assisted telephone interviewing (CATI) surveys of health behaviors such as smoking, drinking, exercise, diet, sexual and acquired immunodeficiency syndrome (AIDS)-related behaviors, attitudes, opinions, beliefs, and knowledge. The CATI-based survey is designed to track knowledge, opinions, and behaviors over time among 18 to 60-year-olds resident in central Scotland and, recently, parts of England. Data collection started in July 1987. By the end of May 1988 face-to-face interviews had been conducted with 2,396 persons and computer-assisted telephone interviewing with approximately 4,000 persons. By the end of March 1989 computer-assisted telephone interviewing had been carried out on a continuing basis, with some 8,995 interviews attained.

The study examines different methodological approaches. In data collection there are three key issues: (1) comparisons between CATI and traditional survey techniques; (2) sampling issues related to continuous data collection; and (3) reliability and validity of data collected on sensitive behaviors. In data analysis there are six key issues: (1) comparing data collected by different methods; (2) mixing data collected by different methods; (3) weighting data to account for telephone coverage; (4) analysis of data that are dynamic; (5) accounting for multiple outcomes; and (6) mathematical analysis.

This paper concentrates on risky and socially stigmatized behavior, particularly sexual behavior related to AIDS, in relation to comparisons of techniques and collecting sensitive data. The overall strategy is to examine differences in reporting by mode of interview; then, within the face-to-face interviews to examine the effect of telephone coverage. The general hypothesis is that if there is no reporting difference between those who have and do not have a telephone then differences between computer-assisted telephone interviewing and face-to-face interviews may be explained by either true mode differences or sampling variation or both. If there is a difference by mode then the question is, which mode sets the standard? Differences on sensitive questions by mode are examined in detail. Using CATI to collect sensitive data provides reliability of reported data over time. Differences due to mode effect may be taken into account by weighting telephone data to reflect class and telephone coverage bias.

Mode Comparisons

The data on more than 3,400 adults aged 18 to 44 years living in central Scotland during the period July to

Table 1. Telephone coverage by age in four face-to-face survey areas

	Areas			
Age	P[a]	G[b]	F[c]	STS[d, e]
18-21	86.6	26.7	65.1	65.5
22-29	81.8	37.7	50.6	68.0
30-39	92.4	47.0	65.8	77.3
40-44	88.9	61.4	74.1	83.5

[a] N = 624, $p < 0.01$, G = −.21
[b] N = 493, $p < 0.0001$, G = −.32
[c] N = 496, $p < 0.01$, G = −.24
[d] N = 753, $p < 0.01$, G = −.24
[e] System 3

David V. McQueen is with the Research Unit in Health and Behavioural Change, University of Edinburgh, United Kingdom.

Special thanks go to research assistants Beatrice Robertson, Rebecca Smith, and Jane Hopton, research fellows Candace Currie and Daan Uitenbroek, and the Unit's administrative secretary Emmanuelle Tulle-Winton. The work is funded by a grant from the Scottish Home and Health Department, the Scottish Health Education Group, and the Economic and Social Research Council. Any statements about the data are the author's and do not necessarily represent the opinion of the funders or the Research Unit in Health and Behavioural Change.

Table 2a. Perception of AIDS as a threat to the nation's health by survey area: Males

AIDS a threat	Areas				
	CATI[a]	STS[b]	P[c]	G[d]	F[e]
Yes	64.7	69.7	74.4	88.1	88.9
No	15.9	12.9	11.5	4.8	6.8
Could become so	19.5	14.0	13.2	6.0	3.4
Do not know	. . .	3.4	0.9	1.2	0.9

[a] N = 511
[b] N = 364
[c] N = 263
[d] N = 193
[e] N = 191

Table 2b. Perception of AIDS as a threat to the nation's health by survey area: Females

AIDS a threat	Areas				
	CATI[a]	STS[b]	P[c]	G[d]	F[e]
Yes	73.1	83.8	85.4	86.4	86.2
No	9.1	6.5	5.1	5.6	8.5
Could become so	19.5	7.2	9.2	7.2	4.2
Do not know	. . .	2.5	0.3	0.8	1.1

[a] N = 662
[b] N = 378
[c] N = 361
[d] N = 299
[e] N = 304

September 1987 provide comparisons. Three samples of households from three geographical areas, selected on the basis of estimated poor telephone coverage, were interviewed by traditional methods; two areas were chosen to represent worst cases for coverage, a very deprived central Edinburgh area (termed "G") with 42.4 percent coverage ($N=493$); and a deprived suburban region (termed "F") with 61.9 percent coverage ($N=496$). The third area (termed "P") was mixed, with some deprivation and coverage of 87.3 percent ($N=624$). Because of concerns for both the respondents

Table 3. Comparison of mode of interview with reported sexual behaviors, by sex

Sexual behaviors	Percentage		
	CATI[a]	Face-to face[b]	Levels of significance
Changed behavior due to AIDS			
Male	15.5	12.4	ns
Female	9.1	8.2	ns
Reported other partners during past year			
Male	11.8	14.5	ns
Female	4.3	7.8	*
Four or more partners in past 5 years			
Male	22.9	14.7	**
Female	5.2	3.8	ns

[a] Males—N = 511; Females—N = 662
[b] Males—N = 364; Females—N = 378
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
ns = not significant

Table 4. Comparison of telephone owners with reported sexual behaviors, by sex

Sexual behavior	CATI[a]	Face-to face[b]	Levels of significance
Changed behavior due to AIDS			
Male	15.5	10.2	ns[c]
Female	9.1	9.1	ns
Reported other partners during past year			
Male	11.8	13.0	ns
Female	4.3	7.5	ns[d]
Four or more partners in past 5 years			
Male	22.9	12.3	***
Female	5.2	3.6	ns

[a] Males—N = 511; females—N = 662
[b] Males—N = 364; females—N = 378
[c] $p < 0.06$
[d] $p < 0.08$
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

and interviewers the initial assessment of the three areas led to the decision that explicit questions regarding sexual practices relevant to AIDS would be ill-advised in the first instance. Therefore primarily attitudinal questions as well as questions about other health-related behaviors were asked.

A detailed review of the three surveys led to the decision to undertake a fourth survey, which consisted of a probability sample of the population of Central Scotland, aged 18 to 44 years, interviewed face-to-face, using the full CATI-based questionnaire, including more explicit sexual behavior questions related to AIDS. To maximize the match to the concurrent CATI interviews and maintain independence, data were collected by System 3 Scotland (STS), an experienced Scottish survey research organization; the sampling frame and strategy attempted to match that of the CATI system. The resulting sample was biased toward lower socioeconomic status (SES) and had a coverage of 72.8 percent ($N=743$).

Tables 1 to 5 present some of the results and are arranged in the following order: telephone coverage, comparison of computer-assisted telephone interviewing with all face-to-face survey areas; effect of telephone ownership (CATI versus STS); effect of telephone ownership in lower class groups.

In general the age profiles of the four face-to-face samples and the CATI sample were very similar (Table 1). These and other demographic details are provided

Table 5. Telephone coverage, by occupational class, in face-to-face interviews

Phone coverage	Occupational classes				
	A & B[a]	C[b]	C2[c]	D[d]	E[e]
Phone ownership	97.6	81.3	76.7	65.2	39.2
Nonphone ownership	2.4	17.7	23.3	34.8	60.8

[a] N = 85
[b] N = 198
[c] N = 215
[d] N = 138
[e] N = 107
NOTE: N = 743, $p < 0.001$, G = +.53

Table 6. Reported sexual behaviors within classes D[a] and E[b], by phone ownership

Phone coverage	Types of behavior			
	A[c]	B[d]	C[e]	D[f]
Phone ownership	13.8	48.4	36.1	94.5
Nonphone ownership	13.2	45.5	34.0	96.4

[a] N = 138
[b] N = 107
[c] Respondents who had more than four partners in the past 5 years, ns
[d] Respondents who have ever used condoms, ns
[e] Respondents who use condoms now, ns
[f] Respondents who say there is nothing they do in their daily life which puts them at risk of getting AIDS, ns

by Research Unit in Health and Behavioral Change, unpublished data, 1987, 1988. The relationship of age to phone ownership is quite marked in the two more deprived areas (G and F) as seen in Table 1, but when a higher level of socioeconomic status is obtained the age effect disappears. With respect to perceptions, some variability occurred in communication and concern about AIDS as reported by men and women in the five surveys. In general there was more difference in men than in women and Table 2 illustrates this. Many comparisons were made between data collected by computer-assisted telephone interviewing and face-to-face; the vast majority of respondents (88.3 percent of computer-assisted telephone interviewing and 89.8 percent of face-to-face) reported not having changed behaviors due to what they know about AIDS. The difference between the two modes of interview was not significant. Table 3 illustrates this and shows sex differences. A similar pattern followed questions on intentions to change behavior. Face-to-face interviewees reported slightly more likelihood of having had other sexual partners ($p < 0.01$). Nonetheless, Table 3 shows that CATI respondents reported more partners in the past 5 years. When one compares CATI responses with STS respondents who reported owning a telephone, responses are not substantively nor significantly different (Table 4). Finally, a number of analyses considered the question of telephone ownership with particular emphasis on those lower classed occupational groups (D and E) where telephone ownership is low. Obviously there is a strong and significant correlation ($p < 0.001$; Gamma = -0.53) between telephone ownership and class in Scotland (Table 5). The question is whether this results in significant differences in reported behavior, and the answer seems to be no, thus introducing the applicability of appropriate weighting strategies (Table 6).

How responses to questionnaire items differ by mode of interview has required a detailed examination of each questionnaire item, taking into account slight differences in population distributions. In general, attitudinal questions are more volatile and subject to greater "interviewer effect." As would be expected, the CATI method tends to reduce this effect. Although general behavioral questions tend to elicit similar responses in both modes of interview, evidence is accumulating that sensitive questions are answered with a higher reporting tendency in CATI. Several items on sexual behavior tend to have higher reporting rates (Table 7). The question of validity remains one for further research, given the paucity of information on current sexual practices.

Table 7. Higher rates of reporting on sensitive questions, by mode of interview

	Percentage		
	CATI	Face-to face	Difference
Four to ten partners in last 5 years	9.7	6.4	+3.3
Many partners	3.3	2.6	+0.7
Reported past sexual activities with same sex	2.2	0.9	+1.3
Current homosexual behavior	45.5	28.6	+16.9
Ever used condoms	56.5	51.1	+5.4
Use condoms now	32.4	31.6	+0.8
Risk from AIDS	6.9	4.2	+2.7

The effect of telephone ownership and response to interview items should be seen as: Can one predict the response to a questionnaire item if one knows phone ownership? The answer is: No, not very well once class and age factors are taken into account. This critical factor makes a weighting strategy possible and useful. Weighting that takes into account telephone ownership in lower social groupings can be devised, but should be in constant review to account for changes in coverage and to provide more accurate weighting for the continuous data. Currently, a poststratification weighting that takes into account coverage by class is applied; the resulting differences between weighted and unweighted data are generally quite small and statistically insignificant; nonetheless weighting is useful when comparing the CATI data with other surveys, although weighting has little effect on long term trends in behavioral change.

Table 8. Reported change in behavior, by variables related to AIDS: Highly significant associations in 18-44 year olds, July-February 1988[a] and March-December 1988[b]

Variables	Levels of significance		Gamma	
	Jul-Feb	Mar-Dec	Jul-Feb	Mar-Dec
Talk with friends	***	***	0.43	0.48
Concern someone close will get AIDS	***	***	.49	.39
Together less than 5 years	***	***	.37	.45
Other partners past year	***	***	.49	.58
Number of partners past 5 years	***	***	.38	.47
Risk of getting AIDS	***	***	.1	.59[c]

[a] N = 2,181
[b] N = 3,338
[c] June-December 1989
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9. Reported change in AIDS-related behavior, by variables related to knowledge, attitudes, and beliefs: March-December[a]

Variables	Levels of significance	Gamma
Knowledge	***	0.32
Source	***	.27
Insect bites	ns	.16
Donate blood	ns	—
Eating food	*	—
Problem in community	***	.40

[a] N = 3,338
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Behavior Related to AIDS: CATI Data

In the CATI-based questionnaire respondents are asked if they have changed any behaviors due to what they know about AIDS; if they respond positively they are asked what they have changed. In the data collected from July 1987 through February 1988, approximately 12 percent (263) reported changing their behavior; from March to December approximately 10 percent of those respondents aged 18 to 44 years reported in the affirmative (Tables 8 and 9). There are three chief concerns: (1) what is being changed; (2) why these individuals are changing their behavior; and (3) the characteristics of the changers.

Table 8 summarizes some key associations that relate to reported behavioral change. These variables are highly significant and strongly associated (Gamma) with change. Furthermore, these associations remain strong over time. Respondents who report changing their behavior talk about AIDS with family and friends, but it is clear that interpersonal relations and conversations among friends are much more important than family discussions. Over 80 percent of reported changers expressed high concern over themselves or someone close to them getting AIDS in contrast to about 50 percent of nonchangers. Of those expressing high concern, most (about 80 percent) reported high use of condoms, fewer partners, and safe sex as a response to this concern.

Furthermore, although most demographic variables had only a minor relationship to this concern, social class did have a significant ($p < 0.001$) inverse relationship.

Stable, monogamous relationships without a history of multiple partners correlated strongly to the reported changing of behavior, with these respondents reporting much less change in behavior. Finally, perceived risk of AIDS, based on the response to the question, "Is there anything that you do in your daily life which you think puts you at risk of getting AIDS," was very significant in respondents reporting change.

This emerging pattern for behavioral change with regard to these social factors remains steady in the data over time, that is, the pattern continues strongly in the data collected from March through December 1988; regression analyses verify that prediction of changes in behavior may be based principally on the variables in Table 8.

As mentioned earlier, items related to knowledge, attitudes, and beliefs were added to the CATI questionnaire beginning in March 1988. Although few of these variables reach levels of significance or strong relationships, some are worth noting (Table 9), notably the items "knowledge about AIDS" and "AIDS could become a problem in the respondent's community." Significantly, those individuals who assess that they have a lot of knowledge about AIDS are more likely to change their behavior; this is consistent with the question on whether the respondent reports changing behavior. More importantly, the strong relationship to behavior change reported in response to the perception of AIDS becoming a problem in the respondent's community is in marked contrast to the less significant response to the perception of AIDS as a serious threat to the nation's health ($p < 0.05$).

Reported intention to change behavior may also be viewed as useful data. In the mid-July 1987 to end of February 1988 data, 9 percent (N = 170) reported an intention to change; this pattern remains similar in the March 1988 to December 1988 data. Table 10 shows the pattern of significant variables related to this expression of intention, and the general pattern which is similar to that of respondents who reported changing behavior, the major exception being the variable of perceived risk of getting AIDS. Further analysis may reveal the role of

Table 10. Reported intention to change, by variables related to AIDS: July-February[a]

Variables	Levels of significance	Gamma
Talk with friends	***	0.38
Concern someone close will get AIDS	***	.31
No steady partner	***	.46
Together less than 5 years	***	.44
Other partners past year	***	.57
Number of partners past 5 years	***	.32
Risk of getting AIDS	ns	—

[a] N = 2,181
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11. Reported change in or intention to change behavior, by demographic variables: July-February[a]

Variables	AIDS change	(Gamma)	Will change	(Gamma)
Age	***	(0.23)	**	(0.21)
Education	ns		ns	
Unemployed	**	(.05)	**	(.12)
SES	ns		ns	
Marital status	***	(.32)	***	(.40)
Sex	***	(.22)	*	(.16)

[a] N = 2,175
NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12. Reported change or intention to change, by reported marital status, 18-44 years, July-February[a]

Status	AIDS change %	AIDS change (number)	Will change %	Will change (number)
Married (N=1,136)	7.5	(85)	4.0	(46)
Member of unmarried pair (N=93)	9.7	(9)	8.6	(8)
Separated (N=69)	13.0	(9)	18.8	(13)
Never been married (N=752)	17.7	(133)	11.2	(84)
Divorced (N=105)	21.0	(22)	17.1	(18)
Widowed (N=20)	15.0	(3)	5.0	(1)

[a] N=2,175

risk in converting those who intend to change to actual changers.

Associations between demographic variables and both reported change and intention to change behavior are substantively weak, even though in some cases significant (Table 11). Men are more likely to be changers than women. Both reported changers, and those reporting intention to change were younger; this change pattern has a direct linear relationship to age. Although social class is insignificant, there is a significant though weak association with being unemployed. Marital status has an obvious strong association which is detailed further in Table 12.

Stability of the Data Over Time

Given nearly 2 years of continuously collected CATI data, many over-time comparisons have been carried out. Some of these have been reported in *dataupdates* (1988 and 1989, unpublished data); others remain as exploratory data runs and graphs. A thorough analysis of the data by time requires several considerations: (1) considerable time for preparing the data so it is methodologically suitable; (2) a stabilized data collection so that sample differences and poststratification weighting are minimized; and (3) a long enough run time with identical questions. Despite these analytical hurdles trends have emerged; Table 13 presents a summary of trends in the data from July 1987 to May 1988. Data sampling times have been reviewed on a daily, monthly, and bimonthly basis. The table reports the high and low percentages of the item over the time period on a monthly basis. A rough estimate of the trend direction is offered as well as a subjective overall estimate of the trend's characteristics.

Conclusions

With respect to the comparison of telephone versus face-to-face interviewing, many of the same methodological issues that have been considered in detail in North America have arisen in Britain, and it is doubtful that in the long run they will be resolved much differ-

Table 13. Time trends: Sexual behavior, all respondents: July 87-May 88

Question	Hi %	Low %	Mean %	Diff %	Trend Dir	Trend Char
Have you changed anything in your everyday life due to what you know about AIDS? [Yes]	14.3	8.0	11.1	6.3	unclear	variant
Is there anything you will change? [Yes]	11.7	7.4	9.1	4.3	level	steady
Do you have a steady partner at the moment? [Yes]	81.3	72.2	77.4	9.1	level	steady
How long have you been together? [>5 years]	75.1	64.3	69.0	10.8	up	variant
During the past year have either of you had any other partners? [Yes]	9.0	5.9	7.4	3.1	level	steady
Over the past 5 years, about how many partners did you have altogether?						
Only 1	66.6	58.5	63.4	8.1	level	steady
2-3	19.3	17.1	18.1	2.2	level	steady
4 or more	15.1	9.4	12.3	5.7	unclear	variant
None	8.9	4.5	6.2	4.4	level	steady
Have you ever used condoms during sexual activities? [Yes]	61.8	54.7	59.9	7.1	up	variant

ently in Britain. What is of interest in telephone interviewing as an approach to assessing AIDS-related sexual behavior is that there seems to be little evidence to date of any advantage of face-to-face interviewing, even within the lower class groups.

From the beginning of this study the concern has been with the feasibility of asking questions about intimate sexual behavior that could relate to the AIDS problem. The strategy was to introduce questions about AIDS and sexual behavior over a period of time to obtain feedback on any difficulties. The CATI system is ideal for such a strategy because of the ability to review responses on a frequent basis. This has resulted in four major revisions of the questionnaire structure. Initially, only attitudinal and knowledge type questions about AIDS were asked; later, after a brief warning of the explicit nature of the questions to come, sexual practice and behavior questions were asked and male respondents only were asked about homosexual practices. To date, neither respondents nor interviewers have experienced problems with these sexual behavior questions, and over a long time only 2 to 5 percent of respondents have expressed a desire to not answer these sensitive questions.

Initially, CATI merely added speed to survey interviewing, but gradually the potential for using the system as more than a substitute for pencil and paper has emerged. This potential cannot be easily separated from conceptual and theoretical considerations. Three key concerns are: (1) stability and change in behavior over time; (2) the context in which behavior changes or remains stable; and (3) the collection and interpretation of so-called sensitive information, for example sexual behavior.

Many research questions need to be answered before the potential of computer-assisted telephone interviewing can be realized. Nonetheless, the assertion is that the CATI-based system is very appropriate to the particular problem of sensitive questions related to the spread of AIDS in the general population. It is a system well suited to tracking the dynamics of health-related behaviors over time; examining inter-relationships among fluctuating behaviors over time; and providing baseline data for utilization.

FEATURE PAPER

Effects of Mode of Data Collection on the Validity of Reported Drug Use

William S. Aquilino and Leonard A. LoSciuto

Introduction

Relatively little research has been done concerning the efficacy of telephone surveys for highly threatening interview topics such as sexual behavior and drug usage. Several researchers have expressed skepticism about the ability of telephone surveys to secure honest reporting of sensitive or illicit behavior (Johnston & O'Malley, 1985; Sudman, 1976; Freeman & associates, 1982). Sensitive surveys have typically relied on costly and time-consuming face-to-face interviewing. This research addressed the question: To what extent does the validity of sensitive surveys differ by mode of interview? Specifically, will the validity of drug-use estimates differ between face-to-face and random digit dialing (RDD) telephone surveys? If so, to what extent are differences attributable to sample characteristics or to the interview situation itself?

The exclusion of nontelephone households in RDD sampling is a potentially serious bias for drug use surveys. Respondents from nontelephone households are more likely than those with telephones to be nonwhite; of lower income and education (Groves & Kahn, 1979); never married, divorced, or separated (Tull & Albaum, 1977); single heads of household under age 35; and to live in rural nonfarm areas (Freeman & associates, 1982). To the extent that race, income, education, and marital status are correlated with drug use, RDD sampling may introduce bias in drug-use estimates.

William Aquilino is with the Center for Demography and Ecology, University of Wisconsin, Madison. Leonard A. LoSciuto is with the Institute for Survey Research, Temple University, Philadelphia.

Survey data were collected by the Institute for Survey Research, Temple University. The study was supported by Grant No. R01-DA04280 from the National Institute on Drug Abuse, Bethesda, Maryland, and by PHS National Research Service Award No. AG00129-03 from the National Institute on Aging. The personal interview data were collected by the New Jersey Household Survey on Drug Abuse, funded by the Department of Public Health, State of New Jersey.

Apart from sampling bias, interview modes may differ in susceptibility to response set bias, most notably socially desirable responding in the face of threatening survey items (Mensch & Kandel, 1988). Use of illicit drugs, such as marijuana and cocaine, and heavy versus social drinking are assumed to be socially undesirable behaviors for the majority of respondents. Thus, to the extent that social desirability affects responses and respondents feel threatened by such interview questions, underreporting of drug use is the largest threat to the validity of drug-use surveys. The more socially unacceptable the substance, the greater the underreporting should be. The influence of the interview milieu on response sets should become greater as items become more threatening to respondents (Sudman & Bradburn, 1974).

Telephone and face-to-face modes may differ in their ability to reduce tendencies toward socially desirable responding. The face-to-face interview can take advantage of one of the best methods for reducing response set effects due to social desirability, the use of self-administered questionnaires (SAQs) (Sudman & Bradburn, 1974). The personal drug survey reported here used self-administered answer sheets to maximize respondents' privacy in reporting drug use. Mode differences for threatening questions would seem to be more likely when face-to-face surveys incorporating self-administered questionnaires are compared to standard telephone surveys. Thus, it was predicted that a telephone survey would yield lower estimates of drug use than an in-person survey, and that the degree of underreporting would vary with the sensitivity of the question.

Study Design

Personal Survey

Recently, several state governments have undertaken surveys to chart statewide trends in drug use. One such

survey, conducted in 1986–1987 for the State of New Jersey, provided the face-to-face interview data for this study. For this survey, a multistage area probability sample of New Jersey's civilian, noninstitutionalized household population aged 18 to 34 years was drawn. The sample consisted of 4,571 residential addresses. A screening rate of 90.5 percent was achieved, and interviews were completed in 79.1 percent of the households determined to include an eligible respondent aged 18 to 34 (N = 1,042 completed interviews).

The personal interview averaged 45 to 60 minutes. Drug categories included tobacco, alcohol, marijuana, cocaine, opiates, hallucinogens, and the nonmedical use of prescription drugs. For all but tobacco, answers to drug use items were recorded by respondents on self-administered answer sheets. Interviewers read the instructions at the start of each drug sequence, and at the respondent's request read the questions aloud while the respondent completed the answer sheet. Answer sheets were sealed in an envelope in the respondent's presence, after completion of the interview. No names were recorded on questionnaires or answer sheets. The field period extended from June 1986 through January 1987.

Telephone Survey

The overriding concern in designing the telephone survey was comparability to the in-person survey. Thus it was paramount that the telephone survey attempt to (1) achieve high screening and interview response rates, (2) minimize item nonresponse, (3) preserve question content and meaning, and (4) guarantee confidentiality to respondents. The goal was the completion of at least 2,000 interviews with respondents 18 to 65 years old. This paper reports substantive results for 18- to 34-year-olds only, the same age group available in the face-to-face survey.

To maximize response rates, a 25-minute telephone interview was constructed by adapting only a subset of the in-person interview sequences. The complete sequence of items about tobacco, alcohol, marijuana, and cocaine was asked. These drugs were selected because they are relatively prevalent, have potential for abuse, represent both legal and illegal substances, and form a continuum from relatively nonthreatening (smoking) to threatening (heavy drinking, marijuana, and cocaine use).

The Waksberg procedure (Waksberg, 1978) of random digit dialing (RDD) was employed for telephone sampling to maximize the proportion of residences contacted during screening, thus reducing survey costs and field time. The sample consisted of 6,932 telephone numbers; 2,075 interviews were completed with adults age 18 to 65 years, of whom 864 were 18- to 34-year-old respondents. A screening response rate of 71.5 percent was achieved, with an interview response rate of 80.3 percent. The interview rate was slightly higher for 18- to 34 year-old respondents (82.5 percent) than for 35- to 65-year-olds (78.8 percent). The data collection period for the telephone survey was October through December 1987.

Response rates of the two surveys differed in the screening rate only (71.5 percent telephone, 90.5 percent in person). The number of screening refusals was higher by telephone than in person (14.9 to 3.3 percent). The interview completion rate with selected respondents was actually slightly higher by telephone (82.5 percent) than in person (79.1 percent).

Results and Discussion

Results will be reported for black and white respondents only. The telephone survey obtained only 60 interviews with Hispanics (representing Puerto Ricans, Cubans, and Chicanos) in the 18- to 34-year-old range, a sample too small to allow for adequate demographic controls and tests for interactions. The findings below are based on telephone interviews with 104 blacks and 636 whites, and personal interviews with 158 blacks and 747 whites.

Sample Characteristics

Demographic characteristics of the telephone and face-to-face samples are presented by race in Table 1. There are substantial differences by mode in the demographic characteristics of blacks, fewer differences for whites. For blacks, the RDD sample yielded greater proportions of married (27 to 17 percent) and lower proportions of divorced or separated respondents (5 to 16 percent) than the face-to-face sample. Unemployed blacks were significantly underrepresented in the telephone sample (4 to 15 percent), whereas the full-time employed were overrepresented (78 to 57 percent). The largest difference for blacks was in income: by telephone, only 16 percent of blacks reported personal incomes below $7,000, compared to 44 percent face-to-face. By telephone, 61 percent of blacks reported incomes of $15,000 or more, compared to only 34 percent in the face-to-face interview. For the 18- to 34-year-old population, then, the switch from area probability to RDD sampling shifted the socioeconomic status (SES) of black respondents substantially upward. The demographic profile of whites in the two survey modes is nearly identical.

The exclusion of nontelephone households did not account for demographic differences for blacks in the random digital dialing and face-to-face samples. Demographic profiles were re-calculated after dropping the nontelephone households from the face-to-face sample. Although over 13 percent (21 cases) of black respondents were dropped from the face-to-face sample, the demographic distributions remained essentially the same as for the full sample comparison.

Bias in Drug-use Reports

Four substances of high use prevalence were chosen for these analyses: tobacco, alcohol, marijuana, and cocaine. Dependent variables were selected to furnish a representative picture of lifetime and current usage within the four drug categories. In analyzing drug data, responses from both surveys were pooled into one data set, with mode of interview added to the pooled data set as a categorical variable. Demographic characteris-

Table 1. Sample characteristics by race and interview mode: Percentage distributions

	Black		White	
	Telephone	Personal	Telephone	Personal
N of cases	(104)	(158)	(636)	(747)
Age				
18–21	19	25	18	21
22–25	18	20	22	21
26–29	26	27	24	25
30–34	37	29	37	33
Sex				
Male	34	34	45	44
Female	66	66	55	56
Marital status				
Married	27	17*	42	42*
Widowed	2	1	—	—
Divorced or separated	5	16	7	6
Cohabiting	7	6	5	2
Never married	60	60	46	49
Education				
Less than high school	10	18	5	7
High school graduate	42	44	37	38
Some college	36	26	26	29
College graduate	12	12	32	27
Work situation				
Employed fulltime	78	57*	72	67
Employed parttime	10	9	13	13
Unemployed	4	15	4	5
Other (not in labor force)	9	19	11	15
Student status				
Full-time student	11	10	10	13
Part-time student	10	6	11	10
Not enrolled	80	84	79	77
Income				
Under $7,000	16	44*	20	32*
$7,000 to $14,999	24	22	16	21
$15,000 to $29,999	50	28	40	32
$30,000 or more	11	6	24	15

* $p < 0.05$ For contingency table chi-square test of statistical independence (sample characteristic × mode of interview).

tics were entered as control variables in multiple classification analyses (MCA).

Preliminary analyses indicated significant race by mode of interview interactions for both alcohol and marijuana. Therefore the decision was made to analyze mode effects for blacks and whites separately. Table 2 displays the unweighted drug use estimates by mode of interview for blacks, whites, and both groups combined. In addition, the MCA net differences between interview modes controlling for sex, age, employment status, income, education, student status, and marital status are displayed.

Tobacco. Tobacco use is the least threatening survey topic of the four substances and should have been the least susceptible to interview mode effects. This prediction is supported by the data. There are no significant mode effects for blacks or whites on the five measures of smoking. For blacks, the smoking estimates are consistently lower in the telephone survey. However, the net differences by mode are greatly reduced when the demographic controls are introduced into the analysis. For whites, the estimates for recent (past year) and current (past 30 days) smoking are identical without demographic controls. These findings suggest that telephone surveys, corrected for demographic biases, can produce estimates of tobacco use comparable to face-to-face surveys for both black and white populations.

Alcohol. Items concerning alcohol use and abuse are more sensitive than smoking, therefore these questions should be more reactive to interview mode effects. Consistent with this prediction, there were substantial mode effects for drinking after controlling for demographic characteristics of the samples. The telephone survey yielded lower estimates for all five drinking measures. The results are particularly striking for drinking habits of blacks over the 30 days before the interview. The number of drinking days was over twice as high in the personal than in the telephone survey; estimates of total

Table 2. Drug use estimates by race and mode of interview

	Black			White		
	Telephone	Personal	Difference[a]	Telephone	Personal	Difference[a]
Tobacco (cigarette smoking)						
Smoked 5 packs or more lifetime (%)	42	50	−3	54	51	+2
Smoked in last 12 months (%)	39	46	0	42	42	0
Smoked in last 30 days (%)	36	44	0	36	36	−2
Number cigarettes smoked per day	3.9	4.6	−0.1	5.8	6.1	−0.4
Number of years smoked daily	3.5	3.7	−.1	4.1	3.9	+.1
Alcohol						
Drank in last 30 days (%)	55	64	−5	79	83	−5*
No. of drinking days in last 30 days	1.7	4.2	−2.5**	5.4	5.4	−.4
Total drinks in last 30 days	5	17	−13**	18	18	−1
No. of days had 5 or more drinks	0.2	1.5	−1.6**	1.2	1.4	−.2
Drunk once or more last year (%)	25	33	−5	52	56	−5*
Marijuana						
Used once or more in life (%)	49	64	−16*	66	67	−2
Used last 12 months (%)	16	29	−6	23	25	−1
Used last 30 days (%)	8	21	−11*	12	13	0
Used 10 or more times in life (%)	21	34	−13*	39	38	+1
No. of days used in last 30 days	0.7	2	−1.3	0.9	1.2	−.2
No. of joints a day in last 30 days	.1	.7	−.7*	.3	.2	+.1
Cocaine						
Used once or more in life (%)	21	24	−1	34	29	+4
Used last 12 months (%)	12	13	+1	10	14	−2
Used last 30 days (%)	4	7	−2	5	6	−2
Used 10 or more times in life (%)	9	7	0	17	11	+6**
No. of days used in last 30 days	.4	.3	+.2	.3	.2	0

[a]Net difference by interview mode (telephone-personal) in multiple classification analyses controlling for sex, age, employment, income, education, student status, and marital status.
NOTE: *** $p < 0.01$, ** $p < 0.05$, * $p = 0.06$

drinks for the month and the number of days respondents had five or more drinks were substantially higher in the personal interview.

For whites, the proportion of respondents who drank in the last 30 days and the proportion who admitted getting drunk at least once in the past 12 months were significantly lower in the telephone interview, although the magnitude of the differences was small (about 4 percentage points). The black differences on these variables, although of the same net magnitude, were not significant. On the whole, the telephone did a much worse job of estimating habits of blacks than of whites. It is clear that uncritical acceptance of the telephone estimates would lead to underestimation of black alcohol use and to overestimation of black-white differences in drinking.

Marijuana. The race by mode interaction is most evident in reports of marijuana use. For blacks the marijuana results followed the pattern for alcohol. The telephone survey furnished consistently lower net estimates of blacks' lifetime and recent marijuana use than did the face-to-face survey. The number of current users was much higher in the personal (19 percent) than in the telephone interview (8 percent); the number of lifetime users was about 30 percent higher in the personal mode. For blacks the telephone mode significantly underestimated both the days and the amount of recent marijuana use.

For whites the remarkable feature of the marijuana estimates is the extent of similarity between the two modes. The net differences between modes were zero or near zero in all six measures of marijuana use. For this statewide sample, then, there would be no bias in marijuana use estimates for whites associated with a shift from in-person to telephone interviewing. The prediction of greater social desirability effects for telephone than for face-to-face surveys was supported by the data for black but not for white marijuana use.

Cocaine. The pattern of cocaine results is similar to alcohol and marijuana, however the mode differences are not significant. The telephone survey furnished slightly lower estimates of blacks' lifetime, last year, and current cocaine use than did the face-to-face survey. For whites, lifetime use estimates were higher by telephone, whereas last year and current use estimates were lower for the telephone than in-person mode. The lone significant main effect was for whites: the number who admitted using cocaine more than 10 times was higher in the telephone survey (17 to 10 percent). Overall, the telephone mode did not produce a significant downward bias in cocaine use estimates for either group. Although expected to be as threatening as questions about marijuana, the most notable finding for cocaine is the near absence of significant mode effects for both blacks and whites.

Question order effects may have contributed to the

lack of significant cocaine findings. In both surveys the cocaine sequence followed items on marijuana use, and previous research has shown that nearly all cocaine users also use marijuana (Yamaguchi & Kandel, 1984; Lazarro & associates, 1988). Virtually all admitted cocaine users in the surveys described here admitted lifetime marijuana use as well. Therefore, before encountering the cocaine items, cocaine users had already been asked to report some illicit drug use, and nearly 100 percent had done so. Admitting marijuana use may have decreased the likelihood that use of another illicit drug, cocaine, would be denied. Thus, mode of interview effects may be the strongest when asking for the first admission of illicit or undesirable behavior during the interview.

Race and Interview Mode Effects

Our results suggest that for whites the telephone survey yielded drug use estimates comparable to the face-to-face survey. Where telephone estimates for whites were significantly lower than in person—for the current use of alcohol and drunkenness—the percentage differences between modes were small. The RDD sample also reproduced the face-to-face demographic profile for whites. The dearth of mode effects for whites is especially noteworthy given the large sample sizes (over 1,300 cases) and therefore relatively high power to detect differences.

The conclusions for blacks are very different. The telephone survey produced a significantly higher SES profile for blacks than did the personal survey. Further, even after controlling for socioeconomic status and demographic characteristics, the telephone survey furnished substantially lower estimates of blacks' current alcohol consumption and marijuana use compared to the face-to-face estimates. Reliance on the telephone survey alone would result in potentially inappropriate conclusions concerning racial differences in alcohol and marijuana use.

Why would blacks display more sensitivity to mode-of-interview effects than whites? Mensch and Kandel (1988) reported that minorities in general are more prone than whites toward socially desirable responses in drug use surveys, regardless of interview characteristics. It is possible that blacks and other minorities are more reactive than whites to variations in survey methodology, especially in surveys of illicit or undesirable behavior.

Although this study cannot determine the causes of the interview mode effects described above, sample characteristics alone do not seem to account for them. As described above, controlling for demographic characteristics does not attenuate mode effects for alcohol and marijuana use by blacks, or for alcohol use by whites. To examine the impact on reported drug use of RDDs nontelephone exclusion, analyses were re-run dropping the nontelephone households from the face-to-face sample. None of the significant mode effects disappeared when controlling for telephone status, although the size of the net differentials was moderated to a slight degree. Controls for sampling bias did not equalize the drug use estimates for the two survey modes.

Table 3. Weighted estimates of drug use by race and mode of interview

	Black		White	
	Telephone	Personal	Telephone	Personal
Tobacco (cigarette smoking)				
Smoked 5 packs or more in life (%)	42	47	51	50
Smoked in last 12 months (%)	42	41	41	42
Smoked in last 30 days (%)	39	38	35	37
Number cigarettes smoked per day	4.4	4.2	5.5	6.1
Number of years smoked daily	3.5	3.4	3.7	3.7
Alcohol				
Drank in last 30 days (%)	52	62	77	84
No. of drinking days in last 30 days	1.8	4.9	5.3	5.6
Total drinks in last 30 days	4	19	18	19
Number of days had 5 or more drinks	0.3	1.9	1.3	1.5
Drunk once or more last year	26	29	53	59
Marijuana				
Used once or more in life (%)	51	64	66	66
Used in last 12 months (%)	16	25	25	26
Used in last 30 days (%)	7	19	13	13
Used 10 times or more in life (%)	17	34	39	39
Number of days used in last 30 days	.6	1.7	1.0	1.1
Number of joints a day in last 30 days	.1	0.6	0.3	0.2
Cocaine				
Used once or more lifetime (%)	19	25	33	30
Used in last 12 months (%)	9	12	12	14
Used in last 30 days (%)	3	7	5	6
Used 10 or more times in life (%)	6	9	17	11
No. of days used in last 30 days	.4	.3	.3	.2

Appropriate weighting of cases within each sample also did not alter the results (Table 3). Case weights reflected adjustments for household selection probabilities (including number of telephone lines), differential nonresponse, and for census profiles of New Jersey's age, race, and sex composition. Using weighted data, the mode differences for blacks were slightly greater, and an additional variable (number of days of marijuana use) showed significant mode effects in the predicted direction. There were no differences in the results for whites when weighted data were used.

Apart from sampling issues, characteristics of the interview situation itself may play a role in altering responses to threatening survey items. Mode differences in provision of privacy and anonymity are central issues in surveys concerning illegal behavior or embarrassing topics. The self-administered answer sheets for drug use reporting in the personal mode, where interviewers read the questions but do not see the answers, may give respondents a greater degree of anonymity than the telephone interview, where the unseen interviewer hears the report directly.

Applicability to Other Surveys

In surveys concerned with sensitive or embarrassing topics, such as sexual behavior, sexually transmitted diseases, and illicit or other socially undesirable behavior, the biases introduced by RDD sampling may be highly correlated with the behavior the survey is trying to estimate. It is imperative in sensitive surveys to correct for the upward SES bias of random digit dialing and for the nontelephone household exclusion. In addition, it is important to recognize that bias introduced by socially desirable responding may vary by mode of interview and that bias associated with mode of interview may vary among racial and ethnic groups. Surveys whose purpose is to estimate racial and ethnic group differences in sensitive or illegal behavior may substantially overestimate such differences by switching from in person to telephone survey modes. Although the telephone survey is attractive for cost, quality control, and efficiency, researchers interested in sensitive topics should not assume that the telephone mode will furnish data fully comparable to the face-to-face interview.

References

Freeman, H. , Kiecolt, K., Nicholls, W., & associate. (1982). Telephone sampling bias in surveying disability. *Public Opinion Quarterly,* 46, 392-407.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Academic Press.

Johnston, L., & O'Malley, P. (1985). Issues of validity and population coverage in student surveys of drug use.In B. Rouse, N. Kozel, & L. Richards (Eds.). *Self report methods of estimating drug use: Meeting current challenges to validity.* (NIDA Research Monograph No.57). Washington, DC: Public Health Service, Department of Health and Human Services.

Lazarro, C., LoSciuto, L., Porcellini, L., & associate. (1988). Assessing the validity of a sequential drug use pattern using the 1985 National Household Survey on Drug Abuse. In *Proceedings of the American Statistical Association,* New Orleans.

Mensch, B. S., & Kandel, D. B. (1988). Underreporting substance use in a national longitudinal youth cohort: Individual and interviewer effects. *Public Opinion Quarterly,* 52, 100-124.

Sudman, S. (1976). Sample surveys. *Annual Review of Sociology,* 2, 107- 120.

Sudman, S., & Bradburn, N. (1974). *Response effects in surveys: A review and synthesis.* Chicago: Aldine Publishing Co.

Tull, D. S., & Albaum, G. (1977). Bias in random digit dialed surveys. *Public Opinion Quarterly,* 41, 389-395.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association,* 73, 40-46.

Yamaguchi, K., & Kandel, D. (1984). Patterns of drug use from adolescence to young adulthood: Sequences of progression. *American Journal of Public Health,* 74, 668-682.

DISCUSSION PAPER

Measuring Behavior Related to Risk of Acquired Immune Deficiency Syndrome

Seymour Sudman

Introduction

Although acquired immune deficiency syndrome (AIDS) is the major public health problem of our time and has generated a flood of research studies using survey methods, of which these papers are examples, there does not appear to be anything in these studies that significantly changes what is currently known about how to conduct surveys on sensitive issues. This may not be an exciting conclusion, but it is comforting to those who appreciate confirmation of earlier results.

To summarize these papers and put them into context, let me briefly touch on the following issues:

1. What are the political consequences of collecting survey data on sensitive topics?
2. Will people answer the questions?
3. The quality of the answers as a function of:
   (a) question wording
   (b) method of administration
   (c) interviewers

The Political Consequences of Collecting Sensitive Data

The paper by Laumann and associates illustrates quite well what can happen to any study that collects data on a sensitive policy issue. Ideologues will oppose the data collection on principle because it may produce data harmful to their position. They will use arguments about the uselessness of the data, invasion of privacy and problems with the research design or anything else they can think of to abort the study. Opposition can come from politicians as in this case or as in the case of the seroprevalence pilot study proposed for Washington, DC that Weeks discussed earlier in this volume, or it can come from the groups being studied who then argue that they are being studied to death and no action has been taken.

Those who believe that sensible policy must be based on information will do all they can to combat the censorship mentality that says "my mind is made up—don't confuse me with the facts." Specifically, the study that Laumann and Gagnon have described is worthy of support from AIDS policy makers. The model examines life course and network considerations and seems both theoretically sound and of practical value to policy makers. Earlier in this book, several discussants pointed out the importance of developing a careful conceptual framework before designing a questionnaire. The study by Laumann and colleagues is an outstanding example of a well-developed conceptual framework.

It should be noted that even extensive advance interaction with interest groups will not guarantee smooth sailing for a survey on a controversial issue. Certainly, for both the Health and Sexual Behavior and the Household Seroprevalence surveys significant efforts were made to interact with a broad range of interested participants. As in the story of Sleeping Beauty, one uninvited guest may be able to cast a curse on the project.

If opposition surfaces, all is not lost. It is sometimes possible to change people's minds or to ignore the opposition. Modifications in the study design that do not cripple it may work. In pilot studies, changing the location is possible and in the extreme case, finding a different source of funding.

Most of us would prefer to do our research and leave the controversies to others. Unfortunately, when doing research on sensitive issues we sometimes need to fight for the right to do something we believe in.

Will Respondents Answer Sensitive Questions?

Many nonresearchers and perhaps even a few researchers still worry that people will be unwilling to

Seymour Sudman is with the Survey Research Laboratory, University of Illinois, Urbana.

answer questions about sensitive issues such as sexual activity, drug use, and AIDS. All previous evidence as well as that presented in McQueen's paper refutes this concern. In central Scotland, certainly not a hotbed of sexual permissiveness, refusal rates on AIDS-related questions ranged from 2 to 5 percent. This is quite close to the kinds of refusal rates found here in the United States on surveys of sensitive topics, although for some questions refusal rates might approach 10 percent. As we all know, refusal rates are highest on that most sensitive of all questions—income.

Does even asking such questions invade the privacy of respondents? Most of us think not, especially when respondents are told in advance that there will be questions on sensitive topics and that they are at liberty to refuse to answer any question that they find too personal. When such information is provided, it reassures respondents and, if anything, reduces rather than increases refusal. It should be noted that informed consent does not require describing all the questions before the interview begins. It is better to do this during the interview as different topics arise. Callahan made the useful distinction between consent to surgery which must be given before the operation starts, and consent to an interview that can be withdrawn at any time.

Question Wording

It is disappointing that in the section on Measuring Behavior Related to Risk of AIDS there was virtually no discussion of questionnaire design issues and particularly of the possible contributions that might be made by recent advances in the applications of cognitive psychology to survey methods. The papers by McQueen and Aquilino and most of Cain's paper deal with methods of administration. The rest of the Cain paper and the Campbell paper deal with interviewers. These will be discussed next, but first some comments on the paper by Laumann and co-workers that does discuss questionnaire design issues.

These authors are relatively optimistic about being able to get high quality data on sexual activities, but the reasons for their optimism are not well spelled out. It is probably true that most respondents will find the tasks easy because they have a single or small number of partners. Unfortunately, the respondents of highest interest and at highest risk would seem to have the greatest difficulty in reporting accurately. Specifically, frequent, casual, same-sex encounters will be difficult to remember on an individual basis and respondents will probably adopt an estimation heuristic. The estimate may or may not be accurate depending on the regularity of the event and the heuristic used. Sometimes it is not necessary to get precise estimates of activity. In this case, however, it is likely that the epidemiologic models that have been developed are probably highly sensitive to measurement errors.

Several papers in this volume deal with methods for evaluating questions. Currently we favor the procedure discussed by Royston and used at the National Center for Health Statistics, the Census Bureau, and the Bureau of Labor Statistics, that is, the cognitive laboratory. This involves the use of think-aloud methods for determining what the respondent thinks the questioner wants and how the respondent goes about answering. Perhaps this was done in developing the questionnaire. If not, it should be.

It is well known that memory decays with time. Thus, the paper would have been strengthened by more discussion of the testing that preceded the decision to use a 5-year period when studying sexual relationships. For most people this might be easy, but again we should worry about those at highest risk.

As a final comment on wording, how did the research team develop the phrases describing sexual activity that are to be placed on cards? Bradburn and I found that people using their own words reported better than those who were asked to use standard words for threatening activities.

The first section in this volume discusses alternative ways of questionnaire evaluation, and there is general agreement that the use of multiple methods is superior to using any single method. The use of multiple methods for testing AIDS questionnaires is especially appropriate given the special problems they present.

Method of Administration

If one is asked to make a general remark about method of administration, based on the papers in this volume as well as much earlier work, the remark would be that method of administration usually has little or no effect on quality of response. Although one can postulate that there should be differences caused by different levels of privacy and different amounts of time spent on the task, for most variables one sees no differences. Essentially, that is also what McQueen found in his careful studies in Scotland.

It at first appears as if Aquilino has found differences between telephone and self-administered forms where an interviewer is present, but this may well be caused by sample or question wording differences. The standard procedure for separating sample differences from methods effects is to ask respondents in the personal interview whether they have a telephone and then to compare only telephone households. This is what McQueen did. Aquilino reports that a similar analysis was run on the New Jersey data, but the results were not changed. It is not clear whether this analysis was run for blacks and whites separately.

What makes one wonder is that no or very small effects are seen for whites and very large differences for blacks. It is not at all clear why this should be and almost certainly was not hypothesized in advance. This might be a real effect, but one would want to see additional confirmation.

Other possible explanations may be that the 20 percent difference in cooperation rates resulted in sample differences that were not accounted for by the weighting. There may also have been differences in context effects because the telephone interview was shortened. It is also unclear why there are method differences for blacks for

alcohol and marijuana, but not for cigarette smoking and cocaine use. Some of the observed differences may simply be caused by sampling variability because the sample sizes in the two black treatment groups are small.

Additional studies of methods of administration effects are proposed in the Cain paper. Given the small sample sizes, it is unlikely that anything very interesting will occur. One final comment on telephone and face-to-face comparisons. When doing these, it is important to distinguish between mode effects and those that might be caused by differences in the experience, training, and quality of interviewers. Thus, any organization in making comparisons between its standard mode of data collection and an alternative method is likely to demonstrate superiority for its standard mode. This has sometimes been observed when Federal agencies test their standard face-to-face interviews using highly trained and experienced interviewers against other methods using new interviewers.

There is a special concern about the test of the effects of differences in method of administration between face-to-face and self-administered. The design calls for first having the interviewer answer the questions and then having the respondent answer them again on a self-administered form. Although Kinsey used methods of asking the same question several times, most respondents resent being asked the same questions more than once. The typical response is " I told you that already!"

Interviewers

The Cain and Campbell papers discuss interviewer training and effects. First, a quick comment on interviewer matching. The literature generally shows effects when the visible interviewer characteristic is germane to the topic being studied. I would predict that there would be gender effects on reports of sexual activity. There is no reason, however, to expect race effects. Even the gender effects may be small. It has been noted that interviewer effects are much stronger on attitudinal than on behavioral questions.

The Campbell paper has no experimental design aspects, but is a sensible discussion of typical issues in training interviewers. Although interviewer selection is not discussed, Bradburn and I found that a useful method for eliminating interviewers who would be uncomfortable with the subject matter was to observe their behavior during training. Those interviewers who were uncomfortable during the training were not asked to interview. Those who remained were extremely professional as indicated by the taped interviews we obtained. In my view, it is this professional tone that the interviewer sets that prevents this from becoming an erotic encounter.

The decision to use experienced field managers to conduct the training and subject-matter specialists as consultants for this study is exactly what one does for any study. What is innovative is the use of reading material and oral sessions to make interviewers comfortable with the sexual words used. As suggested earlier, those who cannot handle the training will be unable to do the interviews. What I found most interesting about Campbell's discussion was not the differences from a typical survey, but the similarities. Indeed, we do not need to abandon or change our standard survey methods when dealing with AIDS and other sensitive issues. We simply need to use them better.

Measuring Behavior Related to Risk of AIDS

Marcie L. Cynamon and Jennie J. Kronenfeld, Recorders, and
Edward O. Laumann, Chair

Much of the discussion centered on whether differences found in reporting of specific high-risk behaviors such as drug use could be attributed to the differences in mode (telephone versus face-to-face interviews with a self-administered add-on) or to nonresponse bias (see Aquilino, in this volume). Aquilino reported that an examination of his data after removing nontelephone households from the face-to-face interview data revealed no significant changes in the findings for whites or blacks.

Several possible reasons were suggested as contributing to differences in respondent reporting of drug use. One was level of social desirability. A second was the use of self-administered answer sheets by respondents during face-to-face interviews—a procedure that offered a certain amount of privacy. A third factor suggested was racial matching of interviewers to respondents, although the Aquilino data did not support an effect from race or sex matching.

The issue of age matching in surveys of sensitive behaviors was addressed and mixed results were reported. Campbell reported on the results of focus groups conducted by the National Opinion Research Center. Focus group facilitators and participants were matched on age, sex, race, and ethnicity. When shown pictures of potential interviewers, all groups invariably chose the picture of an older white female as the interviewer with whom they would feel most comfortable. Again the issue of social desirability was raised. Other questions included whether individuals of the same sex, race, and approximate age would be comfortable reporting accurately their levels of activity (or inactivity) and whether social distance would increase comfort and thus cooperation and accuracy of reporting. Due to the constraints of the sample size, age matching is not planned as one of the methodological tests for the National Health and Sexual Behavior (NHSB) pretest but race and sex matching will be used (see Campbell, in this volume).

One of the problems likely to be encountered in either survey mode—especially in studies of sensitive behaviors—is that many of the individuals at higher risk of practicing behaviors of interest are likely to be missed in studies of households. These individuals are likely to be unconnected to households in the traditional sense and extraordinary efforts may be needed to find them or to identify them with a household. Extensive probing as to whether there are occasional household occupants who have no other usual place of residence may be necessary to reduce noncoverage bias.

It was pointed out that the surveys discussed in this session were not designed to include respondents on the fringe of or outside the framework of households. The increasing importance of reaching these individuals, however, led to discussion of potential methods. One suggestion was to adapt some tactics being developed to measure the homeless population, such as stationing interviewers in parks or other areas of congregation. Another suggestion was the use of a modified capture-recapture method to estimate the size of the population at risk (Spencer, 1989).

Both Campbell and McQueen stressed the need to debrief interviewers extensively to assist in the evaluation and revision of the survey questionnaire. McQueen—who has developed a fifth version of his interview schedule on AIDS-related issues—reported that by keeping an undisturbed core of questions, he has been able to compare with previous iterations the effects of question-and-response order.

To address language sensitivity, Campbell pointed out that the NHSB plans to use commonly accepted words to refer to sexual activities. Interviewers will be trained to understand and be comfortable with sexual terms but will not use them because it is deemed undesirable for interviewers to pretend to mimic respondents. Steps have been taken to make reporting more comfortable for the respondent. For example, a flashcard will be used so that the respondent need never actually use sexual

Marcie L. Cynamon is with the Division of Health Interview Statistics, National Center for Health Statistics. Jennie J. Kronenfeld is with the Department of Health Administration, University of South Carolina. Edward O. Laumann is with the Division of Social Sciences, University of Chicago.

terms. It was suggested that interviewers record the existence of a different language base at the beginning of the interview for later analysis.

Questions were also raised about the issue of respondent recall. A 5-year sexual history is being used in the NHSB for modeling purposes; but whether respondents have sufficient recall to make this a worthwhile time frame is not known. Gagnon reported on successful collection of sexual history data by emphasizing periods of cohabitational relationships. With this method, detailed questions are asked about partner experience using the period of cohabitation as the memory anchor. Sexual activity before the formation and after the dissolution of relationships can then be recalled in context. These approaches—such as focusing on the sexual experience rather than the sexual act and wording questions so that neither heterosexuality nor homosexuality is implied—are expected to improve both recall and reporting. It was also pointed out that recent data from the General Social Survey indicate that the majority of the population has had zero or one partner in the past year, a finding that suggests recall will be an issue for a smaller segment of the population than previously expected.

Also discussed were the closely related issues of cost and data quality common to all surveys. The point was made that some level of error is inherent in every survey and a delicate balance must be achieved between reduction of error and practical economic considerations. Thus when quality is assessed by examining sources of error, the study's basic purpose must be considered. A survey designed to estimate the proportion of people at risk for HIV infection must expend more resources to control nonresponse bias, for example, than a survey designed to describe levels of knowledge about AIDS.

Participants also addressed the issue of privacy and its effect on both response and nonresponse bias. This broad area ranges from protecting from subpoena the identity of respondents engaged in longitudinal studies to convincing respondents of the confidential or anonymous nature of the data collection. Horvitz discussed the concerns expressed by minority populations during the planning of the National Household Seroprevalence Survey. These concerns led to the decision to design the study so that the responses were anonymous rather than confidential.

The point was made that a respondent may not know the differences between confidentiality and anonymity and, if so, that it may affect the decision to participate in a survey. Whether data are more accurately reported when one safeguard is promised over the other or how this affects response bias is also unclear. McQueen observed that people perceive a higher level of privacy in telephone interviews than in face-to-face interviews. Cain reported that observations from field work indicate respondents have greater confidence in the confidentiality of data when the interviewer records data by computer rather than with paper and pencil. While some researchers reported that they thought some telephone respondents felt more assurance of confidentiality in a telephone interview involving computers, McQueen found the opposite to be the case in Scotland.

Considerations for Further Research

Unresolved questions that warrant further research include the following:

1. What effect does mode of data collection (telephone, personal interview, self-administration) have on reporting sensitive AIDS-related behavior?
2. What is the potential of direct computer entry for improving reporting?
3. Can optimal interviewer characteristics be identified and, if so, what are they?
4. How does the variability of nonresponse bias by mode affect survey estimates of sensitive behavior rates or other survey findings?
5. How can nonmembers (or fringe members) of households best be captured for surveys that measure the population at risk for AIDS?

Reference

Spencer, Bruce. (1989). On the accuracy of estimates of numbers of intravenous drug users. In C. Turner, H. Miller, & L. Moses (Eds.). *AIDS, Sexual Behavior, and Intravenous Drug Use* (pp. 429-446). Washington, DC: National Academy of Sciences.

CONCLUDING DISCUSSION PAPER

A Total Survey Error Approach to AIDS-Related Survey Research

Robert M. Groves

Introduction

Applications of survey research in new fields often start with solely descriptive studies. The move to population surveys often arises because of perceived inadequacies of small-scale experiments or convenience-sample observational measurements. As basic questions about the prevalence of various characteristics come to be answered, questions of cause are addressed. Relationships between one variable and another are the focus of data collection efforts. These causal relationships are addressed by fitting statistical models to the data, which are consistent with certain causal explanations of the focus of the research.

Descriptive and analytic uses of survey research often bring with them different concepts of error. Descriptive uses of survey data have in mind a finite population of limited temporal and spatial extents. Analytic modeling often uses the survey data to test theories applicable to populations of infinite size—ones existing in the past, present, and future. These two uses of survey data also tend to focus on different error sources in the data (Deming, 1953; Anderson & Mantel, 1983).

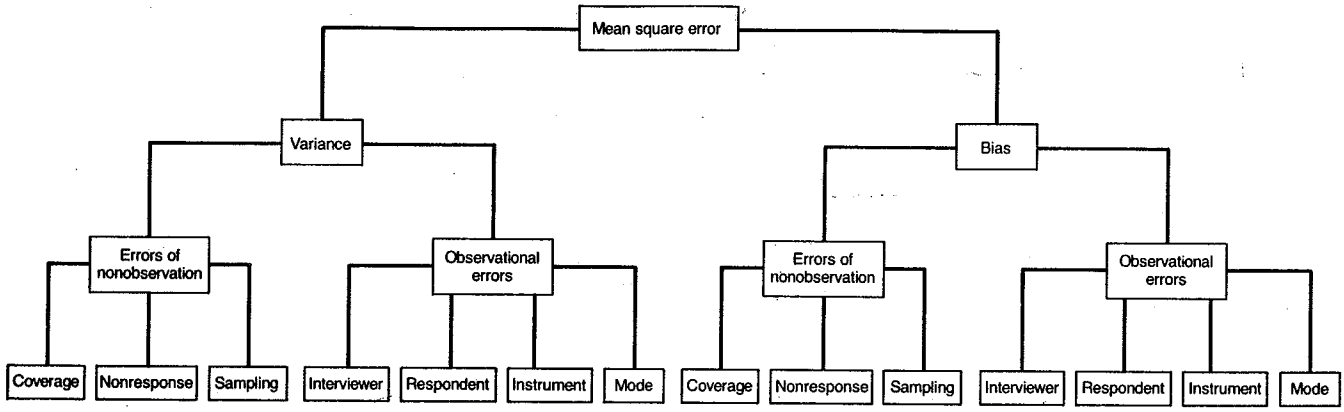
Within each of these groups lies another division—those seeking to reduce or eliminate errors and those seeking to measure the magnitude of the errors remaining in the data. Although these groups need not be mutually exclusive, the existing overlap is a small one. These two groups are referred to as "reducers" and "measurers" (Groves, 1987, 1989). Often, periodic methodological studies are performed to measure the magnitude of a particular error source under some survey condition (for example, telescoping error with different reference periods for doctor visits), then these studies are used for several years to guide efforts to reduce the particular error source (for example, using reference periods of shorter duration). Although one could easily argue that every survey should attempt to measure all error sources, there are financial burdens of measuring errors as well as ones for reducing errors (for example, costs of interpenetrated interviewer assignments to measure interviewer variance are nontrivial). However, at the beginning of a field's use of the survey method researchers are particularly uninformed about likely errors associated with different design options. This seems to be the situation in acquired immunodeficiency syndrome (AIDS)-related surveys, which clearly affects the ability to reduce errors of various sorts.

This paper attempts to review research methodological issues relevant to AIDS-related surveys in the context of a "total survey error" perspective. Figure 1 presents the conceptual framework of total survey error. Both fixed (biases) and variable (variance) errors are permitted within the approach. Sources of observational and nonobservational errors are acknowledged. The nonobservational errors arise from coverage error, the failure of a sampling frame to include some members of the population; nonresponse errors, the failure to obtain data from some sample members; and sampling errors, the failure to measure all persons in the population. Coverage error and nonresponse error are typically ignored in most approaches to error among modelers, although notions of external validity in psychometrics (Cook & Campbell, 1979) and selection bias modeling approaches in econometrics (Heckman, 1979) have similar aspects. Errors of observation can be divided into sources of the interviewer, the respondent, the questionnaire, and the mode of data collection. As with psychometrics, notions of measurement error variance are permitted. They include notions of "simple response variance" arising from an indeterminacy in the response formation process of the respondent (Fellegi, 1964). They also include effects of interviewers on responses, which are seen to vary with different interviewers who might conduct the survey. Inherent in the notion of var-

Robert M. Groves is with the Survey Research Center, The University of Michigan, Ann Arbor.

This work was supported by Grant No. MH43564 from the National Institute of Mental Health.

Figure 1. Conceptual framework of total survey error



iable errors is the assumption of replication of the survey under the same essential conditions but with the assigned unit altered (for example, a different interviewer, a different way of asking the question).

Absent from the total survey error perspective (which belies its title) is much attention to the question as a source of measurement error variance. Psychometric approaches to measurement error (see Figure 2) posit that each question measuring a particular concept is one of a large number of such questions, each of which attempts to tap the same unobservable trait of a person. Following this notion, reducing measurement error variance by using multiple questions for each concept is desirable. In addition, the researcher can measure error variance by variation of responses across the questions is possible. However, psychometric approaches usually fail to incorporate formally the influence of the interviewer and the mode of data collection.

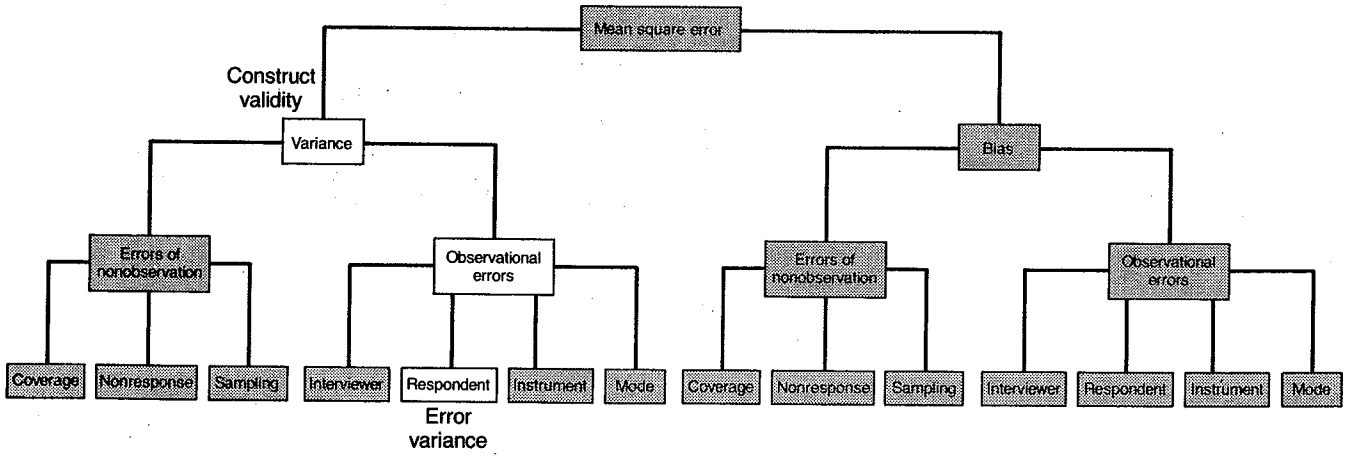
In addition to different concepts of error, the languages of error in the two approaches differ. Instead of the "bias" and "variance," notions of "reliability" and "validity" abound in the psychometric literature. The translation between the concepts is not a straightforward one (Groves, 1989). The papers in this book on AIDS-related surveys exhibit a diversity of design concerns, languages of errors, descriptive versus analytic goals, and concern about measurability of error. This probably flows from differences between descriptive and analytic goals of the studies and between researchers trained in the survey statistics versus psychometric measurement traditions. This is to be expected in a field that has newly formed and is interdisciplinary in nature.

A review of the papers from a total survey error viewpoint suggests several lines of investigation that might be fruitful. Rather than presenting a comprehensive examination, this paper comments on several broad themes.

Errors of Nonobservation

Although the incidence of AIDS-related disorders increases daily, the problem still inflicts only a small por-

Figure 2. Conceptual framework of errors in psychometric measurement theory



Shaded concepts are not central to viewpoint of psychometrics for individual measurement.

tion of the population. The rarity of many of the characteristics and activities measured in AIDS-related surveys complicates the application to them of past survey methodological research findings. For example, much of the survey methodological literature addressing nonresponse error focuses on nonresponse rates (Goyder, 1988). This literature is dominated by efforts to increase the response rate. For a linear survey statistic, nonresponse error is a multiplicative function of the nonresponse rate and the difference in statistics for nonrespondents and respondents. Merely increasing the response rate does not assure reduction of nonresponse error. The paper by Weeks et al. in this volume permits a simple illustration of this point. Assume, as they do in Table 1, a "natural" level of cooperation among human immunodeficiency virus (HIV) negatives that is approximately three times greater than that among HIV positives (actually 0.96:0.30 = 3.2). One way that ratio can manifest itself is by each effort to increase response rate being roughly three times more successful among the HIV negatives than the HIV positives. That is, each increase of the overall response rate might be composed of disproportionately more HIV negatives, not just because they are much more plentiful in the sample, but because they are more cooperative. Assume that roughly 0.01 of the full sample is HIV positive and the remainder HIV negative.

Figure 3 shows how, under these assumptions, the response rates of the two groups (HIV positive and negative) increase as the overall study response rate increases. Note that the response rates for both groups increase as the survey proceeds to achieve a higher response rate, but that the rate for the HIV positives is always lower than that of the HIV negatives. Under these assumptions, only the last efforts of the field period, the most extreme measures of persuasion or locating efforts, are successful at bringing many of the HIV positives into the sample. For example, at an overall 0.90 response rate for the study, the response rate for the HIV positive group is only around 0.30.

Figure 3. Response rate for HIV− and HIV+ under Weeks model of participation

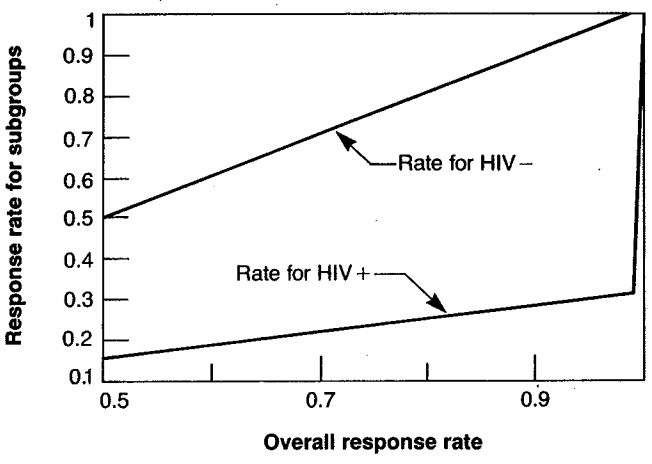


Figure 4. Ratio of bias to true value for proportion HIV+ given Weeks model of participation

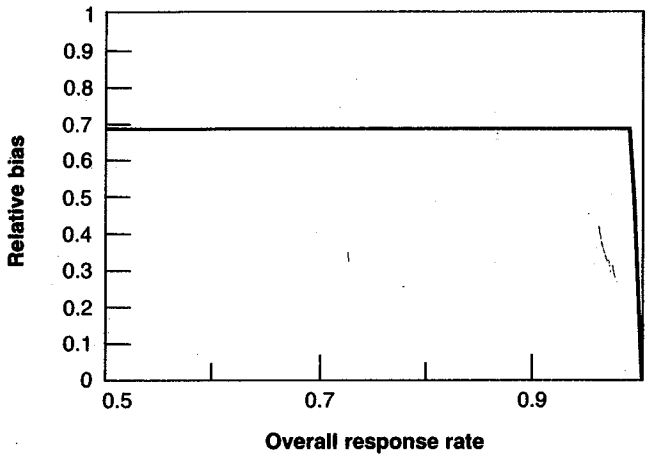


Figure 4 illustrates the nonresponse error properties of such a situation. It plots the absolute value of the ratio of the bias of the sample estimate to the true value of the proportion positive (0.01). The x-axis of the graph is the overall study response rate. The graph shows that there is no change in the bias of the sample estimator as the response rate increases until the very end of the effort, when the extreme final measures are taken and the HIV positives are brought into the respondent group. This is an example of how nonresponse rate is unrelated to nonresponse error except within a very limited range of response rates (ones incidentally unlikely to be obtained).

Is this a proper model for the relationship between nonresponse rate and nonresponse error? Probably not. It suggests no change in the point estimate of the sample statistic as the number of calls increases. It is only one possible model of the survey participation process, which can produce a very high relative bias of the sample estimate, with very high overall response rates in the survey. The Capell and Schiller paper shows counter evidence on this point. Their data do not permit us to assess, however, how the change in values of point estimates with different callbacks relates to the nonresponse bias that remains at the end of the study.

The moral of this illustration is that the traditional beliefs among survey researchers about what constitutes an acceptable response rate may not be wisely transported to AIDS-related surveys. The danger arises from the rare population features of many estimates in such surveys. What should be done? The new field itself must estimate relationships between the proxy indicator of error and the error itself. This requires special studies, with unusual (and expensive) design components.

On another matter related to nonresponse, the relationship between reluctance to participate and true values on the survey variables might be of particular importance to AIDS-related surveys. Rates of infection are highest in intravenous (IV) drug users and those engaging in specific sexual activities with a wide variety of

partners. The sequelae of IV drug use are such that it is likely that these persons are highly transient, of relatively low income, and disproportionately not linked to households. The traditional survey technique of using household population sampling frames clearly involves larger noncoverage and nonresponse errors for AIDS-related surveys than for most household surveys. For that reason the Bradford and Keeter paper on interviewing male prostitutes and a variety of network reporting schemes might have more attractive cost-error properties in this field than in other fields.

On the nonresponse side, the paper by Weeks et al. describes unusual efforts to reduce nonresponse (unfortunately, without a measurement of its effectiveness) through use of videotape presentations shown to the sample household. Other papers describe unusual callback and refusal persuasion efforts. There is growing evidence (Steeh, 1981; DeMaio & associates, 1986; Groves, 1989) that the U.S. nonresponse is increasing. Unfortunately, AIDS-related surveys thus enter the picture in a nonbenign environment. This seems to be a widespread phenomenon in the Western World, exacerbated in some countries by public debates about the role of social measurement in the country (for example, West Germany, Sweden). There is little reason for optimism about the near term costs and benefits of continuing only attempts to reduce nonresponse. I have argued that proportionately more of data collection resources should be invested in understanding and adjusting for the inevitable nonresponse that is present in our efforts (Groves, 1989).

This argument stems from the observation that efforts to decrease nonresponse are both terribly expensive and unfruitful and that they are approaching the limits of ethical behavior on our part. Our current posture is an example of the "role differentiation" concept raised by Callahan in this volume. The societal ingredients that make surveys a straightforward system of knowledge acquisition may be passing, and we may have to adapt to these changes by increasing the sophistication of our adjustment schemes.

Every method of qualifying conclusions from a sample survey because of unit nonresponse implicitly or explicitly involves a model of survey participation. This model identifies the predictors of the likelihood of cooperation, given a request to participate. These models are implemented differently in use of discrete weight groups using data external to the survey, weighting by the reciprocal of estimated likelihood of response, or selection bias models. However, all of them might be described as implementations of models of survey participation.

One way to acquire data for the estimation of these models is through auxiliary data collection efforts in the survey. The Centers for Disease Control's National Seroprevalence Survey, has moved somewhat in this direction. The theoretical foundation of the work consists of concepts found to be influential on compliance in the literature of the social and cognitive psychologies of compliance and persuasion (Cialdini, 1984; Petty & Cacioppo, 1981). Indicators of those concepts are observed by interviewers for both respondents and nonrespondents (for example, number of counterarguments presented to the interviewer before final decision to accept or reject the request, various environmental conditions, distractions during the introduction, and so forth) Empirical predictive models of survey participation are constructed, given these indicators. These become the basis of adjustment models during survey analysis. Such a technology is in its infancy, and its merits and weaknesses are not yet understood. It has the attraction that it seems to be more consistent with ethical principles that have traditionally guided our work and with a more practical and cost efficient solution to the growing rejection of surveys by the populace.

Errors of Observation

Both cognitive and social psychological approaches to survey measurement error seem to be relevant to risk behavior regarding AIDS. Most of the papers addressing measurement error have landed on the bias side, examining implicitly systematic tendencies to underreport behaviors on sensitive topics. The social psychological side of this issue has been well investigated on cross-section populations (Bradburn & associates, 1979; DeMaio, 1984). That work was notably unsuccessful on one dimension—inference that social desirability effects could be reduced by some design change (question wording, interviewer behavior) stem solely from higher reporting rates of traits judged socially undesirable by the researcher.

Efforts at practical use of the Crowne-Marlowe scale appear to have failed for the most part (DeMaio, 1984). Despite the general fear of researcher attitudes determining measured direction of bias, in cross section work this technique does not seem to have misled the field too badly. With special populations, a different story may apply. Rare populations are often of interest because they exhibit subcultural traits different from those of the larger society. The act of needle sharing as a bonding ritual among IV drug users is such an example, and the social desirability influence on responses to sexual activity questions may vary in magnitude and direction across subgroups. Thus, the traditional criterion of "more is better" is particularly worrisome for risk behavior questions in subgroups.

There is another, related point for social desirability or threat effects on risk behavior measures. The social desirability effect as a source of measurement error is a classic example of errors completely correlated with the true value for the respondent. That is, only those respondents who truly have the undesirable trait are subject to the influence toward measurement error. (Note that social desirability effects also act on behaviors!) The elimination of the effects through design changes should only act on those subject to the effect. In rare traits we would expect, therefore, only small changes in point estimates with the elimination of social desirability bias. This might apply to many risk behaviors. This point is relevant because one of the ways that social desirability effects have been reduced in other substantive areas is the deliberate loading of the question (for example, "We know that many people weren't able to vote in the elec-

tion either because they were ill or couldn't get transportation to the polls, or for some other reason. How about you . . .?"). This loading in rare traits risks overloading.

Other interesting reports address biases introduced into survey data by interviewers or by the inflation of the variance of survey estimates because of different ways interviewers implement the questionnaire. The paper by Campbell and others and the Cain paper describe concerns motivated by these reports. Past studies focusing on bias have examined sociodemographic effects of interviewers (Schuman & Converse, 1971; Groves & Fultz, 1985) and interviewer behavior (Cannell & associates, 1981). The typical finding is that when the topic is relevant to the sociodemographic characteristic, higher quality data might be obtained by interviewers perceived by respondents to have similar traits as their own. Gender of the interviewer is an obvious target of concern in AIDS-related surveys. The theoretical puzzle however is that the sexes may induce different effects (for example, a single male overestimating sexual activity to a male interviewer but underestimating it to a female). Matching of respondent and interviewer attributes seems inappropriate here; rather, the randomization design of Cain seems more attractive. It is unfortunate that there are no plans to continue the randomized design past the pretest phase.

The interviewer expectations literature (Sudman & associates, 1977) and the interviewer variance literature (Groves & Magilavy, 1986) take a measurer viewpoint of error rather than a reducer viewpoint. Applying the results of that literature to AIDS-related surveys is not straightforward. The complication arises from the higher density of sensitive questions in these surveys. One track of research on this is the use of prequestionnaires and postquestionnaires administered to interviewers and interpenetrated assignment of interviewers to respondents. The questionnaire is low-cost in both face to face and telephone surveys. The interpenetration is a relatively straightforward and low cost alternative in the telephone mode (Stokes, 1986).

Relationships Among Errors

One unique contribution of the total survey error perspective is the attention to relationships among several error sources. Here attention is both to relative values of the two different error sources and to how attempting to decrease one may affect the value of another.

Several papers in this volume involve the use of telephone surveys in AIDS-related measurement. This could be considered an explicit tradeoff of coverage and measurement error. To my knowledge we have not yet measured on AIDS-related surveys the noncoverage error properties of the telephone. (The National Health Interview Survey AIDS supplement and the eventual National Opinion Research Center study on sexual behavior might be of use here.) The paper by Fleishman and others using client-based Management Information System data might be one approach to learning the proportion of high risk and HIV positive persons who do not have telephones. The work of Thornberry and Massey (1988) suggests that telephone subscription would be disproportionately low in these groups. In addition to the several studies comparing telephone and face to face interviewing on response distributions and missing data rates, it would be useful to measure simultaneously noncoverage error. This is clearly a good area to combine comparisons of costs and errors of alternative modes.

The discussion of race and gender matching of interviewer and respondent in the Cain work is a possible source of tradeoff of nonresponse and measurement error. To estimate the measurement errors associated with race and gender of interviewer might incur a higher nonresponse by the survey.

Finally, the use of sample designs that do not give known probabilities of selection to all members of the target population are attractive in AIDS-related surveys when a convenient, locatable group has many of the desired properties on risk behavior variables. This is a problem of the tradeoff of sampling error measurability, coverage error, and nonresponse. High response rates might be attainable with such a sample, but inference is limited by the restrictions on the sample.

Summary

AIDS-related surveys are pushing the methodology beyond its traditional limits. They pose all the familiar problems of an area of inquiry pursued by researchers from multiple disciplines, some having descriptive and some having analytic goals. Model builders are more interested in variable measurement errors than in most errors of nonobservation. Describers are more interested in bias terms from nonobservation and observation.

It seems unwise to use the existing methodological literature to determine best procedures for AIDS-related surveys. When no standards or history of methodological support for a procedure exist, we should incorporate measures of errors in designs. We need to become measurers when there is little evidence of reducer success. This means that methodological experiments must be the rule for the early work of AIDS-related surveys.

These methodological experiments must give unusual attention to errors of nonobservation while continuing attention to measurement errors.

References

Anderson, D., & Mantel, N. (1983). On epidemiologic surveys. *American Journal of Epidemiology,* 118, 5, 613-619.

Bradburn, N., Sudman, S., & associates (1979). *Improving interview method and questionnaire design.* San Francisco, Jossey-Bass.

Cannell, C., Miller, P., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.). *Sociological methodology.* San Francisco, Jossey-Bass.

Cialdini, R. (1984). *Influence: The new psychology of modern persuasion.* New York, Quill.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Boston: Houghlin Mifflin.

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C. Turner & E. Martin (Eds.). *Surveying subjective phenomena* (vol. 2, pp. 257-282). Beverly Hills, CA: Russell Sage.

DeMaio, T, Marquis, K., McDonald, S., & associates. (1986). Cognitive and motivational bases of census and survey response. *Proceedings of the Annual Research Conference* (pp. 271-295). U.S. Bureau of the Census.

Deming, W. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association,* 48, 224-255.

Fellegi, I.P. (1964). Response variance and its estimation. *Journal of the American Statistical Association,* 59, 1016-1041.

Goyder, J. (1987). *Silent minority: Nonrespondence in sample surveys.* Boulder, CO: Westview Press.

Groves, R. (1987). Research on survey research data quality. *Public Opinion Quarterly,* 51, 156-172.

Groves, R., (1989). *Survey errors and survey costs.* New York: Wiley & Sons.

Groves, R., & Fultz, N. (1985). Gender effects among telephone interviewers in a survey of economic attitudes, *Sociological Methods and Research,* 14(1), 31-52.

Groves, R., & Magilavy, L. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly,* 50(2), 251-266.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica,* 47, 156-172.

Petty, R. E., & Cacioppo, J. T. (1981). *Communication and persuasion: Central and peripheral roots to communication change.* New York: Springer-Verlag.

Schuman, H., & Converse, J. (1971). The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly,* 35(1), 44-68.

Steeh, C. (1981). Trends in nonresponse rates. *Public Opinion Quarterly,* 45, 40-57.

Stokes, L. (1986). Estimation of interviewer effects in complex surveys with application to random digit dialing. *Proceedings of the Second Annual Research Conference of the U.S. Bureau of the Census* (pp. 21-31). Washington, DC: U.S. Bureau of the Census.

Sudman, S., Bradford, N., Blair, E., & associates (1977). Modest expectations: The effects of interviewers' prior expectations on responses. *Sociological Methods and Research,* 6(2), 171-182.

Thornberry, O. T., & Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. In R. Groves, P. P. Biemer, L. Lyberg, & associates (Eds.). *Telephone survey methodology.* New York: Wiley & Sons.

Key Methodological Problems: An Agenda for Research

Floyd J. Fowler, Jr., Conference Chair

Although the session summaries in this volume present specific research considerations from individual session presentations, discussant papers, and floor discussion, some methodological issues meriting further study emerged from the conference as a whole. These are summarized below.

Question Evaluation

The evaluation of survey questions is one of the areas most in need of attention. It is well documented that questions are routinely asked in major health surveys that have one, and usually more than one, of the following characteristics:

- Interviewers do not, and often cannot, read the questions as worded.
- The questions are not consistently understood by all respondents.
- Respondents are unable to provide the information necessary to answer the question.
- Respondents are unwilling to provide the information required by the question.
- The answers do not mean what the researchers think or hope they mean.
- The questions do not measure what the researchers think or hope they measure.

Various techniques, some experimental and others reasonably well developed, were presented about how researchers could evaluate questions, either before or after they are used in a survey. Among these are the following:

- intensive laboratory interviews
- focus group discussion
- coding behavior in pretest interviews
- reinterviews with respondents
- asking parallel forms of questions
- correlational analyses
- comparing survey answers with comparable information derived from other sources, such as clinical evaluations or medical records

Each of these strategies may provide evaluative information about one or more of the potential question problems listed above. However, at the moment, there is no agreement about which evaluative techniques should be used routinely and, for those that should not be routine, the conditions under which they are necessary. In most cases, there are no standard protocols or guidelines on how to use these techniques. Researchers are currently experimenting and developing their own methods. Most important, although the potential value of each of the above techniques is accepted by methodologists, few surveys employ more than one or two of these techniques to evaluate the questions used, and many surveys use none of them.

No agreement exists on what should be reported about question quality. In the absence of such standards, there is no incentive to researchers to try to collect the evaluative information they need in order to determine whether their questions are good.

A set of systematic evaluations of these different approaches to question evaluation is needed that assesses the sensitivity and specificity with which the techniques identify question problems and the cost effectiveness of different ways of using them to develop a more systematic, scientific, and effective set of procedures for question evaluation.

Validity of Survey Data

Studies comparing survey reports with comparable information derived from records point out two key issues in any study attempting to produce descriptive data about a population: First, it is critical to define very carefully what it is that is being counted or described. Further, it is important to critically evaluate the ability and willingness of the person answering questions or completing forms to provide the required information.

Comparisons of record data with survey-based data consistently reveal discrepancies which stem from differences in definition or ways of counting. For example,

when counting visits to doctors, numerous issues about counting rules are encountered. Should all of the following—psychiatrists, psychologists, ophthalmologists, optometrists, acupuncturists, faith healers—count as doctors? What if a person sees a physician's assistant or nurse in a doctor's office? If a person receives services from a laboratory technician and an x-ray technician, does each service count as a separate visit or should all be included in one medical event? Does the counting rule change if the patient goes back to have an x-ray on the day following the initial contact with the doctor? If a researcher simply asks a respondent for the number of visits to doctors in some time period, there is every reason to expect inconsistency among respondents in their definitions of what is included, and there is every reason to think that the researcher will not know what respondents did or did not include in their count.

Another important source of discrepancy between record-based data and survey reports is associated with the information available to those actually answering the questions. Some respondents simply may not know how their health provider is set up corporately (an independent practice association or health maintenance organization, for example). Others may be unaware of the details of their health insurance coverage. Then too, critical facts about an individual patient's medical history or symptoms may be unknown to respondents completing medical records. All these situations may result in discrepancies between record and survey data to a greater or lesser degree.

Studies comparing survey reports with data derived from records or other sources are valuable in helping people understand the nature of what is and is not reported in surveys. Although there were several such studies done on a large scale in the 1960s, such studies have been comparatively rare in the past 20 years. However, when they are done, as documented in several papers in this volume, they highlight the definitional issues—the issue of what is being counted—and also focus attention both on the ability and willingness of various respondents to provide needed information. There is a continuing need for such studies.

Collecting Data from Older People

Although several issues were discussed with respect to surveys of older people, two issues stood out as most in need of future attention. First, a continuing problem in surveys of older people is having a frame from which to sample. Most lists of older people have significant omissions; household-based samples are somewhat inefficient (only about 20 percent include someone 65 or older). Also, household-based samples exclude those in nursing homes—an important group for health services research.

The availability of a list of Medicare beneficiaries as a sample frame for some research purposes provides an important new approach to sampling older people. One important need is to develop experience with using that sample frame, to understand the strengths and limitations of sampling in this way, and to develop effective ways of enlisting cooperation from people sampled from that frame.

Regardless of the sampling frame, perhaps the most important methodological problem in doing research on older Americans is nonresponse. Response rates both for personal interviews with the elderly and for telephone followup interviews after a personal contact are similar to those for other groups. However, older adults are distinctively less receptive to contact through random digit dialing (RDD). One needed area of research is to better understand the reasons for that since RDD sampling is a very popular and cost-effective way to do surveys requiring screening. Research is also needed to better document the biases that result from differential nonresponse among older people.

Regardless of the mode of data collection, however, a certain amount of nonresponse will result from illness and an inability to be a respondent for other reasons. In those instances, the potential of proxy respondents to provide information needs further study. Questions regarding the use of proxies include the following:

- What kind of persons (that is, what relationships they have to respondents) can provide reliable proxy information?
- How does proxy information correspond with data collected by self-report?
- Are there generalizations about the kind of data that can be collected by proxy? In particular, to what extent, if at all, can proxy respondents give meaningful answers dealing with subjective states, such as feelings, knowledge, or opinions?

One suggestion that merits further consideration is to explore the value of routinely collecting some information from both respondents and from significant other people in households so that the strengths, weaknesses, and degree of correspondence from the two kinds of reports can be evaluated. Potentially data from both sources could be combined in some way to produce the best possible estimates.

Although nonresponse and proxy reporting are particularly relevant for studies of older people, they are also relevant for surveys of all populations. Most comparative studies have shown that proxy information is not as "accurate" as self-reports. However, when the choice is between no data at all and a report from a proxy respondent, the standards for the quality of data from proxy respondents are different. Increasingly, nonresponse is becoming one of the most important sources of error in surveys. Consequently one of the priorities for research should be further exploration and evaluation of the potential of proxy respondents to reduce the error in survey estimates caused by nonresponse.

Research Related to AIDS

Asking questions about potentially sensitive behaviors and gathering samples of relatively rare or hard-to-find populations are not new issues to health survey research. However, methodological knowledge about how to produce effective and reliable survey data on AIDS-related topics remains scarce.

Both conference sessions on AIDS focused on the need for methodological research rather than on documented ways to solve measurement problems. The four issues that stood out during these sessions are described below:

1. What are the uses and limitations of standard telephone surveys based on random digit dialing? This is the methodology most often used for surveys in the United States, and researchers are using this methodology to the greatest extent possible for studies related to AIDS. The RDD method seems to be a feasible way to collect data about general population knowledge and attitudes. However, there is some doubt as to whether such surveys can provide useful data about the behaviors that put people at risk of contracting AIDS and about the population groups whose behavior is most critical to the spread of AIDS. Questions to address include the following:

- What biases are introduced by leaving out people who cannot be associated with housing units or who live in housing units that lack telephone service?
- What is the significance of nonresponse to telephone surveys, and to what extent is nonresponse to RDD studies distinctively biased in ways that affect estimates?
- To what extent can sexual and drug use practices that affect risk of AIDS be measured effectively in a telephone survey?

Data are being collected and results reported on the basis of RDD telephone surveys, and a better understanding of the meaning and possible biases of data emanating from these studies is needed.

2. Groups of special relevance to studies of AIDS should be sampled in ways other than standard household-based sampling. Prostitutes, intravenous drug users, and gay men are examples of three groups who may be hard to sample using standard household sampling techniques. Research is much needed on the strengths and limitations of alternative ways of sampling such people.

3. Research is also needed on how to design interview schedules to provide useful data about high-risk sexual behavior. The limit of survey research is what people are willing and able to answer. Questions can be designed to help make interviewers and respondents comfortable in interviews about sexual behavior. However, careful thought and empirical research are needed to determine whether the comfortable ways of asking questions can also provide the required data. In addition, the way questions are organized to deal with sexual behavior affects both their analytic usefulness and the ability of people to recall and report accurately. It is quicker and in some ways easier to ask summary questions about risky behavior, but it is not possible to classify a behavior—such as having sexual intercourse without a condom—as risky without having information about the particular partner involved.

At the moment, researchers are using various ways to devise questions that characterize sexual activity. A great deal more work is needed to understand the implications of the different approaches to asking such questions.

4. Finally all surveys, to one degree or another, suffer from error due to nonresponse, limitations of the frame from which samples are selected, questions that are not clear to all respondents, respondents who are unable or unwilling to answer questions accurately, and interviewers who influence the answers they obtain. Survey research related to AIDS is distinctive only in the degree to which each of these potential sources of error can be seen as important and a major source of bias in the estimates. Although good survey practice tries to minimize the effect of various sources of error on data estimates, there is a limit in the extent to which error can be eliminated. Thus, it is always good practice to take the special steps needed to measure error to the extent possible.

At the First Conference on Health Survey Research Methods in 1975, Horvitz presented a paper on the total survey design perspective. In his presentation, he noted that while there are good estimates of sampling errors, estimates of the impact of various sources of nonsampling error in survey data tend to be very poor or nonexistent. Unfortunately, that situation continues to exist today with respect not only to AIDS-related research but with respect to survey research in general. The most pressing need for survey methodology is the need for better estimates of how the specific wording of questions, the recall period used, nonresponse, and interviewer behavior and procedures affect survey estimates. It is through such estimates of error that the quality of the data collected can be evaluated and, of equal importance, the priorities for designing better data collection procedures can be set.

Conference Participants

Lu Ann Aday, Ph.D.
Associate Professor of Behavioral Sciences
School of Public Health
University of Texas
P.O. Box 20186
1200 Herman Presler
Houston, TX 77225
(713) 792-4372

William Aquilino, Ph.D.
Post-Doctoral Fellow
Center for Demography and Ecology
2435 Social Science Bldg.
University of Wisconsin
Madison, WI 53706
(608) 263-4020

Sandra H. Berry
Director
Survey Research Group
The RAND Corporation
P.O. Box 2138
1700 Main St.
Santa Monica, CA 90406
(213) 455-2478

Ron Biggar
Survey Statistician
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7100

Johnny Blair
Manager of Operations and Sampling
Survey Research Laboratory
University of Illinois
1005 W. Nevada St.
Urbana, IL 61801
(217) 333-6154

Judith B. Bradford, Ph.D.
Assistant Professor and Associate Director
Survey Research Laboratory
Virginia Commonwealth University
Box 3016
901 W. Franklin St.
Richmond, VA 23284
(804) 367-8813

Donald J. Brambilla, Ph.D.
Senior Research Scientist
New England Research Institute
9 Galen St.
Watertown, MA 02172
(617) 923-7747

Virginia S. Cain, Ph.D.
Sociologist
Demographic and Behavioral Sciences Branch
Center for Population Research
National Institute of Child Health and Human Development
6130 Executive Blvd.
Bethesda, MD 20892
(301) 496-1174

Joan Callahan, Ph.D.
Associate Professor
Department of Philosophy
University of Kentucky
1415 Patterson Office Tower
Lexington, KY 40506
(606) 257-1861

Kathleen A. Calore
Senior Analyst
Health Economics Research, Inc.
75 Second Ave., Suite 100
Needham, MA 02194
(617) 444-8910

Barbara Campbell, Ph.D.
Senior Survey Director
National Opinion Research Center
1155 E. 60th St.
Chicago, IL 60637
(312) 702-8496

Charles Cannell, Ph.D.
Professor and Research Scientist Emeritus
426 Thompson St.
The University of Michigan
Ann Arbor, MI 48106
(313) 936-0092

Frank J. Capell
Epidemiologist
Office of AIDS
California Department of Health Services
714/744 P St.
Box 942732
Sacramento, CA 94234-7320
(916) 445-0553

DonnaRae Castillo
Senior Writer/Editor
Division of Research Dissemination and External Liaison
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-2904

Dorothy Cerankowski
Senior Field Supervisor
Center for Survey Research
University of Massachusetts
100 Arlington St.
Boston, MA 02116
(617) 956-1150

James R. Chromy, Ph.D.
Vice President of Statistical Sciences
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
(919) 541-6228

Steven B. Cohen, Ph.D.
Senior Biostatistician
Division of Intramural Research
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-4836

Marcie L. Cynamon
Special Assistant to the Director
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7085

Jack Elinson, Ph.D.
Principal Investigator
Pediatric Resource Center
Outcome Study
Medical and Health Research Association of New York City
40 Worth St.
New York, NY 10013
(212) 393-1310

Z. Erik Farag, Ph.D.
Director
Division of Research Dissemination and External Liaison
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-2904

Manning Feinleib, M.D., Dr. P.H.
Director
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7016

John A. Fleishman, Ph.D.
Assistant Professor
Department of Community Health
Project Director
Center for Gerontology and Health Care Research
Brown University
Box G-B214
Providence, RI 02912
(401) 863-3211

Floyd J. Fowler, Jr., Ph.D.
Senior Research Fellow
Center for Survey Research
University of Massachusetts
100 Arlington St.
Boston, MA 02116
(617) 956-1150

Howard E. Freeman, Ph.D.
Professor of Sociology
University of California, Los Angeles
Los Angeles, CA 90024
(213) 206-6721

John H. Gagnon, Ph.D.
Professor of Sociology
State University of New York at Stony Brook
Stony Brook, NY 11794
(516) 632-7734

Judith Garrard, Ph.D.
Associate Professor
Division of Health Services Research and Policy
School of Public Health
University of Minnesota
420 Delaware St., S.E.
Minneapolis, MN 55455
(612) 624-6151

Robert M. Groves, Ph.D.
Associate Professor of Sociology
Program Director and Senior Research Scientist
Survey Research Center
The University of Michigan
Ann Arbor, MI 48106-1248
(313) 936-0027

Larry A. Hembroff, Ph.D.
Associate Professor
Survey Director
Center for Survey Research
Social Science Research Bureau
301 Olds Hall
Michigan State University
East Lansing, MI 48824-1111
(517) 353-3255

Daniel Horvitz, Ph.D.
Distinguished Institute Scientist
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
(919) 541-6450

William D. Kalsbeek, Ph.D.
Associate Professor
Department of Biostatistics
School of Public Health
University of North Carolina
Rosenau Hall 201 H
Chapel Hill, NC 27514
(919) 962-3249

Graham Kalton, Ph.D.
Chairman
Department of Biostatistics
Survey Research Center
The University of Michigan
Ann Arbor, MI 48109
(313) 764-5450

Robert M. Kaplan, Ph.D.
Professor and Acting Chief
Division of Health Care Sciences
UCSD School of Medicine
La Jolla, CA 92093
(619) 534-6058

Judith D. Kasper, Ph.D.
Assistant Professor
Department of Health Policy and Management
The Johns Hopkins University
624 N. Broadway
Baltimore, MD 21205
(301) 955-2487

Barbara Kerr
Program Analyst
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782

Dorothy W. Kingery, Ph.D.
Director
Survey Research Center
University of Georgia
544 Graduate Studies Bldg.
Athens, GA 30602
(404) 542-6110

Elaine Kokiko
Executive Vice President
Moshman Associates, Inc.
Suite 410, North Tower
7315 Wisconsin Ave.
Bethesda, MD 20814
(301) 229-3000

Mary Grace Kovar, Dr. P.H.
Special Assistant for Data Policy and Analysis
Office of Vital and Health Statistics System
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7105

Jennie J. Kronenfeld, Ph.D.
Director
Health Survey Laboratory
Professor
Department of Health Administration
School of Public Health
University of South Carolina
Columbia, SC 29208
(803) 777-6096

Richard A. Kulka, Ph.D.
Associate Director
National Opinion Research Center
1155 E. 60th St.
Chicago, IL 60637
(312) 702-1200

Edward O. Laumann
George Herbert Mead Distinguished Service Professor of Sociology
Dean
Division of Social Sciences
University of Chicago
1126 E. 59th St., Room 110
Chicago, IL 60637
(312) 702-8798

William B. Lohr
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-3091

Nancy A. Mathiowetz, Ph.D.
Senior Survey Methodologist
Division of Intramural Research
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-4836

Ian McDowell, Ph.D.
Associate Professor
Department of Epidemiology and Community Medicine
University of Ottawa
451 Smythe Rd.
Ottawa, Ontario K1H-8M5
Canada
(613) 787-6480

David V. McQueen, Ph.D.
Director
Research Unit—Health and Behavioral Change
University of Edinburgh
17 Teviot Place
Edinburgh, EH1 2 QZ
United Kingdom
011-44-031-225-5402

Doris R. Northrup
Vice President
Westat, Inc.
7823 Custer Rd.
Bethesda, MD 20814
(301) 251-8214

Janet D. Perloff, Ph.D.
Visiting Research Associate Professor
University Center for Policy Research
State University of New York at Albany
135 Western Ave.
Albany, NY 12222
(518) 482-0871

Stanley Presser, Ph.D.
Professor
Department of Sociology
Director
Survey Research Center
The University of Maryland
College Park, MD 20742-1315
(301) 454-5564

Mark Reiser, Ph.D.
Assistant Professor
Department of Decision and Information Systems
College of Business
Arizona State University
Tempe, AZ 85287-4206
(602) 965-5486

Willard Rodgers, Ph.D.
Associate Research Scientist
Survey Research Center, Institute for Social Research
The University of Michigan
Ann Arbor, MI 48106-1248
(313) 764-5450

Patricia Royston, Ph.D.
Mathematical Statistician
Office of Research and Methodology
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7111

Ken R. Smith, Ph.D.
Associate Professor
Department of Family and Consumer Studies
Director
Survey Research Center
University of Utah
2120 Annex Bldg.
Salt Lake City, UT 84112
(801) 581-5459

Seymour Sudman, Ph.D.
Walter H. Stellner Distinguished Professor of Marketing
University of Illinois
1005 W. Nevada St.
Urbana, IL 61801
(217) 333-4273

Cynthia Taeuber
Chief
Age and Statistics Branch
Population Division
Bureau of the Census
Washington, DC 20233
(301) 763-7883

Cynthia Thomas, Ph.D.
Senior Research Associate
Montefiore Medical Center
H. & L. Moses Hospital Division
111 E. 210th St.
Bronx, NY 10467
(212) 920-6481

Owen Thornberry, Ph.D.
Director
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20782
(301) 436-7085

Lois M. Verbrugge, Ph.D.
Research Scientist
Institute of Gerontology
The University of Michigan
300 N. Ingalls
Ann Arbor, MI 48109-2007
(313) 764-3493

Daniel Walden, Ph.D.
Senior Research Manager
Division of Intramural Research
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-4836

Richard Warnecke, Ph.D.
Professor
Department of Sociology
Director
Survey Research Laboratory
University of Illinois at Chicago
P.O. Box 6905
Chicago, IL 60680
(312) 996-6130

Michael F. Weeks
Project Director
National Household Seroprevalence Survey
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709
(919) 541-6000

Norman Weissman, Ph.D.
Director
Division of Extramural Research
National Center for Health Services Research and Health Care Technology Assessment
5600 Fishers Ln.
Rockville, MD 20857
(301) 443-2345

Deborah M. Winn, Ph.D.
Deputy Director
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Hwy.
Hyattsville, MD 20872
(301) 436-7085

☆U.S.GOVERNMENT PRINTING OFFICE:1989-626-157/10586

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. NCHSR 89-15	2.	3. Recipient's Accession No. PB90-100082/AS

4. Title and Subtitle
Health Survey Research Methods -- Fifth Conference Proceedings
DHHS Publication No. (PHS) 89-3447

5. Report Date September 1989

6.

7. Author(s)
Floyd J. Fowler

8. Performing Organization Rept. No.

9. Performing Organization Name and Address

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.
(C)
(G) HS06081

12. Sponsoring Organization Name and Address
DHHS, PHS, OASH, National Center for Health Services Research and Health Care Technology Assessment (NCHSR)
Publications and Information Branch, 18-12 Parklawn Building
Rockville, MD 20857 Tel.: 301/443-4100

13. Type of Report & Period Covered

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

These proceedings are the result of the fifth conference on Health Survey Research Methods, held at Keystone, Colorado, May 2-4, 1989. Twenty-five feature papers, 10 discussion papers, 5 session summaries, and a conference summary address the following major topics: "Strategies for evaluating questions," "Validity of reporting in surveys," "Collecting data from samples of older adults and nursing home populations," "Samples for studies related to AIDS," and "Measuring behavior related to risk of AIDS." In addition, the conference summary also discusses key methodological problems that merit further research. Among these are the need for a set of systematic evaluations of different approaches to question evaluation assessing both sensitivity and specificity; the need for studies comparing survey reports with data derived from records; the need for research on the potential of proxy respondents to reduce survey error caused by nonresponse; and the need for methodological research on AIDS. Finally it was pointed out that "the most pressing need for survey methodology is the need for better estimates of how the specific wording of questions, the recall period used, nonresponse, and interviewer behavior and procedures affect survey estimates."

17. Document Analysis a. Descriptors

NCHSR publication of research findings does not necessarily represent approval or official endorsement by the National Center for Health Services Research and Health Care Technology Assessment or the U.S. Department of Health and Human Services.

b. Identifiers/Open-Ended Terms
health services research, survey methodology, survey design, data collection procedures, sampling error

c. COSATI Field/Group

18. Availability Statement:
Releasable to the public. Available from National Technical Information Service, Springfield, VA 22161 Tel.: 703/487-4650

19. Security Class (This Report) Unclassified

20. Security Class (This Page) Unclassified

21. No. of Pages 277

22. Price