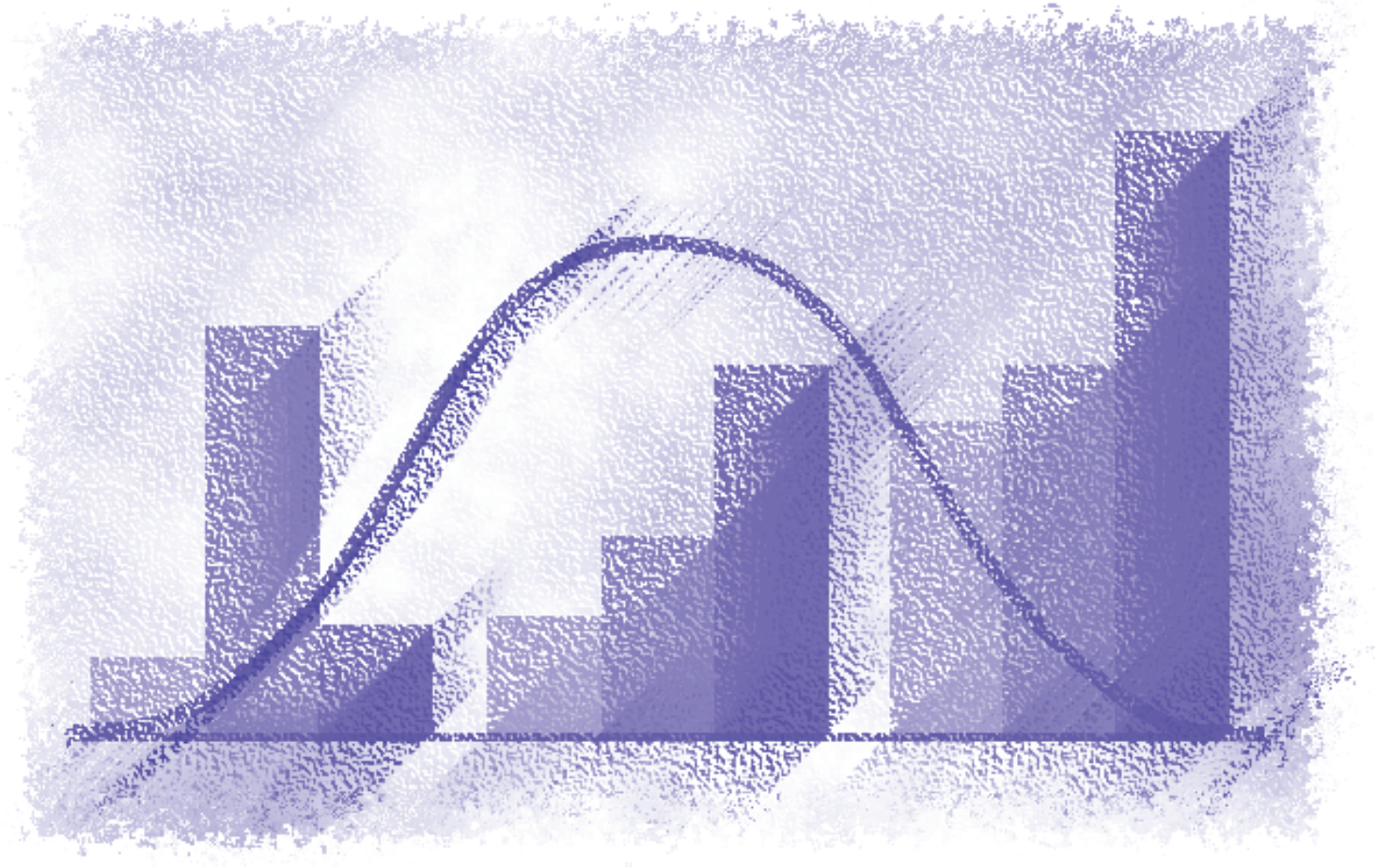


Seventh Conference on

HEALTH SURVEY RESEARCH METHODS

CDC

SAFER • HEALTHIER • PEOPLE™

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Seventh Conference on

HEALTH SURVEY RESEARCH METHODS

Edited by
Marcie L. Cynamon and Richard A. Kulka

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
February 2001

DHHS Publication No. (PHS) 01-1013

Dedicated to Seymour Sudman (1928–2000), University of Illinois, who was a charter member of the Health Survey Research Methods conference series. He attended all seven conferences and chaired the third conference in 1979. His contributions to the field of health survey research methods are both legion and enduring, and his intellect, sense of humor, and dedication to improving health survey research will be sorely missed.

The Seventh Conference on Health Survey Research Methods (HSRM) was held in Williamsburg, Virginia, on September 24–27, 1999, continuing a series of meetings that began in 1975 to discuss new, innovative survey research methods to improve health survey research data. The HSRM conferences bring together researchers from a variety of disciplines, including those who are at the forefront of survey methods research, are responsible for major health surveys, and use survey data to develop and implement health policy and programs. As with the previous HSRM conferences, the overarching goal was to review, critique, and add to our body of knowledge about survey methodologies to improve the quality of health survey methods and enhance the value and utility of the data that surveys provide for policymakers responsible for shaping health practice, policy, and programs. This is accomplished by

- Selecting presentations that represent progressive survey methods research relevant to providing quality data related to the nation's health
- Providing a forum for critical discussion of the presentations, using a format that combines formal discussion by expert reviewers and open discussion by all conference participants
- Preparing and publishing a complete written summary of the conference, compiled from the formal papers, comments from invited discussants, and general discussion by participants, in the form of a formal conference proceeding

The Seventh Conference focused on

- Major survey activities within and outside the federal statistical system
- The potential impact of ongoing methodological research in surveys
- How survey methods affect the usefulness of survey data in addressing the information and policy needs of those charged with planning, delivering, and improving health assessments, services, and research to the nation

The Conference brought together four key groups of stakeholders who contribute to and/or benefit from ongoing health survey research to review and critique the current state of survey methods in this area and develop concrete recommendations for how these might be enhanced, improved, and better focused. These groups comprised

- Researchers from various disciplines who are engaged in survey methods research
- Researchers and administrators responsible for the major health surveys in the federal statistical system
- Researchers and practitioners who use survey data to assess health policies
- Those in the government who make these policies

In all, 76 persons attended this conference. Twenty-seven papers were selected from 80 submissions. The selected papers were presented in five topical sessions:

1. Collecting Data from Children and Adolescents (5)
2. Racial and Ethnic Populations: Cross-Cultural Considerations (5)
3. Comparability of Data across Different Modes of Data Collection (6)
4. Validity of Results (6)
5. Needs for State and Local Data of National Relevance (5)

There was also a special panel session on "Policy Challenges for the Future—International, National, and State Surveys." A chairperson and two rapporteurs were assigned to each session and the special panel, and there were two formal discussants for each paper session.

These themes represent both continuity with and change from previous meetings. While the goals of reducing survey error and increasing the utility of the data remain constant, the methods to achieve these goals continue to evolve. For example, earlier conferences focused on the importance of using standardized procedures. In more recent years, a growing need has emerged to adapt and tailor survey procedures to successfully include diverse geographic and other population subgroups, such as young children, ethnic minorities, and persons with specific characteristics.

Background and History of the Health Survey Methods Conferences

In 1975, a group of researchers representing both academic institutions and government research agencies met informally to discuss the strengths and limitations of health survey data and how the data could be improved. Specifically, the discussion centered on developing a mechanism that would provide a forum for discussion of the results of methodological research in health surveys. This forum would allow for the effective communication of research findings to a large body of researchers engaged in broad areas of health research, ranging from health services to epidemiology. They concluded that a conference specifically devoted to health

survey research methods was needed for the following key reasons:

(a) Researchers using, developing, and/or evaluating survey research methods were widely dispersed, both geographically and in their work settings and disciplines. Hence, effective communication among those working on survey methods relevant to health surveys was difficult, and the results of their work were often inaccessible to others conducting health surveys.

(b) Interdisciplinary communication was further impeded by the absence of a specific forum where regular discussion of survey methodology relevant to health data collection could be a central focus. Even in those forums that existed primarily for discussion of survey methods, such as the annual meetings of the American Association for Public Opinion Research (AAPOR) and the Survey Research Methods Section of the American Statistical Association, the focus was broader than health survey methods alone.

(c) Methodological findings—particularly work in progress, negative results, and studies of methodologies that do not work—were not routinely reported in traditional journals, although they may have important implications for those engaged in health research. Moreover, methodological findings in major reports that were tangential to the main substantive research questions or reporting requirements were often cryptically reported.

That first conference, held at Arlie House in Arlie, Virginia, in 1975 and led by Leo Reeder, was attended by 50 substantive researchers and methodologists with common interests in health survey research. The specific goals of the conference were to (1) identify common survey research problems and describe the current philosophy regarding these problems, (2) determine which issues merited the highest priorities for funding, (3) identify key policy issues that could be informed or developed using survey data, and (4) disseminate these results and their implications to the widest possible audience of data users and survey researchers. Participants at the first conference decided that the meeting and its proceedings would be worthwhile to the larger research community and that similar meetings should be held over the next few years with their proceedings published.

The 1975 conference was sponsored by the National Center for Health Statistics and the National Center for Health Services Research (now called the Agency for Healthcare Research and Quality). Its format consisted of open discussion on specified broad topics that was recorded for publication. This format continued for the next two conferences. Since then, numerous government agencies, foundations, private nonprofit research organizations, and universities have contributed financial and/or administrative support to these meetings. The period between meetings has lengthened beyond the original intention to conduct them approximately every two years, because all funds must be raised by the planning committee to cover the costs of each conference and of publishing and disseminating the proceedings.

The composition of the planning committee has also changed over the years. As funding sources expanded, so did the committee. Participation at the early meetings was by invitation only. Today a more expansive approach is used, balancing invited and contributed presentations rather than relying solely on open discussion. Following each session is a formal discussion by one or two persons, followed in turn by a general floor discussion. Both formal and informal discussions are captured for inclusion in the proceedings. This has served to broaden the conference topics to encompass new areas of inquiry and identify important areas for future consideration. It is also an acknowledgment that the field of health survey research is growing rapidly, as are its challenges.

Conference Themes

While the focus of these conferences has always been on survey research methods, with an emphasis on nonsampling error, specific themes have varied from year to year. The first five conferences each had a session devoted to total survey design. All seven have addressed issues of the validity of survey data, both in the form of independent sessions and as a component of the topics of questionnaire design, respondent recall and burden, and validation of survey responses through records or other external sources. Most have had a session on sample design and the problem of locating rare, minority, or hard-to reach populations. Each conference has had a session on mode of data collection, with the newly emerging Web-based data collection considered for the first time at the seventh conference. The primary goal in each case was to present what was known and unknown about sources of survey error and how to minimize it. Unlike other professional conferences, presenting negative findings or "unsuccessful" results is viewed as appropriate and encouraged.

The major policy issues of the day determined how these themes were integrated into each conference. The impact of the Privacy Act of 1974 on response rates was a topic of considerable concern at the second conference, held in 1977. In 1979, the government was beginning to collect data on access to and cost of health care. Reflecting those objectives, several sessions in the third and fourth conferences addressed the design and implementation of surveys on cost of and access to health care services. By 1989, interest had shifted toward data that are inherently very difficult to collect. Concerns about HIV/AIDS required the development of questionnaires that delved into topics far more sensitive than any that had appeared to date in government-sponsored health surveys. While measurement issues related to access and medical expenditures were an important topic for discussion, the primary focus of the fifth conference was the total design of surveys related to homelessness, AIDS-related risk behaviors and measuring the prevalence of HIV/AIDS, and the new strategies being devised to address these major health problems.

Issues surrounding cognition began to emerge during the third conference and gradually became more dominant. Improving the reliability of data through the use of diaries and memory aids was addressed at the third and fourth conferences in the context of provider and patient surveys. The application of cognitive psychology was an area of focus at the two subsequent conferences, with sessions that explored the validity of question wording, cultural effects of interpretation, and improving pretesting techniques.

Although each of these principal foci of previous conferences continued in one form or another, the sixth conference reflected an increased interest by federal programs in using existing data sets and surveys in creative ways for program evaluation and obtaining policy-relevant data—e.g., through the use of add-on items to existing surveys, administrative data, and other strategies—along with the opportunities and challenges that such strategies provide. While not an area of focus during the seventh conference, the trend toward survey integration—the multiple use of sampling frames and questionnaires to enhance the utility of data—of federal surveys has continued.

The choice of themes for each conference reflects a strategic balance between (1) reflecting important shifts in the focus of major heath policy issues in the years immediately prior to the conference, (2) anticipating possible changes in the health policy research agenda in the near term, (3) identifying major developments in survey research methods relevant to addressing those emergent or forecast issues, and (4) adding to what is known and addressing new areas that had been featured briefly at previous conferences.

In selecting the most important methodological issues to address in the seventh conference, members of the planning committee were faced with the same constraints evident in previous conferences. In effect, it is not possible to fully anticipate the future health policy issues that survey research methods will be called on to address, but that is precisely the challenge presented to health survey researchers in these conferences. To address that challenge, the structure and content of the conference must both anticipate future directions and assess the extent to which current survey methods are adequate to address health policy questions currently in play—both intractable problems that have been with us for many years in one form or another (e.g., the need to ask sensitive questions) and those that have evolved over the past few years.

Thus, the planning committee suggested that the overarching theme for the seventh conference should be identifying the new questions of each type and providing a systematic assessment of the extent to which our survey methods are adequate to address them. The featured papers solicited, selected, and presented (including the special panel) did, we believe, achieve the desired balance and ultimate goal. In each session these issues were raised in the context of problems faced by one or more large-scale federal statistical surveys and were regarded as presenting significant new challenges, both immediately and over the next several years. Moreover, in most cases these issues represent logical extensions of methodological issues raised and discussed at one or more of the previous conferences, and most are the subject of ongoing or planned methodological research.

Marcie L. Cynamon
Special Assistant
National Center for Health Statistics
Centers for Disease Control and Prevention

Richard A. Kulka
Research Vice President
Statistics, Health, and Social Policy
Research Triangle Institute

ACKNOWLEDGMENTS

This conference would not be possible without the generous contributions from agencies committed to its success. As the conference increases in size and scope, so does the demand for funding. We are particularly grateful to our growing list of supporters: Agency for Health Care Policy and Research (currently, the Agency for Healthcare Research and Quality), Health Care Financing Administration, Health Resources Services Administration, National Cancer Institute, National Center for Health Statistics, National Institute on Alcohol Abuse and Alcoholism, National Institute on Drug Abuse, Robert Wood Johnson Foundation, Substance Abuse and Mental Health Services Administration, and Survey Research Center of the University of Michigan.

The conference planning committee is composed of representatives from the academic and research communities and the funding agencies, who met periodically to develop the structure and content of the conference. We want to recognize the efforts of the committee members and thank them for their dedication to the myriad arduous organizing tasks: Lu Anne Aday (University of Texas at Houston), Barbara Bailar (National Opinion Research Center), Richard Campbell (University of Illinois at Chicago), Steven Cohen (Agency for Healthcare Research and Quality) Brenda Edwards (National Cancer Institute), Floyd J. Fowler, Jr. (University of Massachusetts Boston), Joseph Gfroerer (Substance Abuse and Mental Health Services Administration), Michael Hilton (National Institute on Alcohol Abuse and Alcoholism), Arthur Hughes (National Institute on Drug Abuse), James Lepkowski (University of Michigan), Katherine Marconi (Health Resources and Services Administration), Nancy Mathiowetz (University of Maryland), and Richard Warnecke (University of Illinois at Chicago).

All sessions during the three-day conference were plenary, and participation by all attendees is an essential part of the conference. Relevant presentations followed by insightful formal and informal discussion are the hallmark of the Health Survey Research Methods conferences, and this one is no exception. We thank the participants from academia and the government for sharing their stimulating research for these proceedings. The quality of the presentations was universally excellent.

No conference can succeed without the driving force of one very special person who makes it all happen. For us, that person is Diane O'Rourke of the Survey Research Laboratory of the University of Illinois. In spite of our many inadvertent efforts to derail her, Diane managed to stay focused as she arranged planning meetings, distributed the Call for Papers and subsequent abstracts, selected the site, negotiated travel arrangements, oversaw the budget, coordinated the logistics of preparing the proceedings manuscript, and catered to everyone's insatiable needs. This was Diane's second time organizing the HSRM conference and, amazingly, she has agreed to organize the next one. Many, many thanks. Diane was assisted greatly by Kristen Hertenstein, whose good humor and knowledge kept us on track. Ingrid Graf helped out at the conference, ensuring that all went as planned.

As co-chairs of the conference and the organizing committee, we have enjoyed tremendously the opportunities we were given to work with a wonderful group of professionals whose interest in survey research is an inspiration. We took great pleasure in overseeing the conference and in working together over the years it took to plan. We look forward to future conferences as the field of health survey research methods continues to evolve.

Marcie L. Cynamon

Richard A. Kulka

CONTENTS

FOREWORD iii

ACKNOWLEDGMENTS vii

SESSION 1: COLLECTING DATA FROM CHILDREN AND ADOLESCENTS 1

Collecting Information about the Health Experiences of Publicly Insured Adolescents
Patricia M. Gallagher, Floyd Jackson Fowler, Jr., and Diana Elliott 3

Improving Adolescent Health Care Surveillance
Jonathan D. Klein, Caryn A. Graff, John S. Santelli, Marjorie J. Allan, and Arthur B. Elster 11

Young Children's Reports of Their Health: A Cognitive Testing Study
Anne Riley, George Rebok, Christopher Forrest, Judy Robertson, Bert Green, and Barbara Starfield 19

Innovative Strategies for Increasing Active Parental Consent in School-Based Drug Education Research
Jennifer Hawes-Dawson, Gail Zellman, Sarah Cotton, and Marvin B. Eisen 27

Design and Methodological Issues in a National Longitudinal Study of Children in the Child Welfare System
Kathryn Dowd, Paul Biemer, and Michael Weeks 35

DISCUSSION PAPER
Comments on Sampling Issues in Collecting Data from Children and Adolescents
Sandra H. Berry 43

DISCUSSION PAPER
Advantages and Limitations of Using Children and Adolescents as Survey Respondents
Nicholas Zill 47

SESSION SUMMARY
Discussion Notes, Session 1
David Maglott and Elsie Palmuk, Rapporteurs 51

PANEL SESSION: POLICY CHALLENGES FOR THE FUTURE: INTERNATIONAL, NATIONAL, AND STATE SURVEYS
Chair: Lu Ann Aday; Panel Members: Lu Ann Aday, Cathy Schoen 53

PANEL SUMMARY
Discussion Notes, Panel Session
D. E. B. Potter and Richard Strouse 65

SESSION 2: RACIAL AND ETHNIC POPULATIONS: CROSS-CULTURAL CONSIDERATIONS 67

Culture and Item Nonresponse in Health Surveys
Linda Owens, Timothy P. Johnson, and Diane O'Rourke 69

Cross-Cultural Adaptation of Survey Instruments: The CAHPS® Experience
Robert Weech-Maldonado, Beverly O. Weidmer, Leo S. Morales, and Ron D. Hays 75

Readability of CAHPS® 2.0 Child and Adult Core Surveys
Leo S. Morales, Beverly O. Weidmer, and Ron D. Hays 83

A Challenge to the Cross-Cultural Validity of the SF-36 Health Survey:
Maori, Pacific, and New Zealand European Ethnic Groups
Kate M. Scott, Diana Sarfati, Martin I. Tobias, and Stephen J. Haslett 91

Methods for Increasing Recruitment and Retention of Ethnic Minorities
in Health Research Through Addressing Ethical Concerns
Vickie M. Mays 97

DISCUSSION PAPER
Issues in Turning Concerns about Culture and Survey Error into Scientific Questions
Robert M. Groves 101

DISCUSSION PAPER
Racial and Ethnic Populations: Cross-Cultural Considerations in Health Survey Research
Robert L. Santos 105

SESSION SUMMARY
Discussion Notes, Session 2
Terry DeMaio and Diane Makuc, Rapporteurs 111

SESSION 3: COMPARABILITY OF DATA ACROSS DIFFERENT MODES OF DATA COLLECTION 113

Comparing Telephone and Face-to-Face Interviewing in Terms of Data Quality:
The 1982 National Election Studies Method Comparison Project
Melanie C. Green and Jon A. Krosnick 115

Mode of Administration Considerations in the Development of Condition Specific Quality of Life Scales
Todd H. Rockwood, Robert L. Kane, and Ann Lowry 123

Mode Differences in Reports of Alcohol Consumption and Alcohol-Related Harm
Lorraine T. Midanik, John D. Rogers, and Thomas K. Greenfield 129

Obtaining HIV Test Results with a Home Collection Test Kit in a Community Telephone Sample
Dennis H. Osmond, Joseph Catania, Lance Pollack, Jesse Canchola, Deborah Jaffe, Duncan MacKellar, Linda Valleroy . . 135

The Methodological Implications of Conducting Web-Based Research
Elizabeth T. Miller 143

Physician Response in a Trial of High-Priority Mail and Telephone Survey Mode Sequences
Danna L. Moore, Jim Gaudino, Pat deHart, Alan Cheadle, and Diane Martin 149

DISCUSSION PAPER
Discussion of Papers on Mode Effects
Norman M. Bradburn 155

DISCUSSION PAPER
Comparability of Data across Different Modes of Data Collection
Colm O'Muircheartaigh 159

SESSION SUMMARY
Discussion Notes, Session 3
Mary Grace Kovar and Judith Lessler, Rapporteurs 163

SESSION 4: VALIDITY OF RESULTS 165

Methodological Issues in Measuring the Uninsured
Joanne Pascale 167

Methodological Differences in Measuring Health Care Coverage
Rachel Harter, Alma Kuby, and Whitney Moore 175

Identifying Children with Special Needs
Floyd Jackson Fowler, Jr., Patricia M. Gallagher, and Charles J. Homer 181

Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records
Stephen J. Blumberg and Marcie L. Cynamon 189

The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits
Robert F. Belli, James M. Lepkowski, and Mohammed U. Kabeto 197

Use of Geographic Contextual Variables in Examining Survey Item Validity
Catharine W. Burt 205

DISCUSSION PAPER
Discussion Notes
Graham Kalton 211

DISCUSSION PAPER
Discussion Notes
Seymour Sudman 215

SESSION SUMMARY
Discussion Notes, Session 4
Timothy Johnson and Willard Rodgers 217

SESSION 5: NEEDS FOR STATE AND LOCAL DATA OF NATIONAL RELEVANCE 219

Pooling State Telephone Survey Health Data for National Estimates:
The CDC Behavioral Risk Factor Surveillance System, 1995
Ronaldo Iachan, Jane Schulman, Eve Powell-Griner, David E. Nelson, Peter Mariolis, and Carol Stanwyck 221

State Estimates of Substance Abuse Prevalence: Redesign of the National Household Survey on Drug Abuse (NHSDA)
Joseph Gfroerer, Doug Wright, and Peggy Barker 227

The National Immunization Survey: A Surveillance System for State and Local Estimates of Childhood Vaccination Levels
Philip J. Smith, Michael P. Battaglia, Danni Daniels, Victor G. Coronado, and J. N .K. Rao 233

Targeting Approaches to State-Level Estimates
Jennifer H. Madans, Trena M. Ezzati-Rice, Marcie Cynamon, and Stephen J. Blumberg 239

DISCUSSION PAPER
Needs for State and Local Data of National Relevance
James M. Lepkowski 247

SESSION SUMMARY
Discussion Notes, Session 5
Donald Camburn and Arthur Hughes, Rapporteurs 251

CONFERENCE THEMES AND CONCLUSIONS 255

PARTICIPANTS LIST 257

Collecting Data from Children and Adolescents

One of the major trends identified by the Steering Committee in planning this conference was a growing demand for questions on increasingly sensitive topics and, especially, engaging younger and younger respondents in such research. More generally, committee members noted that an increased emphasis and focus on gathering data from and about children and adolescent populations has become a major feature of both the current and near-term future landscape of health survey research, citing numerous major studies either currently in progress or planned. Collecting data from (or even about) children and adolescents obviously poses some significant, special methodological challenges—e.g., access, consent, sampling frames and coverage, age-specific interview issues (including comprehension, attention span, etc.), and interviewer effects—that must be squarely faced as these important new surveys go forward.

Each of the feature papers in this session addresses one or more of these key challenges, and the discussion that followed their presentation (both formal and from the floor) added or underlined some additional concerns and challenges. One overarching theme was a clear recognition that there are indeed many important questions that cannot be answered adequately without collecting the relevant information directly from children and adolescents themselves, rather than from their parents, caregivers, records, or other "proxy" sources. At the same time, however, the barriers to doing so can be formidable.

The first three papers predominantly address the first question by (1) comparing the direct reports of adolescent and parent reports of the teens' health care experiences, (2) assessing the validity and reliability of adolescent self-reports on their preventive health care visits, and (3) exploring in depth the degree to which very young children can provide adequate reports on their own health. The final two papers focus more on the very significant challenges associated with sampling and gaining access to children, including especially (1) soliciting participation from institutional gatekeepers (e.g., schools and agencies), (2) requirements and difficulties associated with obtaining informed consent, and (3) other Institutional Review Board (IRB) and human subject issues.

Collecting Information about the Health Experiences of Publicly Insured Adolescents

Patricia M. Gallagher, Floyd Jackson Fowler, Jr., and Diana Elliott

Introduction

Researchers are very concerned that parents are not good reporters about the experiences of their adolescent children. More and more, people who want to get accurate information about the experiences of adolescents feel it is important to interview teenagers themselves (Hess et al., 1998; Stussman, Willis, & Allen, 1993). As part of our continuing work with the Consumer Assessment of Health Plans (CAHPS®) project to develop survey instruments to measure consumer experiences with health plans, we carried out a number of methodological experiments to understand better the nature of the problem of learning about the health care experiences of adolescents and to explore some alternative ways to collect data about adolescent experiences.

The goal of the CAHPS project is to gather comparable data from samples of health plan members about their experiences in getting medical care. A particular challenge, not unique to CAHPS, is how best to collect data about adolescents. This issue is especially salient for those who gather information about Medicaid members, where more than 16% are aged 6 to 17 (Pamuk, Makuc, Heck, Reuben, & Lochner, 1998). We already have data from a 1997 pilot study of a sample of privately insured adolescents and their parents, which permitted comparison of parent and teen answers (Gallagher & Fowler, 1998). We found that there are differences between parent and adolescent reports of teenagers' health care experiences. The results of focus groups, plus the pilot study, suggested that some questions were best answered by teens, others by their parents.

Our next step was to do a larger study, designed to test alternative ways to collect data about adolescents, that focused on a more complex population of teenagers. This paper reports the results of that study.

There were three overarching goals in this study of publicly insured adolescents. The first was to assess alternative protocols for collecting data from parents and their teenaged children. The protocols tested included self-administration by mail, interviewer administration by telephone, and a mixed-mode approach in which attempts were made to interview nonresponders to a mail protocol by telephone.

The second goal was to compare adolescents' responses by mode to learn more about the comparability of data collected when teens complete a self-administered instrument by mail with data collected when they respond to an interviewer-administered questionnaire by phone. For populations that do not respond well to mail survey requests, collecting data by telephone, either as a primary mode of data collection, or to interview mail nonresponders, can be important. However, such a strategy is only appropriate if data collected by mail and telephone are comparable and can be combined.

The third goal was to obtain better information on the data consequences of the decision about whether parents or adolescents are asked to report on the teenagers' health care experiences. If comparable information about the teen can be collected either from the parent or the adolescent, the additional complications associated with surveying adolescents can be avoided.

Methods

Sample Design

The sampling frame was provided by the Division of Medical Assistance (DMA), which oversees the administration of Medicaid in Massachusetts through the MassHealth program. Families in which at least one teenager, age 13 through 17, had health insurance through MassHealth were sampled. Families were assigned to one of four treatment conditions as outlined in Table 1. One sample was contacted by mail only (n = 600), another by telephone only (n = 600), and the remaining two by a combination of mail and telephone (two samples of n = 800 each). In just one sample, the dual-mode control group, the parent reported about the sampled adolescent's health care. Otherwise, both teenagers and their parents were asked to complete interviews about the teen.

In an attempt to reduce complications introduced by the historically imperfect contact information provided by Medicaid, samples for the mail-only and telephone-only protocols were restricted to enrollees for whom both a mailing address and a telephone number were available from Medicaid records. The dual-mode samples were cross-sectional.

The authors are at the Center for Survey Research, University of Massachusetts, Boston. The research reported here was supported by a cooperative agreement from the Agency for Health Care Policy and Research. We also gratefully acknowledge the assistance and cooperation of the Division of Medical Assistance, Commonwealth of Massachusetts.

Table 1. Study design

Sample	Sample Criteria	Respondent(s)	Protocol	*n*
1	Known contact information	Teen and parent	Telephone	600
2	Known contact information	Teen and parent	Mail	600
3	Cross-section	Teen and parent	Dual-mode	800
4	Cross-section	Parent only	Dual-mode	800

Questionnaire Design

The survey instruments were based on the CAHPS 2.0 Child Core questionnaire and included questions about the health plan enrollee's interactions with health care providers and with the health plan. We had evidence from our previous research that parents and adolescents differed in their reports about the teens' experiences with these two types of interactions. In the absence of a gold standard identifying which set of responses best reflects reality, we assumed that parents would be more accurate reporters of health plan interactions and that teen reports of their own experiences with doctors would be better than those of parents. The questionnaire for adolescents centered on questions about their interactions with providers, while the parent instrument primarily contained health plan-related questions.

Because of known linguistic diversity in the sample, all respondents were offered the opportunity to respond either in Spanish or English. The self-administered questionnaires were dual-language instruments, printed in English on one side and Spanish on the other. All contact materials, including the information sheet and reminder postcard, were also presented in a dual-language format. Bilingual interviewers were available to conduct telephone interviews in either Spanish or English.

Data for this study was collected during the spring and summer of 1999.

Data Collection Protocols

Mail Mode

All correspondence was addressed to the parent. Parental compliance with survey instructions was considered tacit consent for adolescent participation. Contact by mail followed standard mail survey research protocols. First, a questionnaire packet was sent to the household. This packet, for all but the control sample, contained an information sheet that asked the parent to do three things: (1) complete the questionnaire entitled "Questions for Parents," (2) give the named child the "Questions for Teens" questionnaire to fill out, and (3) return both completed instruments in an enclosed postage-paid envelope. The packet for the control sample contained a single instrument for parents to complete about their child's health care.

Seven to 10 days after the initial mailing, a thank you/reminder postcard was sent. Approximately 2 weeks after the reminder postcard was mailed, replacement questionnaire packets were mailed to all nonresponding households. If only one of the pair of teen/parent questionnaires had been returned, we sent an individualized follow-up letter identifying the missing respondent, along with the appropriate questionnaire to be completed. About a month later, nonresponders were contacted by telephone. For the mail-only protocol, these calls were reminders to return the mailed questionnaires or to offer a remail, while the dual-mode study members were offered the opportunity to complete telephone interviews.

Telephone Mode

About a week prior to the start of telephone fieldwork, parents in the telephone-only group were sent an information sheet that outlined the study objectives and sponsor and advised them that they would soon be contacted and asked to participate in a short interview. A week later, professional interviewers attempted to contact these households by telephone. The goal was to interview a parent or guardian about the sampled teenager's health care. Once the interview was complete, the interviewer explained to the parent that she or he would like to interview the adolescent directly to ask questions about the child's interactions with health care providers. If the parent consented, the interviewer asked to speak with the teen to learn whether the child was willing to complete an interview.

For both telephone interviews and reminder calls, no fewer than 6 calls were placed; in many cases, considerably more. To ensure adequate coverage, daytime and evening calls were made on different days of the week, including both weekend and weekday attempts. No interviews were attempted with teenagers whose parents or adult guardians had not given explicit consent for the adolescent interview.

Analysis Plan

To answer the first research question, comparing the feasibility of alternative protocols, response rates by adolescents and parents were calculated for each treatment protocol. These rates were calculated as the proportion of the eligible sample responding; sample members with incorrect contact information were assumed eligible (American Asscociation for Public Opinion Research, [AAPOR], 1998).

The basic analysis for the other two experiments was to compare the distribution of responses for the groups of interest. In the parent/teen response comparison experiment, adolescent responses from the dual-mode protocol were compared with parent responses from the control group, in which parents were proxy respondents for their children. For this experiment, responses to 315 adolescent interviews were compared with those of 369 parent interviews.

To compare teen responses by mode of administration, the responses from adolescents in the mail-only sample were compared with those from teens exposed only to the telephone protocol. There were 196 responses by mail and 194 by telephone available for this analysis.

Results

Returns by Mode

As can be seen in the table of response rates (Table 2), neither the telephone-only nor the mail-only protocol proved superior in obtaining responses from both parents and adolescents. Response rates by telephone (35%) and by mail (33%) were not significantly different. The parents were better responders by telephone than by mail, but this effect disappeared when attempts were made to complete the adolescent half of the interview pair.

Employing a dual-mode protocol was more productive than either of the single-mode approaches. When mail nonresponders were offered a telephone interview, returns were nearly 8% higher than the mail-only approach and almost 6% higher than with the telephone-only approach. The mail portion of the dual-mode approach yielded about the same results as the mail-only protocol, but giving nonrespondents the chance to complete a telephone interview brought the dual-mode response rate up to just over 40%. Complicating the survey process by asking both parents and teens to report on the adolescents' experiences yielded about 7% fewer responses than when just the parents were asked to respond.

Parental denial of permission for adolescent telephone interviews did not prove to be much of a problem. Overall, the parental permission denial rate was 2%. The telephone refusal rates were similarly low; about 5% of parents and about 6% of adolescents refused to be interviewed.

Ineligible cases were identified primarily through telephone efforts. The main reason for ineligibility was that the teenager was no longer enrolled in Medicaid; this was the case for 61% of those not eligible. Another 17% of the ineligibles were institutionalized in a residential treatment facility, group home, or correctional facility. Fourteen percent no longer lived with their parents, and there were indications that at least a few of these adolescents were in group homes or foster care. A few teens were ineligible by reason of age (6% of those not eligible); some had aged out of the eligible range, and others were under 13 (apparently they had incorrect birth

Table 2. Response rates for the MassHealth teen member survey: Overall and by mode

	Initial Sample (*n*)	Completed Interviews (*n*)	Ineligible Sample[1] (*n*)	Refusals (*n*)	Parental Permission Denied (*n*)	Incorrect Contact Info (*n*)	Other Nonresponse[2] (*n*)	Eligible Sample (*n*)	Response Rate (%)
Sample 1—Telephone Mode, Known Contact Information									
Parent	600	239	37	44	—	224	56	563	42.5
Teen	600	194	37	56	22	224	67	563	34.5
Both	600	194	37	56	22	224	67	563	34.5
Sample 2—Mail Mode, Known Contact Information									
Parent	600	216	1	0	—	45	338	599	36.1
Teen	600	203	1	1	—	45	350	599	33.8
Both	600	196	1	0	—	45	358	599	32.7
Sample 3—Dual Mode, Parents and Teens									
Parent—mail	800	270	0	0	—	52	478	800	33.8
Parent—phone	486	87	19	21	—	299	60	467	18.6
Parent—total	800	357	19	21	—	351	52	781	45.7
Teen—mail	800	260	0	1	—	52	487	800	32.5
Teen—phone	485	65	19	26	5	306	64	466	13.9
Teen—total	800	325	19	27	5	358	66	781	41.6
Both—mail	800	249	0	0	—	52	499	800	31.1
Both—phone	486	653	19	26	5	306	65	467	13.9
Both—total	800	314	19	26	5	358	78	781	40.2
Sample 4—Dual Mode, Parents Only									
Parent—mail	800	289	3	0	—	51	457	797	36.3
Parent—phone	455	75	32	27	—	283	38	423	17.7
Parent—total	800	364	35	27	—	334	370	765	47.6

[1] Because the sampling unit was the adolescent MassHealth member, study eligibility was based on the teen's status.
[2] The "Other nonresponse" category includes illness, language difficulties, contact limitations, and in the case of mail responses, failure to complete the correct survey.
[3] Three of the parents completed a mail questionnaire, while the teens responded by telephone.

dates recorded in Medicaid records). In the remaining 2% of the ineligible cases, the adolescent was reported to be developmentally delayed and unable to complete the questionnaire.

Mailing two questionnaires to a household created certain respondent identification difficulties, evidenced by the wrong respondent completing a questionnaire. In 11 cases it appears that the sampled teenagers were parents themselves and they completed the parent questionnaire about their own children. In another 10 cases the adolescent instrument was completed by someone other than the sampled teen. For the most part, this appeared to be parents filling out the questionnaire for their teenagers; in other cases a teenager other than the sampled adolescent filled out the questionnaire, and in a couple of cases, teens in a household completed the instrument as a group. In all, 0.6% of the parent returns and 1.5% of the adolescent returns involved an incorrect respondent. In the telephone interviews, where an interviewer was available to help sort it out, incorrect respondent identification was not an issue.

Respondent Characteristics

Looking at responses to the dual-mode test, adolescent respondents were about evenly divided by gender (52% female), but girls tended to be more likely than boys to respond by telephone (55%), though not at a significant level. Most (about 54%) were white, but this was a diverse group, with about 12% African American, 7% Asian, 5% American Indian, and 28% listed themselves as "Other Race" (respondents were instructed to select all applicable categories). In response to an additional ethnicity question, about 35% self-identified as Hispanic. The opportunity to respond on the telephone increased the response rate for all groups, but this was especially true for white, black, and Native American teens, whereas Hispanic and Asian teens tended to respond by mail. Predictably, the phone mode also increased response rates in households where parents had higher levels of education. Table 3 presents adolescent respondents' characteristics by mode.

Comparisons of Teen Responses by Mode

It is a recurring finding that results obtained on the telephone are more positive than those from mail surveys and that the differences often have to do with self-descriptions (Dillman, Sangster, Tarnai, & Rockwood, 1996). The CAHPS items do not have a large social desirability component, and modal differences in previous studies have proved to be minimal.

Responses obtained from teens in the mail- and telephone-only samples were compared to learn whether there were any effects by mode of administration. For most items there were no differences, but 5 of 33 items demonstrated a significant difference. Counter to what might be expected from a social desirability explanation, in two of the three questions for which there was a positive direction, the teens responded more positively by mail than by phone. These questions asked how much of a problem it was to get necessary care and whether the teens were able to get appointments for regular or routine care as soon as desired. Adolescents, however, were more likely to report by mail than by phone that they always had to wait in a doctor's office more than 15 minutes for an appointment, and that they have a personal doctor. Table 4 outlines results of the mode test by question type.

It is also worth noting that in another four items, differences between teen responses by mode approached significance ($p < .10$). In reponse to a provider interaction question that asked whether doctors discuss how the child is feeling,

Table 3. Adolescent respondent characteristics in dual-mode experiment by mode of administration

	Mail (% of all mail)	Phone (% of all phone)	Total % (*n*)	*p**
Gender				ns
Male	49.2	44.6	48.2 (150)	
Female	50.8	55.4	51.8 (161)	
Hispanic	37.3	23.8	34.5 (102)	<.05
Race				<.05
White	50.2	67.7	53.8 (169)	
Black	11.6	15.4	12.4 (39)	ns
Asian	7.6	3.1	6.7 (21)	ns
Native Am.	3.6	9.2	4.8 (15)	.059
Other race	28.1	21.5	26.7 (84)	ns
Age	x = 14.88	x = 14.83	x = 14.87 (310)	ns
Parent education				<.001
<8th grade	17.5	11.3	16.3 (49)	
Some HS	23.3	9.7	20.6 (62)	
HS	35.8	46.8	38.2 (115)	
Some college	21.3	19.7	20.9 (63)	
College grad	0.4	8.2	2.0 (6)	
Grad work	1.7	3.3	2.0 (6)	

**p* calculated by chi-square test for all but age, where *t*-test comparing means was used.

Table 4. Summary of comparisons by type of question: Adolescent mode test

	Teens by Mode		
Question Type	**Same**	**Different**	**Total**
Provider interaction			
Screening	3	1	4
Substantive	13	1	14
Health status			
Screening	1	0	1
Substantive	5	0	5
Utilization			
Screening	1	0	1
Substantive	2	1	3
Office-related			
Screening	1	0	1
Substantive	2	2	4
Total	28	5	33

Table 5. Items demonstrating significant differences in adolescent responses by mode

Item	Mail	Phone	*p*	*n* (mail/ phone)
Provider interaction				
Problem getting necessary care			.004	129/141
Not a problem	92%	79%		
Have a personal doctor (Screening question)			.009	183/191
Yes	81	69		
Office-related				
Get appointment as soon as wanted			.033	103/107
Always	48	37		
Wait in office 15 minutes or more			.029	132/141
Always	23	14		
Utilization				
Number of ER visits			.006	190/191
None	78	68		
1	16	16		
2	4	9		
3	2	2		
4	0	2		
5–9	0	2		
>10	0	1		

growing, or behaving, teens tended to respond more positively by mail than by phone. Two health status questions (whether the teen sees a doctor more than twice for a condition, and whether the teen takes prescription medicine regularly for a condition) and the global rating of health care were also nearly significant. Tables 5 and 6 present adolescent responses to items demonstrating significant and nearly significant differences by mode.

Comparisons of Parent and Teen Responses

For more than three-quarters (79%) of the 33 items asked of both respondents, there were no significant differences between parents' and adolescents' answers (Table 7). In 3 of the 7 items where differences appeared, the questions centered on the patient-doctor relationship: how often doctors talk with teens about how they are feeling, growing, or behaving; how often doctors explained things in a way the adolescent could understand; and the rating of the personal doctor. The other four questions addressed issues on which the parent could be expected to be a better informant than the child; two asked about phoning the doctor's office for advice during office hours (one a screening question, the other substantive), another about getting an appointment as soon as desired, and the last about taking prescription medicine regularly for a condition.

Screening questions allow for the identification of respondents for whom subsequent target questions apply; not all

Table 6. Items demonstrating nearly significant differences in adolescent responses by mode

Item	Mail	Phone	*p*	*n* (mail/ phone)
Provider interaction				
Doctors talked about feeling, growing, behaving			.061	155/130
Always	53%	44%		
Health status				
Take Rx meds regularly for condition			.062	44/36
Yes	73	53		
Seen doctor at least twice for condition			.056	44/37
Yes	79	60		
Rating of all health care	$\mu = 8.16$	$\mu = 8.53$	.063	129/141

Table 7. Summary of comparisons by type of question: Parent vs. teen responses

	Parent/Teen Responses		
Question Type	Same	Different	Total
Provider interaction			
Screening	3	1	4
Substantive	10	4	14
Health status			
Screening	1	0	1
Substantive	4	1	5
Utilization			
Screening	1	0	1
Substantive	3	0	3
Office-related			
Screening	1	0	1
Substantive	3	1	4
Total	26	7	33

questions apply to all respondents. Teens were far less likely than parents to report that they had called the doctor's office for help (44% versus 25%), thus limiting the number of responses to the substantive item that asks about how often that help was provided. See Table 8 for a comparison of adolescent and parent responses to significantly different items.

Discussion

Incorrect or inadequate respondent contact information drove response rates down. Many (about 40%) of the original records were missing either addresses or telephone numbers. It is likely that the response rates in the single-mode studies would have been lower if sampling had not been restricted to

Table 8. Items with significantly different parent and adolescent responses

Item	Parent	Teen	*p*	*n* (parent/ teen)
Provider interaction				
Doctors talked about feeling, growing, behaving			.004	189/229
Always	42%	54%		
Call for advice during office hours (Screening question)			.000	354/308
Yes	44	25		
Get phone help during office hours			.007	155/77
Always	74	52		
Doctors explain things to teen			.002	181/195
Always	67	52		
Rating of personal doctor (0–10 mean scale)	8.96	8.59	.032	251/229
Office-related				
Get appointment as soon as wanted			.002	194/158
Always	59	45		
Health status				
Take R_x meds regularly for condition			.002	85/70
Yes	72	64		

cases with complete contact information in Medicaid records. Methods that were employed to locate respondents included use of a computerized telephone number and address look-up service; requests for address correction and forwarding by the Postal Service; calls to directory assistance; and mailing postcard requests for telephone number updates to cases for whom we had addresses but no telephone numbers. Even after these extensive efforts to obtain current information, we were unable to get good contact information for nearly 45% of the dual-mode test sample.

Another way to think about outcome rates is to calculate the rate of cooperation. This is the proportion of all eligible units ever contacted who responded (AAPOR, 1998). The cooperation rate for the dual-mode experiment with adolescents and their parents responding was about 75%. This compares favorably with cooperation rates we observed in a privately insured sample of teenagers, where cooperation rates were about 82% by mail and 73% by telephone (Gallagher & Fowler, 1998). Teenagers enrolled in Medicaid and their parents proved to be about as willing to complete questionnaires about the adolescents' health care as families with private health insurance. However, in both cases, population mobility and the quality of the contact information provided by the sponsoring agencies greatly hampered efforts to reach respondents. This was particularly true in the Medicaid population.

While the single-mode approaches yielded about the same response rates, the telephone mode was better for obtaining explicit informed consent from both parents and adolescents.

In the dual-mode sample, more than 15% of the teens and more than 20% of the parents chose to respond in Spanish. It is unlikely we would have achieved the reported response rates without the use of dual-language instruments.

While the CAHPS instruments have demonstrated minimal mode effects in samples of privately and publicly insured adults (Fowler, Gallagher, & Nederend, 1999), mode effects among these adolescents enrolled in Medicaid are more difficult to explain than those seen in adults. The patterns do not fit previous research; here, many significantly different answers were more positive when collected by mail. Although it is not possible to sort this out fully, we can say that it is feasible to administer the instrument to adolescents using a mixed-mode protocol. However, the mode implications are not clear cut; there are some differences, but not many, and the effects that do emerge are somewhat counterintuitive. It may be that differences in the characteristics of those most likely to respond by mail (e.g., Hispanics, and households where parents had lower levels of education) contribute to the observed differences.

The mixed-mode protocol clearly improved response rates but still did not bring overall response to a satisfactory level. The data suggest, however, that if potential respondents can be reached, a dual-mode design is a reasonable strategy.

There are always tradeoffs to be made when making study design decisions. It is not clear cut that collecting data directly from adolescents is preferable to asking parents to be proxy reporters; the results were not strikingly different between these groups. There are a few items where the data differ, and for certain research purposes it may be worthwhile to get some information from teens directly. In other cases, it is debatable whether the parent or the teen is the most appropriate respondent. For items such as those that ask about making appointments, the rating of the personal doctor, or getting advice during office hours, the question of to whose standards health plans should be held accountable is worth considering.

Although self-reports are preferable in general to proxy reports, when decisions about the design of surveys of adolescent health care are being made, it is worth weighing the additional costs, both financial and in response rates, associated with contacting two respondents per household.

References

American Association for Public Opinion Research. (1998). *Standard definitions*. Ann Arbor, MI.

Dillman, D. A., Sangster, R. L., Tarnai, J., & Rockwood, T. H. (1996). Understanding differences in people's answers to telephone and mail surveys. *New Directions for Evaluation, 70,* 45–61.

Fowler, F. J., Jr., & Gallagher, P. M. (1997). Mode effects and consumer assessment of health plans. In *Proceedings of the Survey Methods Section* (pp. 928–933). Virginia: American Statistical Association.

Fowler, F. J., Jr., Gallagher, P. M., & Nederend, S. (1999). Comparing telephone and mail responses to the CAHPS Survey Instrument. *Medical Care, 37*(3), MS41–MS49.

Gallagher, P. M., & Fowler, F. J., Jr. (1998). Collecting information about the health care experiences of adolescents. In *Proceedings of the Survey Methods Section* (pp. 878–882). Virginia: American Statistical Association.

Hess, J., Rothgeb, J., Zukerberg, A., Richter, K., Le Menestrel, S., Moore, K., & Terry, E. (1998). *Teens talk: Are adolescents willing and able to answer survey questions?* Paper presented at the Annual Conference of the American Association for Public Opinion Research, May 14–17, 1998, St. Louis, MO.

Pamuk, E., Makuc, D., Heck, K., Reuben, C., & Lochner, K. (1998). *Socioeconomic status & health chartbook: Health, United States, 1998.* Hyattville, MD: National Center for Health Statistics.

Stussman, B. J., Willis, G. B., & Allen, K. F. (1993). Collecting information from teenagers: Experiences from the cognitive lab. In *Proceedings of the Section on Survey Research Methods* (pp. 382–385). Virginia: American Statistical Association.

Improving Adolescent Health Care Surveillance

Jonathan D. Klein, Caryn A. Graff, John S. Santelli, Marjorie J. Allan, and Arthur B. Elster

Background

Adolescent preventive services guidelines recommend confidential, comprehensive screening and counseling (Elster & Kuznets, 1994; U.S. Preventive Services Task Force [USPSTF], 1996; Green, 1994). Adolescents also face substantial barriers to receiving quality health care, but are rarely asked about their access or about the content of their own care (Klein, Wilson, McNulty, & Scott-Collins, 1999). Current public health surveillance and managed-care quality assurance methods rely on parent report of adolescent care, chart reviews, or administrative databases (Centers for Disease Control and Prevention [CDCP], 1995; Department of Health, Education, and Welfare [DHEW], 1974; Vistnes & Monheit, 1997). However, guidelines for adolescent care recommend confidential discussion of sensitive issues, including sexuality, reproductive health, substance use, mental health, and abuse. Parent report or chart documentation may not accurately reflect the care delivered. Physicians overestimate their delivery of preventive services in surveys and often do not document all of their interactions in charts (Lewis, Clancy, Leake, & Schwartz, 1991; Gemson & Elinson, 1986). Preventive visits may also be more accurately remembered by youth than by providers.

Current surveillance methods for health behaviors rely on adolescent report (Kann et al., 1993, 1998; Brener, Collins, Kann, Warren, & Williams, 1995). To know whether recommended services have been delivered and to improve preventive services for youth, accurate surveillance tools for assessing the content and quality of health services are needed. This paper reports on two studies that assess the validity and reliability of adolescent self-report about their receipt of preventive health screening and counseling services. Additionally, we will explore the implications of our findings for managed-care quality assurance and for public health surveillance activities designed to improve the health of adolescents.

Jonathan D. Klein, Caryn A. Graff, and Marjorie J. Allan are at the Division of Adolescent Medicine, Strong Children's Research Center, Department of Pediatrics, University of Rochester School of Medicine, Rochester, New York; John S. Santelli is at the Centers for Disease Control and Prevention; and Arthur B. Elster is Director, Clinical and Public Health Practice Outcomes, at the American Medical Association.

This research was supported in part by the Centers for Disease Control and Prevention, by the Robert Haggerty Fund, and by R01-HS 08192 from the Agency for Health Care Policy and Research. Dr. Klein is also supported by a Generalist Faculty Scholars Award from the Robert Wood Johnson Foundation.

Methods

Study 1: Validity

A convenience sample of 14- to 21-year-old adolescents were recruited at the time of their preventive care visits, defined as any regular nonacute health care visits; school, sport, or camp physicals; or reproductive health checkups, including prenatal visits. Adolescents were recruited from 15 community-based primary care practices in Monroe County, New York, including 7 pediatric and 3 family medicine suburban private practices, 2 teaching hospital clinics, and 3 urban community health centers. Clinical sessions were monitored for eligible patients, with systematic sampling from all sessions of each provider's practice.

A research assistant approached adolescents in the waiting room, determined eligibility, explained the study, and obtained informed consent from both parents and adolescents or from mature minors >17 years who were seeking confidential/protected services.

To audiotape visits, the research assistant accompanied the adolescents to the exam room with the recorder, and the adolescent was instructed on how to start the audiotape when the provider entered the exam room. Discussion that occurred outside the room was not captured on tape. If the clinician or the adolescent chose to stop the tape for part or all of an interview, these visits were excluded from analysis. Adolescents were randomly assigned to early and late follow-up groups and were surveyed by phone, either 2–4 weeks after their visit or 5–7 months after their visit, about their use of and access to care, as well as about the content of their most recent preventive visit (the "index visit," for this study).

Audiotapes were coded to assess delivery of 33 specific preventive service content areas identified from the CDC/AMA Guidelines for Adolescent Preventive Services (GAPS) (Table 1). Two trained research assistants listened independently to each tape, coding for discussion of each content area. Intraobserver reliability was assessed using Cohen's kappa, which accounts for the agreement between observations due to chance (Landis & Koch, 1977).

The audiotape coding for whether a topic was discussed was used as the gold standard for defining whether a counseling or screening service had been provided. If both of the

audiotape coders agreed that a topic was addressed, the content area was coded as discussed. The proportion of disagreement between the two raters ranged from 1.1% for having discussed anabolic steroids to 25% for having discussed an adolescent's friends (Table 1). Recoding disagreements between raters to either "yes" or "no" codes for whether a topic was discussed, or treating disagreements as missing data, had little or no effect on the sensitivity and specificity of the audiotaped gold standard compared to adolescent telephone interviews (data not shown). Because there were virtually no differences in the magnitude of agreement regardless of the method for treating discordant coding, results are presented with the unresolved cases treated as missing and excluded from the analysis in the interest of space.

Chart reviews were used as the gold standard for determining whether a physical examination or lab test had been provided during the visit, because these procedures were less likely to have been captured on the audiotape. Each chart was reviewed by two coders, and consensus interpretations were assigned, with a third coder mediating any coding disagreements.

Table 1. Inter-rater reliability (Cohen's kappa)/ percent disagreement for content coding of audiotaped visit

Topic	Kappa	% Disagreement
Weight	0.85	6.1
Blood pressure	0.67	15.7
Cholesterol	0.82	2.7
Immunizations	0.84	5.6
Diet	0.81	8.0
Body image	0.62	14.9
Exercise	0.74	12.8
Sleep	0.85	6.7
Teeth	0.82	8.8
Seatbelt	0.90	4.8
Bike helmet	0.89	5.3
Fighting	0.73	6.9
Violence	0.73	6.9
Weapons	0.83	3.2
Cigarettes/smoking	0.94	1.6
Chewing tobacco	0.72	4.0
Alcohol	0.89	3.7
Drugs	0.81	8.0
Steroids	0.74	1.1
OTC Drugs	0.38	20.3
Sex	0.80	5.3
Sexual orientation	0.20	23.5
Birth control	0.75	12.0
Condoms	0.91	4.5
HIV	0.83	7.5
STDs	0.81	9.1
Friends	0.45	27.5
School	0.77	8.3
Family	0.73	13.6
Future plans	0.75	12.5
Suicide	0.79	4.0
Abuse	0.70	3.2
Confidentiality	0.91	4.3

Study 2: Reliability

Test-retest reliability was assessed using a paper-and-pencil school survey method similar to the Centers for Disease Control and Prevention's Youth Risk Behavior Surveillance System. A trained research assistant administered surveys to students in 9th- through 12th-grade English and Health classes in one high school in New York State, with an interval of 14 days between administrations. Parental consent was obtained for adolescents through a mailing that explained the study. Adolescents who chose not to complete the survey were given an alternate activity by their teacher. An anonymous student-generated unique identifier was used to link time 1 and time 2 surveys.

The survey included 91 items, which assessed lifetime, current, 12-month, 30-day, and 1-week self-reported health risks and protective behaviors and 12-month recall of the screening and counseling services received. In addition, the survey addressed the age at each adolescents' first encounter with certain risk behaviors. Agreement between time 1 and time 2 responses were assessed using Cohen's kappa. Median kappa values were used to compare agreement between types of different questions. Multiple linear regression was used to evaluate two models testing the influence on reliability of (1) individual adolescent factors (age, gender, or ethnicity) and (2) question item characteristics (item prevalence, sentence complexity, time frame, and question type).

The study protocols were approved by the University of Rochester Research Subjects Review Board and the Institutional Review Board at the Centers for Disease Control and Prevention. With IRB permission, waiver of documentation of consent was allowed for the reliability study.

Results

Validity

Of 561 eligible adolescents seen for preventive care visits during monitored sessions, 537 (96%) were approached, and 401 (75%) consented and enrolled in the study. After having their visit audiotaped, one participant dropped out of the study. Complete audiotapes were successfully obtained from 374 visits (94% of enrollees), and 354 subjects (89%) completed subsequent telephone interviews. Half (180) of the final sample were interviewed between 2 and 4 weeks of their visit (90% completion rate), and the other half (174) were interviewed 5 to 7 months after their visit (87% completion rate). Chart review data was obtained for all 400 adolescents who completed enrollment in the study.

Seventy-five percent of the adolescents who participated were white, and 59% were female. The mean age of participants was 16 years (S.D. = 1.67 years). There were no differences in gender, age, or ethnicity between adolescents who chose to enroll and those who refused participation.

Intraobserver reliability (Cohen's kappa) between raters ranged from 0.20 for discussing sexual orientation to 0.94 for discussing tobacco (Table 1), reflecting fair to excellent agreement for most items (Landis, & Koch, 1977). Only three items (discussing over-the-counter (OTC) drug use, sexual orientation, and friends) had kappas of 0.45 or less.

Visits and Utilization

Almost all adolescents surveyed (94%) remembered having had a preventive care visit on or near the index visit date. Adolescents interviewed early were more likely than adolescents interviewed late to remember the exact date of their visit (20% vs. 3%; $p < .0001$), and gave a smaller range of possible visit dates. Adolescents interviewed early were also more likely than adolescents interviewed late to identify the date of their visit within a week (76% vs. 24%; $p < .0001$). Most adolescents (94%) accurately identified the site of care delivery, and (84%) identified the clinician they had seen. There were no differences between those interviewed early and those interviewed late in their ability to identify their clinicians and site of care.

Screening and Counseling Prevalence and the Validity of Adolescent Report

The prevalence of screening and counseling during these preventive health care visits, based on coding of all tapes (early and late), ranged from 2% for discussing anabolic steroids to 86% for discussing sex. Adolescents' report was most sensitive for anabolic steroid use, family issues, cigarettes and smoking, exercise, school performance, and physician-patient confidentiality. For items with the highest sensitivity by self-report, the 2- to 4-week group was slightly more accurate than the 5- to 7-month group in each category (data not shown) (Klein et al., 1999).

Examination Validity

Based on chart review data as a gold standard for physical examination and lab procedures, the most often documented examinations included heart (84%), ears (85%), height (86%), and weight (96%). HIV testing (5%), MMR immunizations (4%), urine culture (3%), and drug testing (0%) were least often provided (Table 2). Adolescents also were most likely to report having had their height, weight, and blood pressure measured; having received an immunization (usually a hepatitis B shot, a tetanus shot, or both); and that their ears, heart, lungs, or testes were examined.

Among the 2- to 4-week follow-up group, self-report sensitivity ranged from a low of 5% for having a urinalysis to 100% for having height and weight measurements, a Pap smear, or an HIV test (Table 3). Sensitivity for the 5- to 7-month follow-up group ranged from 4% for a urinalysis to 100% for chlamydia, gonorrhea, and cholesterol testing.

Table 2. Frequencies of adolescent self-reported receipt of a physical examination or procedure compared to chart documentation

	Early Interview		Late Interview	
Examination of:	**Phone %**	**Chart %**	**Phone %**	**Chart %**
Exam				
Weight	99	97	97	95
Blood pressure	98	34	94	29
Height	97	86	94	87
Testes	94	33	96	34
Heart/Lungs	92	84	93	84
Ears	90	86	89	85
Breast	29	35	26	40
Pelvic	24	13	13	11
Lab tests				
Blood test	33	31	19	33
Pregnancy test	20	6	14	5
Pap smear	13	4	9	9
Cholesterol test	13	6	12	5
HIV test	12	6	3	3
TB test	12	7	20	8
Gonorrhea test	11	8	10	8
Chlamydia test	9	8	9	8
Urinalysis	2	63	1	58
Drug test	1	0	2	0
Urine culture	0	3	2	4
Immunizations				
Hepatitis B	70	49	72	49
Immunizations	62	57	45	56
Tetanus	38	19	33	16
MMR	10	6	9	3
Overall Median	24	19	19	16

For the 2- to 4-week follow up group, the specificity of self-report ranged from a low of 1% blood pressure measurement to 100% for having a urine culture (Table 3). Specificity for the 5- to 7-month follow up group ranged from 5% for having blood pressure measured to 100% for a urinalysis. Both early and late groups were least specific at reporting whether they had heart and lung exams or height and blood pressure measurements. The early interview group also was not very specific at reporting testicular examination. For the early group, reports of procedures such as urine cultures, drug testing, pregnancy testing, MMR immunizations, and HIV testing had the highest specificity.

Reliability

In the reliability study, 296 (87%) of 339 eligible adolescents were present at time 1, and 293 (99%) of these completed surveys; 253 (86%) of these adolescents were present in class and completed surveys 2 weeks later at time 2.

Eighty-nine percent of the adolescents who completed the pencil-and-paper survey were white, and 52% were female.

Table 3. Sensitivity/specificity: Adolescent self-report of discussion with health care provider compared to chart data from the encounter

Did your doctor examine / order:	Early Interview		Late Interview	
	Sensitivity %	Specificity %	Sensitivity %	Specificity %
Exam				
Weight	100	50	97	*
Blood pressure	95	1	98	5
Height	100	30	98	40
Testes	95	11	95	*
Heart/Lungs	97	40	98	33
Ears	99	57	99	52
Breast	36	72	38	79
Pelvic	52	84	62	94
Lab tests				
Blood test	78	88	42	93
Pregnancy test	78	98	75	94
Pap smear	100	92	71	96
Cholesterol test	88	95	100	95
HIV test	100	96	*	96
TB test	85	93	77	85
Gonorrhea test	89	92	100	92
Chlamydia test	89	95	100	93
Urinalysis	5	89	4	100
Drug test	*	99	*	98
Urine culture	*	100	*	94
Immunizations				
Hepatitis B	94	88	87	65
Immunizations	94	86	68	86
Tetanus	97	87	84	83
MMR	73	97	50	92
Overall Median	24	19	19	16

*Cell size too small to calculate

Participants' ages were 14–15 years (38%), 16 years (31%), and 17 years of age or older (31%). Forty-five adolescents (18%) had seen a clinician between the two survey administrations and were excluded from analyses for questions with which their responses might change because of the visit. There were no differences in gender, age, or ethnicity between the adolescents who had seen a provider between administrations and those who had not.

Reliability (Cohen's kappa) between time 1 and time 2 responses ranged from 0.94 for having a pelvic exam to 0.33 for having talked with their clinician about physical activity or exercise at their last visit (Table 4). Adolescents were most reliable in their report of having a pelvic exam (0.94), ever smoking (0.93), their height (0.93), and ever having sex (0.90). Whether or not they used smokeless tobacco in the past 30 days (0.39), the number of times they visited a source of care other than their primary care source (0.37), having discussed sexual orientation (0.34), and reported 7-day physical activity or exercise (0.33) were among the least reliable items.

Questions about adolescents' behaviors had a median kappa of 0.75. For questions measuring having received counseling or screening, the median was 0.63. Questions about adolescents' utilization of health services had a median kappa of 0.57. Questions assessing lifetime prevalence and reported age at initial behavior had median kappas of 0.79 and 0.78, respectively. Questions assessing current behaviors also had a relatively high median kappa, 0.70. The median kappa for 30-day recall questions was 0.65, while the median kappa was lower (0.53) for questions that asked about activities that had occurred in the past 12 months. Not surprisingly, questions that prompted recall within a week performed poorly, with a median kappa of only 0.35.

In the multiple regression analyses, with agreement as the dependent variable, neither age nor ethnicity was significantly associated with adolescents' reliability either for reporting the counseling/screening they had received from their health care provider, or for reporting their behaviors; gender had a mild effect, with girls being slightly more likely than boys to report care reliably (Table 5). In contrast, question time frame ($\Delta R^2 = 0.18$), prevalence ($\Delta R^2 = 0.09$), type of question ($\Delta R^2 = 0.12$), and question complexity ($\Delta R^2 = 0.02$) were positively associated with reliability. The full model, assessing question complexity and other factors' association with kappa values (agreement), resulted in an R^2 of 0.54 (Table 5).

Table 4. Question type categories with agreement (Cohen's kappa) for each question and overall category medians

Behaviors: Median kappa 0.75		Counseling Median kappa 0.63		Utilization Median kappa 0.57	
Ever tried smoking	0.93	List items		At last visit did you have a pelvic exam	0.94
Self-report of height	0.93	At last visit did provider discuss birth control use	0.77	Tested for chlamydia in past 12 months	0.81
Ever had sex	0.90	At last visit did provider discuss bike helmet use	0.73	Ever had hepatitis b vaccine	0.76
Age first tried marijuana	0.89	At last visit did provider discuss alcohol use	0.72	Know of a place for confidential care	0.68
Self-report of weight	0.89	At last visit did provider discuss smoking or cigarette use	0.69	Injured while exercising and treated in past year	0.65
Age at first intercourse	0.87	At last visit did provider discuss condom use	0.68	What was your last vist for	0.65
Used condom last time had sex	0.86	At last visit did provider discuss family	0.67	Were you given forms at last visit	0.65
Considered suicide in past year	0.86	At last visit did provider discuss sex	0.66	When was your last routine visit	0.63
Ever tried quitting smoking	0.82	At last visit did provider discuss future plans	0.66	Number of times visited ED in past 12 months	0.62
Number of lifetime sexual partners	0.82	At last visit did provider discuss seatbelt use	0.65	Ever gone to provider without parent's knowledge	0.60
Days smoked in last 30	0.81	At last visit did provider discuss weight	0.63	Last time needed care where did you go	0.57
Birth control method used last time	0.80	At last visit did provider discuss street drug use	0.63	Did you see your regular doctor at last visit	0.54
Times used marijuana in last 30 days	0.80	At last visit did provider discuss HIV or AIDS	0.63	Number of times been to a provider or clinic in past 12 months	0.53
Ever smoked regularly	0.79	At last visit did provider discuss confidentiality	0.63	When was last visit to provider or clinic	0.51
Age at first cigarette	0.78	At last visit did provider discuss how you feel about your body	0.62	Do you go to one place for care	0.51
What are you doing about your weight	0.76	At last visit did provider discuss sexual or physical abuse	0.62	Have you been treated for suicide in past year	0.49
Attempted suicide in past year	0.75	At last visit did provider discuss ways to quit smoking	0.61	Have you been injured while at work and treated by provider in past year	0.48
Number cigarettes smoked per day in last month	0.74	At last visit did provider discuss chewing tobacco or snuff	0.61	Do you have a doctor to go to when sick	0.47
Ever talked about AIDS with parents/family adults	0.74	At last visit did provider discuss friends	0.61	Number of times been to one source of care in past 12 months	0.46
Age at first alcoholic drink	0.73	At last visit did provider discuss emotions or moods	0.61	When was your last visit to your one source of care	0.43
Number of days smoked cigars in last 30	0.72	At last visit did provider discuss healthy eating/diet	0.58	Times visited other sources of care in past 12 months	0.37
Age tried coke for first time	0.72	At last visit did provider discuss suicide	0.58		
Describe weight	0.72	At last visit did provider discuss school	0.57		
Times in a physical fight in past year	0.69	At last visit did provider discuss STDs	0.56		
Bicycle helmet use in past 12 months	0.65	At last visit did provider discuss physical activity or exercise	0.50		
Proofed when buying cigarettes in last month	0.65	At last visit did provider discuss use of steroid pills or shots	0.44		
Days in past 30 had an alcoholic drink	0.64	At last visit did provider discuss sexual orientation	0.34		
Days in past 30 had 5 or more drinks in a row	0.64	At last visit did provider discuss setting a date to quit smoking	–0.02		
Ridden w/driver who was drinking in past month	0.63				
Seatbelt use	0.58	Single items			
Driven a vehicle when drinking in past month	0.47	At last visit did provider discuss risks of STDs	0.76		
Engaged in a fight which required medical treatment in past year	0.43	At last visit did provider discuss HIV/AIDS	0.73		
Used snuff in past 30 days	0.39	At last visit did provider talk about cigarettes/smoking	0.68		
		At last visit did provider discuss condoms to prevent HIV/AIDS	0.64		
		At last visit did you talk privately with provider	0.64		
		At last visit did provider discuss BC to prevent pregnancy	0.62		
		Has your provider ever talked about quitting	0.60		
		At last visit did your provider talk about alcohol	0.56		
		At last visit did you talk with provider about confidentiality for teens	0.53		

Discussion

Our data suggest that adolescents' self-report of the care they have received is a valid and reliable method for determining the content of preventive health service delivery. In reporting about the care they had received five to seven months earlier during preventive care visits, most adolescents remembered having preventive care visits and identified their doctor and site of care.

Adolescents recall discussing steroids, confidentiality, school, exercise, family, and cigarettes/smoking with highest sensitivity and specificity. Other important issues such as alcohol

Table 5. Demographic and question format factors associated with adolescent's self-retest reliability

	Beta	Sig T
Model 1		
Age, ethnicity, sex = Agreement *		
$R^2 = 0.05$		
Age 14–15	0.03	0.66
Age 16	–0.10	0.15
Ethnicity	–0.03	0.64
Sex	–0.17	0.00
Model 2		
Prevalence, complexity, type, and time frame = Agreement (kappa)		
$R^2 = 0.54$		
Question type		
Behavior	0.43	0.01
Counseling	–0.17	0.40
Utilization	–0.07	0.62
Prevalence > 95%	–0.31	0.00
Question complexity	–0.18	0.02
Question time frame		
Age at	0.04	0.68
Current	–0.36	0.03
Ever	0.03	0.78
Month	–0.24	0.03
Week	–0.41	0.00
Year	–0.21	0.06

* Average reliability score computed for each student based on the percent agreement between the same questions at time 1 and time 2.

use, sex, and condoms were also accurately recalled most of the time. Adolescents also were able to report with validity those topics that were not discussed at their visits, including weapons, violence, abuse, bicycle helmets, and cholesterol.

Although we found chart documentation to be a good source of information about immunizations and some laboratory procedures, our data suggest that charts may result in both over- and underreport of the screening and counseling services actually delivered. We had trained clinicians review the charts in our study; however, we did not attempt to validate our interpretation further. Additionally, in a study examining office records as a source of ambulatory care information, 20% of records contained illegible terms or abbreviations interpretable only by the recording physician (DHEW, 1974). Adolescents also may report having discussed issues, even if the screening they received was done by paper-and-pencil survey.

Adolescents are also reasonably reliable in test-retest reporting of their health behaviors and of the screening and/or counseling services they have received. Reliability was good for most questions, regardless of respondent age, gender, or ethnicity. Question recall time span, complexity, and condition prevalence also significantly affect the reliability of adolescents' answers to various items.

Adolescent self-report of drug, tobacco, and alcohol use has previously been shown to have reasonable reliability (Needle, McCubbin, Lorence, & Hochhauser, 1983; O'Malley, Bachman, & Johnston, 1983; Martin & Newman, 1988). In a study using the 1992 Youth Risk Behavior Survey (YRBS) questionnaire, adolescents also reported on a variety of health risk behaviors with reasonable reliability (Brener et al., 1995). As in this earlier study, we also found that adolescents were most reliable in reporting on lifetime and current behaviors. While our study focused on only a subset of the YRBS questions, those questions regarding behaviors which we included performed with kappas similar to those identified by Brener.

In contrast to reports of health behaviors, this study is the first to examine reliability of health services use self-report by adolescents. Previous field tests of the National Health Information Survey examining the validity of self-reported medical care use by a household sample of adults found underreporting of health care encounters by 20%, and as many as 39% of adults incorrectly classify their usual source of care (Jobe et al., 1990; Perloff & Morris, 1989). Our sample was drawn from clinical sources, however, and not from the general population. Thus, our subjects responses about care use, while substantially better in accuracy, are not directly comparable to randomly selected respondents.

Adolescents are most reliable in reporting lifetime or current behaviors, compared to reporting behaviors over shorter recall periods. Both sentence complexity and time frame of recall have the greatest effects on the reliability of adolescent reports. However, adolescents also demonstrate reasonably high reliability for having received care and for having received screening or counseling for most preventive health services. The reliability of service use approaches (and, in the case of some content areas, exceeds) the reliability of behavioral self-report by adolescents. Self-report of utilization and of services received could be used to assess the content of primary care delivered to youth in quality measurement and/or public health surveillance systems.

Measuring Quality

Measurement of adolescent clinical preventive services as they have been received by adolescents has implications for assessing receipt of specific clinical preventive services in public health surveillance systems, for medical care quality assurance systems, and for health services research. The Health Plan Employer Data and Information Set (HEDIS) version 3.0 currently includes several measures for clinical preventive services, including one for an annual preventive care visit for adolescents (National Committee for Quality Assurance, 1996). This is similar to the periodicity of visits recommended by many of the guidelines for adolescent preventive care, including the *Guidelines for Adolescent Preventive Services, Bright Futures,* and the American Academy of Pediatrics (Elster & Kuznets, 1994; Green, 1994; American Academy of Pediatrics, 1996; USPSTF, 1996). However, to assess the quality of care provided, it is important to look at the content of care delivered, not just at utilization measures. For example, recent data suggest that

just over half of all adolescents had the opportunity to talk alone with their provider during health care encounters (Klein et al., 1999); one in three adolescents reported having missed needed care, most often due to confidentiality concerns. Adolescents are known to avoid care for sensitive issues unless their confidentiality is assured (Malus, LaChance, Lamy, Macaulay, & Vanasse, 1987). Each of the referenced guidelines above also call for specific screening and counseling interventions, most of which are also recommended both by the American Academy of Family Practitioners [AAFP] (1994) and by the U.S. Preventive Services Task Force (1996). While some specific screening and counseling interventions were provided during the preventive care visits we audiotaped, the prevalence of screening and counseling services in these visits fall far short of the care that is recommended for adolescents.

Initial findings from our study have led to adoption of several items by the Centers for Disease Control and Prevention (CDC) Youth Risk Behavior Surveillance System (YRBSS). The YRBSS is a biannual national school-based survey of adolescents; most states and several local areas also conduct separately sampled surveys. Two core items have been added to the YRBSS, and additional items are available in a supplemental module available to states and local areas. These items will assess when adolescents last had a care visit and whether they had received preventive counseling about HIV and STDs, or about tobacco use.

Additionally, our work has led to collaboration with the Foundation for Accountability and the National Commission on Quality Accreditation on their Child and Adolescent Health Measurement Initiative (CAHMI). The CAMHI is charged with developing quality measures for child and adolescent health care for use in quality assurance, and to help families, purchasers, and providers improve the quality of care. Questionnaire items from our study have been incorporated into the CAHMI's Adolescent Health Survey instrument, and initial field trials conducted in six managed-care plans in New York, California, and Florida. The goals of these field trials are to compare telephone versus paper and pencil method performance; to compare different case-finding strategies; and to assess the internal reliability of candidate quality measure performance values. In addition to these efforts, several of our health services receipt items have been incorporated into the Consumer Assessment of Health Plans (CAHPS) Adolescent-CAHPS pilot project with the Massachusetts Department of Medical Assistance

Limitations

The validity and reliability study are both limited by the representiveness of their samples, since both the clinicians and the adolescents who agreed to participate may be subject to selection bias and may not be fully representative of either clinicians' performance or all adolescents' recall. Additionally, the validity study is limited by the accuracy of the audiotape coding, both by not being able to see nonverbal communication between providers and patients, and by not being able to capture all of the patient and provider's interactions (for example, discussions on the way to the room or in the hall). Thus the tapes may underestimate the true rates of counseling or screening. The presence of the tape itself also may have affected both adolescents' recall and the content of the discussion. However, these effects likely would have resulted in increased delivery of recommended preventive services; thus, our observations may have inflated the usual performance of these clinicians.

Our reliability study is further limited in that we did not test multiple ways to ask about specific items, to confirm whether or not question structure rather than content affects response reliability. In addition, our results may not be fully generalizable, because our sample was from one high school in New York State.

Conclusion

Adolescent self-report may be a reasonably accurate source of information for public health surveillance and managed-care quality assurance systems about the content of health services adolescents have been provided. In fact, because many of the discussions during adolescent's visits are conducted privately between adolescents and their clinicians, adolescents may be a better source of some kinds of information than either their parents or their charts. Additionally, interviewing adolescents is the only way to assess the preferences of youth with regard to the care they receive.

Surveying adolescents via telephone and through pencil-and-paper surveys about the health services they have received is relatively valid and reliable and is of comparable accuracy to asking about adolescents' recent health behaviors. Use of adolescent self-reports of the content of primary care in managed-care quality assurance and public health surveillance systems has the potential to improve the quality of adolescent care. The questions also have implications for better quality improvement activities, for community needs assessments, for SCHIP evaluation, and for future research on preventive services delivery and health outcomes for adolescents. This study adds support to quality measurement strategies that seek to obtain data directly from youth.

References

American Academy of Family Physicians (1994). *Age charts for periodic health examinations.* Kansas City, MO.

American Academy of Pediatrics (1996). *Guidelines for Health Supervision III.* Elk Grove Village, IL.

Brener, N., Collins, J., Kann, L., Warren, C., & Williams, B. (1995). Reliability of the Youth Risk Behavior Survey Questionnaire. *American Journal of Epidemiology, 141*(6), 575–580.

Centers for Disease Control and Prevention. (1995). *Health risks in America: Gaining insights from the BRFSS.* Atlanta, GA: U.S. Dept. of Health and Human Services.

Department of Health, Education, and Welfare. (1974). *NAMCS: Background and methodology, U.S. 1967–1972.* Rockville, MD: U.S. Dept. of Health and Human Services.

Elster, A., & Kuznets, N. (1994). *Guidelines for adolescent preventive services.* Baltimore, MD: Williams and Wilkins.

Gemson, D., & Elinson, J. (1986). Prevention in primary care: Variability in physician practice patterns in NYC. *American Journal of Preventive Medicine, 2*, 226–234.

Green, M. (Ed). (1994). *Bright Futures, guidelines for health supervision of infants, children, and adolescents* (1st ed.). Arlington, VA: National Center for Education in Maternal and Child Health.

Jobe, J., White, A., Kelley, C., Mingay, D., Sanchez, M., & Loftus, E., (1990). Recall strategies and memory for health-care visits. *Millbank Memorial Fund Quarterly, 68,* 171–189.

Kann, L., Kinchen, S., Williams, B., Ross, J., Lowry, R., Hill, C., Grunbaum, J., Blumson, P., Collins, J., & Kolbe, L. (1998). Youth risk behavior surveillance—United States, 1997: State and local YRBSS coordinators. *Journal of School Health, 68* (9), 355–369.

Kann, L., Warren, W., Collins, J., Ross, J., Collins, B., & Kolbe, L. (1993). Results from the national school-based 1991 Youth Risk Behavior Survey and progress toward achieving related health objectives for the nation. *Public Health Reports, 108* (Suppl 1), 47–67.

Klein, J., Graff, C. A., Santelli, J. S., Hedberg, V. A., Allan, M. J., & Elster, A. B. (1999). Developing quality measures for adolescent care: Validity of adolescents' self-reported receipt of preventive services. *Health Services Research, 34,* 391–340.

Klein, J., Wilson, K., McNulty, M., & Scott-Collins, K. (1999). Gender issues in access to and use of health services by adolescents: The Commonwealth Fund survey of the health of adolescent girls. *Journal of Adolescent Health, 25,* 120–130.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lewis, C., Clancy, D., Leake, B., & Schwartz, J. (1991). Counseling practices of internists. *Annals of Internal Medicine, 114,* 54–58.

Malus, M., LaChance, P., Lamy, L., Macaulay, A., & Vanasse, M. (1987). Priorities in adolescent health care: The teenager's viewpoint. *Journal of Family Practice, 25* (2),159–162.

Martin, G., & Newman, I. (1988). Assessing the validity of self-reported adolescent cigarette smoking. *Journal of Drug Education, 18,* 275–284.

National Committee for Quality Assurance. (1996). *Health plan employer data and information set, 3.0.* Washington, D.C.: U.S. Dept. of Health and Human Services.

Needle, R., McCubbin, H., Lorence, J., & Hochhauser, M. (1983). Reliability and validity of adolescent self-reported drug use in a family-based study: A methodological report. *International Journal of the Addiction, 18,* 901–912.

O'Malley, P., Bachman, J., & Johnston, L. (1983). Reliability and consistency in self-reports of drug use. *International Journal of the Addictions, 18,* 805–824.

Perloff, J., & Morris, N. (1989). Validating reporting of usual sources of health care. In *Health Survey Research Methods: Conference proceedings* (pp. 59–64). U.S. Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment.

U.S. Preventive Services Task Force. (1996). *Guide to clinical preventive services: An assessment of the effectiveness of 169 interventions* (2nd ed.). Baltimore, MD: Williams and Wilkins.

Vistnes, J., & Monheit, A. (1997). *Health insurance status of the civilian noninstitutionalized population: 1996.* MEPS Research Finding No. 1, AHCPR Pub. No. 97-0030. Rockville, MD: Agency for Health Care Policy and Research.

Young Children's Reports of Their Health: A Cognitive Testing Study

Anne Riley, George Rebok, Christopher Forrest, Judy Robertson,
Bert Green, and Barbara Starfield

Introduction

Individuals are uniquely able to report on their own health experiences, and it is likely that children are no exception (La Greca, 1990). Moreover, assessments from multiple respondents are necessary to describe child functioning comprehensively and to predict their health outcomes effectively (Achenbach, McConaughy, & Howell, 1987; Hart, Lahey, Loeber, & Hanson, 1994). Despite the validity of self-report over proxy report and the value of multiple perspectives, when this work was begun, no instruments existed for capturing children's expressions of health and well-being in a systematic manner (Landgraf & Abetz, 1996). The challenge was to provide children themselves with a means for describing their physical and emotional well-being. This study is the first in a project to develop a generic pediatric health status questionnaire for elementary school-aged children.

Regardless of age, to complete a health questionnaire a person must at least have a rudimentary self-concept; understand the basic notions of health and illness; and be able to pay attention, comprehend the questions, discriminate between the response alternatives, recall health experiences, and write a response. These requisite skills guided our investigation.

Four study objectives were developed. First, using a cognitive testing methodology, we attempted to determine whether children from 5 to 11 years of age can answer health survey items. Second, in order to reduce the demand for literacy and ability to handle abstract concepts, we tested the feasibility of a pictorial questionnaire format using cartoon drawings of a "universal" child to illustrate key concepts. Illustrations have been used successfully in a number of child questionnaires (Breton et al., 1999; Fox & Leavitt, 1995; Harter & Pike, 1984; Raviv, Raviv, Shimoni, Fox, & Leavitt, in press; Valla, Bergeron, Bérubé, Gaudet, & St-Georges, 1994; Valla, Bergeron, Bidaut-Russell, St-Georges, & Gaudet, 1997) and have the advantages of maximizing attention to the task and minimizing reliance on younger children's limited vocabularies. Third, we examined several types and numbers of response formats to see which are most easily understood by young children and which they prefer. Finally, we tested children's understanding of specific concepts of their health and wording of the different response formats.

Literature Review

It is known that even children as young as 5 years old can describe internal mental states such as perceptions, emotions, cognitions, and physiological states, but we were unsure whether they could distinguish between different aspects of themselves (good at numbers, but poor at reading), expecting that they would show evidence of "all or none thinking" (Burbach & Peterson, 1986; Byrne, 1996; Harter & Pike, 1984; Stone & Lemanek, 1990).

Language mastery is likely to limit young children's ability to describe their health. Although even young children can respond to questions about pain (Ross & Ross, 1984; Harbeck & Peterson, 1992; McGrath et al., 1996) and nausea (Zeltzer et al., 1988), 5- and 6-year-old children give more variable and less discriminating responses than older children. It is also clear that children's understanding of health-related words, ability to understand complex sentences, and ability to comprehend and match verbally presented sentences with illustrations increase with age (Nelson, 1976; Stone & Lemanek, 1990) and that 5-year-olds are likely to use relational terms, such as "more/less" and "same/different," incorrectly (Donaldson & Wales, 1968).

In terms of their concept of health, children below age 8 were expected to view health in terms of specific health practices and to lack understanding that they could be partially healthy (Natapoff, 1978, 1982). We did not expect children below age 8 to understand that illness is defined by a set of concrete symptoms or to use internal cues to identify the presence of illness (Burbach & Peterson, 1986; Hergenrather & Rabinowitz, 1991; Neuhauser, Amsterdam, Hines, & Steward, 1978; Perrin & Gerrity, 1981). Over our entire age spectrum, children were not expected to be able to think logically about future health or to have a concept of mental health (Natapoff, 1978, 1982).

Multiple aspects of children's abilities rapidly increase with age (Gale & Lynn, 1972; Hagen & Hale, 1973; McKay, Halperin, Schwartz, & Sharma, 1994; Rebok et al., 1997;

The authors are at The Johns Hopkins University, Baltimore, Maryland. Anne Riley, Christopher Forrest, Judy Robertson, and Barbara Starfield are in the Department of Health Policy and Management, School of Public Health; George Rebok is in the Department of Mental Hygiene, School of Public Health; and Bert Green is in the Department of Psychology, School of Arts and Sciences.

This work was supported by the Agency for Health Care Policy and Research grant HS07045 awarded to Dr. Anne Riley. An AHCPR grant awarded to Dr. Barbara Starfield previously supported the development of the CHIP-AE.

Table 1. Demographics of each study sample

		Year of age *N*								
Study	**Total *N***	**5**	**6**	**7**	**8**	**9**	**10**	**11**	***N*/% Boys**	***N*/%Nonwhite**
1	35	11	8	4	5	3	3	1	18 (51%)	11 (69%)
2	19	2	6	5	2	1	2	1	11 (58%)	11 (58%)
3	60	7	12	16	9	6	8	2	35 (58%)	29 (48%)

Wechsler, 1974; Woodcock & Mather, 1990). In terms of recall, children are able to recall routine and novel events accurately for at least 24 hours at age 5 (Ornstein, 1995; Schwab-Stone, Fallon, Briggs, & Crowther, 1994) and can recall novel events for weeks by age 7 (Gathercole, 1998; Ornstein, 1995), but are not good at timing events until age 7 or older (Friedman, 1991).

Clinical experiments in medical settings suggest children find visual analogue scales engaging and understandable, at least for reports of pain intensity (McGrath, 1991; Ross & Ross, 1984). We found no questionnaires for children that use illustrated Likert response scales. Those with illustrations use dichotomous responses, and one uses two sets of dichotomous responses in order to obtain a 4-point scale (Harter & Pike, 1984). Moreover, there are no studies of the effects of age, gender, or race of the illustrated character on the quality of children's responses.

Thus, the literature supported the feasibility of developing a health questionnaire for children, although there were significant gaps about the optimal ways to ask children questions about their health. The content of the questionnaire drew from earlier work conducted by the investigators on the Child Health and Illness Profile—Adolescent Edition (CHIP-AE; Starfield et al., 1993, 1995). In order to support longitudinal assessments of health status from childhood through adolescence, the child version of the CHIP uses the same structure as the CHIP-AE. The CHIP-AE is a self-administered health status measure that adolescents aged 11 through 17 years complete. It comprises 6 domains (Satisfaction with Health, Discomfort, Risks, Resilience, Disorders, and Achievement) and 20 subdomains that were conceptually derived and supported by factor analysis.

General Method

The cognitive testing studies were undertaken with convenience samples of children 5–11 years old, focusing on the optimal ways to ask children questions about their health; the most easily understood response formats; children's understanding of health concepts; and ability to utilize different response options. Parents of children in day care or after-school programs were asked by the day care providers to sign a consent form, which explained that the assessment would be audiotaped and which included several examples of the items to be asked. All children of consenting parents were interested and signed an assent form after the study had been explained to them. The study protocol was approved by the Johns Hopkins institutional review board. All three studies involved administration of the items to each child individually, by trained interviewers, typically in a large classroom after the end of the school day.

A total of 114 children were assessed. The majority of children were African-American, and over one-third were white, non-Hispanic. All children in the first two studies were recruited from three after-school day care programs in the residential areas of Baltimore City that serve low- to middle-income families. In the third study, two-thirds were from one of the after-school programs and 18 were from medical clinics at Johns Hopkins University that serve children with chronic conditions. Although data were not kept on the number of refusals, day care providers reported a very good response to their requests for participation, probably because parents were only required to provide consent. Clinicians also reported a good response, although time constraints associated with the medical appointment prevented participation by some youth. It appeared that the samples were representative of the settings from which they were recruited. Table 1 provides a summary of the age, gender, and race distribution of the sample for each study. The methods used to assess children's comprehension and performance were based on "think-aloud" methodology (Ericsson & Simon, 1993).

Study 1: Questions, Methods, and Results

Study 1 Questions

How can health questions be asked in a way that engages children's interest and focuses their attention?

The specific questions were: (1) Can children translate the intensity of their preferences and frequency of behavior into a scaled response? (2) Can a character be developed with whom most children can easily identify?

Our intent was to illustrate a character with whom children could identify and who represented what the "healthy" and "unhealthy" child at each end of the response scale would experience. The illustrations were drawn by a professional cartoonist, with input and feedback from the investigative team. To avoid problems with having multiple characters and test versions, we aimed to develop a "universal character" that would be age, gender, and race neutral.

Study 1 Methods

Sixteen simple items representing common tasks or activities (e.g., eating ice cream, cleaning one's bedroom) were developed so that we could evaluate children's answers with-

out concern about their ability to understand item content. Children were asked a series of 16 simple questions to see whether they could use the VAS. As an example, for the item "How much do you like ice cream?" the child was instructed to mark an "X" on the line indicating how much they liked ice cream by the distance from the ice-cream-eating child illustration and the non-ice-cream-eating child illustration. The response options ("a lot," "somewhere in between," "not at all") were read to the child but no response labels appeared on the VAS (see Figure 1). We also asked questions designed to assess children's ability to recall events occurring over a period of time (e.g., "How many days have you eaten ice cream in the past week?"). Finally, children were asked whether they thought the illustrated child character was sort of like them, and why; whether the character was a girl or a boy; and whether the character was the same age or younger or older than they were.

Study 1 Results

In general, even children as young as age 5 seemed to be able to use the VAS. They used both ends and the middle of the line. The 5-year-old children appeared to understand the content of the questions, but they had difficulty understanding the concept of a "week." Children younger than 8 were very concrete in their identification with the character. For example, when asked "Is this child sort of like you?" several children responded "No, because she has only four fingers," and, to "Do you think this kid could be your friend?" one 5-year-old girl replied "Yes, I could like someone with only four fingers." In subsequent versions of the questionnaire, the cartoon character had the appropriate number of digits. The gender of the character was identified, for the most part, as a girl by the girls and as a boy by the boys. They typically identified the child character as being within one year of their own age. No children indicated that any of the characters were different from themselves based on race, facial features, or hair.

Study 2: Questions, Methods, and Results

Study 2 Questions

What response formats are most easily understood by children?

Three questions were posed: (1) Is a straight-line visual analogue scale (VAS) more easily used and understood by children than a set of discrete response options presented as circles? (2) If circles are acceptable, can children use four labeled circle response options or only three? (3) Are graduated-size circles preferred to same-size circles?

Study 2 Methods

Twenty items, representing five of the domains of the CHIP-AE, were chosen. Items that presented the most concern about children's ability to understand them were selected. Each item was presented twice to each child to test five response formats. Each response format was used with eight items: a blank-line visual analogue scale (VAS); a hatched-line VAS with three labeled response areas; three labeled, equal-size circle response options; four labeled, equal-size circle response options; four graduated circle response options.

As an example, children responded to the item, "In the past week, how often did you have a stomachache?" using the blank-line VAS response format by marking an "X" on the line indicating frequency by the distance the "X" was marked from the healthy child and unhealthy child illustrations. The labeled VAS line had two hatch marks on it to indicate three options that were read and pointed out to the child: "every day," "some days," or "no days." For the circle format, the child marked the labeled circle response that was most true for him or her. (See Figure 2.)

At the conclusion of the item presentation, children were asked which of two response options was easier to answer and which they liked better. The choices were VAS versus circles; three same-sized circles versus four same-sized circles; and same-sized circles versus graduated circles.

Study 2 Results

The majority (74%) of children preferred the circle response format over the VAS lines. Moreover, 68% preferred the graduated circles over the same-sized circles, and 74% preferred four over three circles. Agreement between each child's two responses to the same items showed that children's responses were consistent 80% of the time on just over half the items when the graduated circles and same-sized circles were presented. On three-fourths of the items they agreed 80% of the time when four versus three circles were presented, whereas only one in four items had more than 80% agreement in either comparison in which the VAS format was involved. With only 19 children and four items for each comparison, reliable statistical estimates are not possible.

Study 3: Questions, Methods, and Results

Study 3 Questions

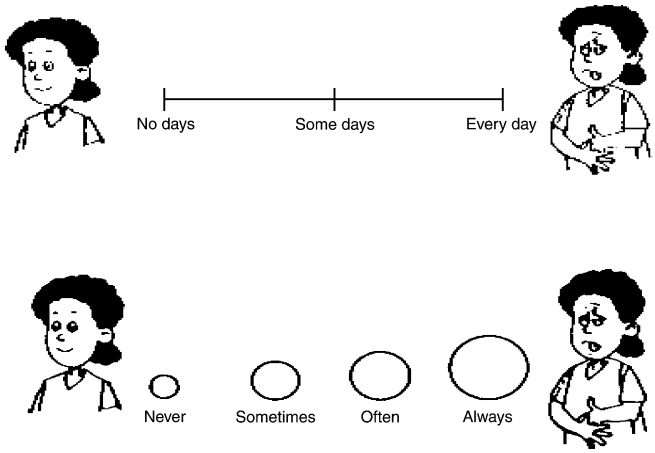
How well do children understand specific concepts of health? How well do they understand the wording of response formats? How many response options do children prefer? Is a specific recall period helpful?

Study 3 Methods

Thirty-two items were presented twice to test alternative wordings, and two risk behavior items were asked only of the 8- to 11-year olds. Four graduated circles, anchored by illustrations at each end, were used for these items. Six items were repeated at the end to test children's ability to use five instead of four response options.

Figure 1. Item: "How much do you like ice cream?" Instructions given to the child were: "If you like ice cream a lot, mark an "X" on the line near this child who is eating an ice cream cone. If you do not like ice cream at all, mark an "X" on the line near the child who is not eating ice cream. If you feel somewhere in between, mark an "X" along the middle of the line."

Figure 2. Item: "In the past week, how often did you have a stomachache?" Instructions given to the child were: "If you had a stomachache *every day*, mark an "X" on the line near this child who looks like (he/she) has a stomachache. If you had a stomachache *no days*, mark an "X" on the line near this child who looks like (he/she) does not have a stomachache. If you had a stomachache *some days*, mark an "X" somewhere in between. The closer the "X" is to the child [point to child with stomachache], the more days you had a stomachache. The closer you mark the "X" to this child [point to child without stomachache], the fewer days you had a stomachache."



The interviewer presented each item to a child. Once the children marked their responses, they were asked why they responded the way they did, and then they were asked to explain the meaning of the key term. Several examples are "healthy," "energy," "pain," "threatened," "shoplifted," and "on a dare." The interviewer presented synonyms and requested examples as needed to probe children's understanding of the items. After the six items with five responses were presented, children were asked whether they thought they responded the same both times; whether it was easier to answer with four or five circles, and why; and whether they liked four or five circles better, and why.

To examine children's understanding of the key terms for each item, a three-point coding scheme was developed where 1 = poor or no understanding of the term, 2 = some understanding, and 3 = clear understanding. Children's explanations of the key terms were coded using the interviewers' notes, referring back to the tape recordings as necessary to clarify responses. Interrater reliability in coding was 78%. Analysis of the data focused on three areas of interest:

1. level of understanding of key terms by age and for the total group
2. tendency to select the extreme responses
3. use of response formats with 4 and 5 graduated circles

Study 3 Results

Level of Understanding

Analysis of the degree of understanding of the 24 key terms presented to children of all ages showed expected age-related trends; the percentage of terms for which there was poor understanding varied inversely with age (Pearson $r = -.70$).

For the total group 17.4% of terms were poorly understood. Five-year-olds had poor understanding of 50.0% of the terms; 6-year-olds understood 25.3% of the terms poorly; 7-year-olds understood 19.0% poorly, and the older children (ages 8–11) had poor understanding of only 3.5% of the terms tested. Understanding by age was significantly different by one-way analysis of variance ($p < .001$, df = 3; $F = 27.1$). In post hoc tests, only the comparison between ages 6 and 7 was not statistically significant.

Similarly, the percentage of terms that were clearly understood increased directly with children's age ($r = .69$). For the total sample 57.9% of terms were clearly understood. Five-year-olds clearly understood 26.8% of terms; 6-year-olds, 47.2%; 7-year-olds, 55.2%; and 8- to 11-year-old children clearly understood 73.5% of the terms presented ($p < .001$, df = 3; $F = 16.2$), with no difference between ages 5 and 6, and ages 6 and 7). There were no significant gender or race differences.

Several key terms were identified as problematic for at least some of the younger children. The word "healthy" was not understood by a majority of 5-year-olds. Younger children and many older children equated "healthy" only with health behaviors, most particularly eating fruits and vegetables.

Table 2 summarizes the percentage of children at ages 5, 6, 7, and 8–11 years with poor understanding of each of the terms and the rank order of the whole group's understanding of the terms tested. The age gradient in understanding is clear, showing that almost all or all of the older children were able to understand each of the terms posed to them and that more than a third of the 5-year-olds did not understand the majority of these words or phrases. The 6- and 7-year-olds understood more than the 5-year-olds but they still had trouble with many concepts. The risk-behavior items were asked only of 8- to 11-year-olds, all of whom understood the word "weapon," but the 8-year-olds did not understand the phrase "to get high."

Tendency to Select Extreme Responses

To examine the range of response options, the percentage of extreme responses (1s and 4s) that each child gave for the 28 items answered by all ages was computed. Overall, the mean percentage of extreme responses for the sample was 63.2% (Pearson correlation with age = –.62). By age group,the mean percentage of extreme responses was 87.1%, 78.9%, and

Table 2. Study 3: Percentage of children with poor understanding of key terms for total group and by age

	Percentage with Poor Understanding					
Key Term	**Age 5 (n = 7)**	**Age 6 (n = 12)**	**Age 7 (n = 16)**	**Ages 8–11 (n = 25)**	**Total Group (n = 60)**	**Rank* in Total Group**
on a dare	85.7	72.7	43.8	8.7	40.4	1
irritable	71.4	50.0	43.8	12.0	35.0	2
for excitement	85.7	45.5	31.3	8.3	31.0	3
get away with	71.4	41.7	26.7	4.2	25.9	4
keep you from doing	71.4	41.7	18.8	8.3	25.4	5†
threatened	71.4	41.7	31.3	0	25.4	6
proud	71.4	16.7	37.5	4.0	23.3	7
temper	42.9	33.3	18.8	12.0	21.7	8
good things	57.1	25.0	31.3	0	20.0	9
energy	42.9	25.0	18.8	8.0	18.3	10†
healthy enough	42.9	41.7	18.8	0	18.3	11
comfortable	57.1	18.2	20.0	4.3	17.9	12
neighborhood	57.1	33.3	6.3	4.2	16.9	13
active games	42.9	33.3	6.3	8.0	16.7	14
real problem	57.1	0	25.0	4.3	15.8	15
are taught	33.3	45.5	6.7	0	14.5	16
nervous	42.9	16.7	12.5	0	11.7	17
other adults	42.9	8.3	12.5	0	10.3	18
numbers	14.3	8.3	21.4	0	8.8	19
itch	28.6	16.7	6.3	0	8.3	20
healthy	42.9	0	6.3	0	6.7	21†
stomachache	28.6	0	12.5	0	6.7	22
pain	14.3	8.3	6.3	0	5.1	23
worried	28.6	0	0	0	3.4	24

*Ranking of key terms by % of children with poor understanding (1 = poorly understood by highest % of children).
†Tied with next numeric rank in series.

61.4% for children aged 5, 6, and 7, respectively, and 50.4% for those ages 8 through 11. Children aged 5 and 6 gave significantly higher percentages of extreme responses than those aged 7 or ages 8–11 ($p < .001$, df = 3; $F = 14.1$). For girls, the mean percentage of extreme responses was 67.5%; for boys, 59.6%, a nonsignificant difference. Inspection of the responses to the five-point scale tested in six items showed that although the 6-year-olds are not confused by a five-point response format, they effectively convert it to a three-point format, using only the middle and both extremes.

Use of Four and Five Response Options

There was no indication that children had difficulty using five response alternatives to respond to questions. For questions that children understood well (How often do you feel really healthy? How often do you have a stomachache? How many TV shows a day do you watch?) responses were consistent between the four- and five-point administration. Fifty-six percent of children thought they answered the five-point response format the same way they answered the four-circle option. Sixty-two percent said they thought the five-circle response alternative was easier to answer than the four-circle alternative, and 67% said they liked the five alternatives better than the four-circle response alternative. Their reasons for liking the five-point response included "It gives more chance to give my answer" and "Because I get more choices."

Effects of Age and Illness

All children remained involved in the health survey task in studies 2 and 3 for at least 30 minutes, many for 45 minutes. Children 6 years and older were generally able to understand quickly what they were supposed to do and that they were to think about their own health. The 5-year-olds, on the other hand, often needed extra guidance to understand what was being asked of them. For illustrations depicting a specific representation of a more general concept (e.g., breaking a rule), young children were overly focused on the specific example provided by the illustration.

There were no statistically significant differences in understanding between the chronically ill and community samples. However, there were differences in reports of health, indicating a trend for the chronic illness sample to have lower satisfaction with health and greater discomfort (especially irritability and restricted activity) than the community children. The lack of statistical significance was related to the small sample sizes.

Discussion

Significant age-related differences in understanding the items and response formats were observed. Five- to seven-year-old children, especially 5-year-olds, had fundamental problems in understanding many basic health concepts, dramatically worse than children aged 8–11. The 5-year-olds needed much assistance with the tasks, did not understand the majority of the key terms, and tended to use only the most extreme responses, effectively describing aspects of their health as good or poor. Although the 6- and 7-year-olds also had difficulty with some terms, they understood the basic nature of the health survey and the items and responded in ways that seemed meaningful. Nonetheless, they also tended to use extreme responses. As expected, the terms that presented the most problems to the younger children were those that were most abstract, such as "healthy," "irritable," and "energy." The 8- to 11-year-old children were almost universally able to understand the tasks and the terms. They preferred the circle to the straight-line format and the graduated circles to describe the increasing frequency or intensity of their response; they were comfortable with up to 5 response options; and they explained their answers in ways that clearly showed they understood. The addition of a specific 4-week recall period to items regarding the experience of symptoms and behaviors virtually eliminated responses that referred to distant experiences. No pattern of gender or race differences in understanding or in use of response options was found. Children were positively engaged by and identified with the illustrated child character and validated its gender, age, and race neutrality.

These results indicate that school-age children can report their health when asked in a format that they find acceptable and understandable. Children as young as age 6 were able to report on virtually all aspects of their health. Cognitive limitations were likely responsible for the lack of comprehension of the task demands among the 5-year-olds (Piaget, 1952; Rebok, 1987). The primary limitation of these cognitive tests is the small numbers of children tested in each study, and further work in this area is heartily encouraged. These results have guided the development and testing of the Child Health and Illness Profile—Child Edition (CHIP-CE).

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232.

Breton, J. J., Bergeron, L., Valla, J. P., Berthiaume, C., Gaudet, N., Lambert, J., St-Georges, M., Houde, L., & Lepine, S. (1999). Quebec child mental health survey: Prevalence of DSM-III-R mental health disorders. *Journal of Child Psychology and Psychiatry, 40,* 375–384.

Burbach, D. J., & Peterson, L. (1986). Children's concepts of physical illness: A review and critique of the cognitive-developmental literature. *Health Psychology, 5,* 307–325.

Byrne, B. M. (1996). *Measuring self-concept across the life span* (pp. 52–58). Washington, DC: American Psychological Association.

Donaldson, M., & Wales, R. J. (1968). On the acquisition of some relational terms. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 235–268). New York: Wiley.

Ericsson, K. A., & Simon, H. A. (1993). *Verbal reports as data: Protocol analyses* (rev. ed.). Cambridge, MA: MIT Press.

Fox, N. A., & Leavitt, L. A. (1995). VEX-R: Violence Exposure Scale for Children—Revised. Unpublished. Institute for Child Study, University of Maryland, College Park.

Friedman, W. J. (1991). The development of children's memory for the time of past events. *Child Development, 62,* 139–155.

Gale, A., & Lynn, R. (1972). A developmental study of attention. *British Journal of Educational Psychology, 42,* 260–266.

Gathercole, S. E. (1998). The development of memory. *Journal of Child Psychology and Psychiatry, 39,* 3–27.

Hagen, J. W., & Hale, G. H. (1973). The development of attention in children. In A. D. Pick (Ed.), *Minnesota Symposia on Child Psychology* (Vol. 7, pp. 117–140). Minneapolis: University of Minnesota Press.

Harbeck, C., & Peterson, L. (1992). Elephants dancing in my head: A developmental approach to children's concepts of specific pains. *Child Development, 63,* 138–149.

Hart, E. L., Lahey, B. B., Loeber, R., & Hanson, K. S. (1994). Criterion validity of informants in the diagnosis of disruptive behavior disorders in children: A preliminary study. *Journal of Consulting and Clinical Psychology, 62,* 410–414.

Harter, S., & Pike, R. (1984). The pictorial scale of perceived competence and acceptance for young children. *Child Development, 55,* 1969–1982.

Hergenrather, J. R., & Rabinowitz, M. (1991). Age-related differences in the organization of children's knowledge of illness. *Developmental Psychology, 227,* 952–959.

La Greca, A. M. (1990). Issues and perspectives on the child assessment process. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 3–17). Boston: Allyn and Bacon.

Landgraf, J. M., & Abetz, L. N. (1996). Measuring health outcomes in pediatric populations: Issues in psychometrics and application. In Spiker, B. (Ed.) *Quality of life and pharmacoeconomics in clinical trials* (2nd ed., pp. 793–802). Lippincott-Raven Publishers: Philadelphia.

McGrath, P. A. (1991). Versatile pain measures for children. *Journal of Pediatric Somatic Medicine, 6,* 175.

McGrath, P. A., Seifert, C. E., Speechley, K. N., Booth, J. C., Stitt, L., & Gibson, M. C. (1996). A new analogue scale for assessing children's pain: An initial validation study. *Pain, 64,* 435–443.

McKay, K. E., Halperin, J. M., Schwartz, S. T., & Sharma, V. (1994). Developmental analysis of three aspects of information processing: Sustained attention, selective attention, and response organization. *Developmental Neuropsychiatry, 10,* 121–132.

Natapoff, J. (1978). Children's views of health: A developmental study. *American Journal of Public Health, 68,* 995–1000.

Natapoff, J. (1982). A developmental analysis of children's ideas of health. *Health Education Quarterly, 9,* 131–141.

Nelson, N. (1976). Comprehension of spoken language by normal children as a function of speaking rate, sentence difficulty, and listener age and sex. *Child Development, 47,* 299–303.

Neuhauser, C., Amsterdam, B., Hines, P., & Steward, M. (1978). Children's concepts of healing: Cognitive development and locus of control factors. *American Journal of Orthopsychiatry, 48,* 335–341.

Ornstein, P. A. (1995). Children's long-term retention of salient personal experiences. *Journal of Traumatic Stress, 8,* 581–605.

Perrin, E. C., & Gerrity, P. S. (1981). There's a demon in your belly: Children's understanding of illness. *Pediatrics, 67,* 841–849.

Piaget, J. (1952). *The origins of intelligence in children.* New York: International Universities Press.

Raviv, A., Raviv, A., Shimoni, H., Fox, N., & Leavitt, L. (in press). Children's self report of exposure to violence and its relation to emotional distress. *Journal of Applied Developmental Psychology.*

Rebok, G. W. (1987). *Life-span cognitive development*. New York: Holt, Rinehart & Winston.

Rebok, G. W., Smith, C. B., Pascualvaca, D. M., Mirsky, A. F., Anthony, B. J., & Kellam, S. G. (1997). Developmental changes in attentional performance in urban children from eight to thirteen years. *Child Neuropsychiatry,* 3, 28–46.

Ross, D. J., & Ross, S.A. (1984). Childhood pain: The school-aged child's viewpoint. *Pain, 20,* 179–191.

Schwab-Stone, M., Fallon, T., Briggs, M., & Crowther, B. (1994). Reliability of diagnostic reporting for children aged 6–11 years: A test-retest study of the Diagnostic Interview Schedule for Children—Revised. *American Journal of Psychiatry, 151,* 1048–1054.

Starfield, B., Bergner, M., Ensminger, M., Riley, A., Ryan, S., Green, B., McGauhey, P., Skinner, A., & Kim, S. (1993). Adolescent health status measurement: Development of the Child Health and Illness Profile. *Pediatrics, 91,* 430–435.

Starfield, B., Riley, A. W., Green, B. F., Ensminger, M. E., Ryan, S. A., Kelleher, K., Kim-Harris, S., Johnston, D., & Vogel, K. (1995). The Adolescent Child Health and Illness Profile: A population-based measure of health. *Medical Care, 33,* 553–566.

Stone, W. L., Lemanek, K. L. (1990). Developmental issues in children's self-reports. In La Greca, A.M. (Ed.) *Through the eyes of the child: Observing self-reports from children and adolescents* (pp. 18–55). Boston: Allyn and Bacon.

Valla, J. P., Bergeron, L., Bérubé, H., Gaudet, N., & St-Georges, M. (1994). A structured pictorial questionnaire to assess DSM-III-R-based diagnoses in children (6–11 years): Development, validity, and reliability. *Journal of Abnormal Child Psychology, 22,* 403–423.

Valla, J. P., Bergeron, L., Bidaut-Russell, M., St-Georges, M., & Gaudet, N. (1997). Reliability of the Dominic-R: A young child

mental health questionnaire combining visual and auditory stimuli. *Journal of Child Psychology and Psychiatry, 38,* 717–724.

Wechsler, D. (1974) *Wechsler Intelligence Scale for Children—Revised*. New York: Psychological Corporation.

Woodcock, R. W., & Mather, N. (1990). W.J-R Tests of Cognitive Ability—Standard and Supplemental Batteries: Examiner's Manual. In R. W. Woodcock & M. B. Johnson, *Woodcock-Johnson Psycho-Educational Battery—Revised,* Allen, TX: DLM Teaching Resources.

Zeltzer, L. K., LeBaron, S., Richie, D. M., Reed, D., Schoolfield, J., & Prihoda, T. J. (1988). Can children understand and use a rating scale to quantify somatic symptoms? Assessment of nausea and vomiting as a model. *Journal of Consulting and Clinical Psychiatry, 56,* 567–572.

Innovative Strategies for Increasing Active Parental Consent in School-Based Drug Education Research

Jennifer Hawes-Dawson, Gail Zellman, Sarah Cotton, and Marvin B. Eisen

Introduction

Considerable research literature documents the negative impact of active parental consent on participation rates and sample bias in school-based studies of adolescents, especially studies on sensitive topics such as drug use or sexual behavior. Yet little has been written about promising strategies for increasing parental response when active consent is either required by federal or state law or by local Institutional Review Boards (IRBs). Consequently, researchers who are required to implement active parental consent methods lack solid guidance about what *does* and *does not* work to enhance parental return of signed permission slips for their children.

This paper describes how we achieved an overall parental response rate of 77% under active (written) consent requirements in 12 large, inner-city schools with substantial minority populations. While the 77% form return rate is higher than that typically reported in the active consent literature, achieving it required an intensive and costly campaign that required a high level of support from school principals, coordinators, and teachers. We describe our consent plan and incentives and discuss their costs and effectiveness in implementing in-school surveys with 6,300 sixth-grade participants in testing a school-based drug prevention program called Lions-Quest Skills for Adolescence. We also examine the repercussions of active versus implicit (passive) parental consent procedures on study outcomes (response rates, data quality, fieldwork procedures, schedules, and costs) and compare our results with similar RAND research and other published studies.

Literature Review

Two methods are commonly used for obtaining parental approval to conduct research with minors: active (written) versus implicit (passive) consent. The first method, active consent, requires that all parents return a signed permission slip to indicate whether they *do* or *do not* want their child to participate in the research. Under active consent, parents who fail to return a consent form as well as those individuals who indicate on the form that they do not want their child to participate in the research are treated as "parental refusals." The second procedure, implicit or passive consent, asks parents to return a form *only if* they do not want their child to participate. Under implicit parental consent, we assume that parents have consented to the research unless they refuse by mail, by phone, or in person. Thus, under implicit parental consent, parents who want their child to participate in the research do not have to take any action—we assume that nonresponse reflects a conscious parental decision to allow their child to participate in the research.

Research on the costs and effectiveness of these two parental consent methods has generally found that active parental consent yields low response rates and sample bias, is expensive and time consuming to implement, and may not be feasible for large-scale studies (Ellickson & Hawes, 1989; Esbensen, et al., 1996; Esbensen, Miller, Taylor, He, & Freng, 1999).[1] The requirement for active parental consent consistently has resulted in samples limited to half the size that would ordinarily be available if passive consent methods were used (Thompson, 1984). Because many parents fail to return a signed consent form, more parents are counted as refusing participation under active consent, which typically yields overall response rates that cluster between 30% and

Jennifer Hawes-Dawson, Gail Zellman, and Sarah Cotton are at RAND, Santa Monica, California.

Marvin B. Eisen is at the Urban Institute.

[1]To date there have been very few published studies that have disputed these general findings about the adverse affects of active consent requirements on response rates and sample bias. Two published papers, one by Ellickson and Hawes (1989) and one by Mobers and Piper (1990), demonstrated that it is feasible—when the sample size is manageable and clustered in sites near the researchers' home offices—to obtain high response rates with aggressive mail and phone follow-ups. Both studies reported response rates in the range of 85–88%; however, both studies employed innovative and costly strategies to boost response rates, including a mix of mail and phone follow-ups, as well as school channels to reach parents. Ellickson and Hawes also aggressively used teachers to reach nonresponding parents, while Mobers and Piper relied heavily on phone call reminders to accept verbal consent from parents who did not respond to mail requests. It is interesting to note that the IRB for the Mobers and Piper study approved an innovative strategy of accepting parents' verbal consent provided that the researchers maintained copies of phone company charge records to provide independent evidence of the phone contact with parents. For both studies, a minimum 30-day process was needed to ensure adequate time for parental response. Informal discussions with survey colleagues at other survey research organizations also indicate that there have been other isolated cases where high response rates have been obtained under active consent requirements. However those studies have several critical elements in common: relatively small sample sizes; highly cooperative districts and schools; good parent contact information, including addresses and phone numbers; and experienced survey staff who are skilled at working in school environments.

60%. Moreover, the children of those parents who do consent are rarely representative of the population being studied. Past studies have shown that active consent also yields study populations that significantly underrepresented important groups—African-Americans, Asian-Americans, Latinos, low achievers, children with less well-educated parents, those at risk for engaging in problem behaviors, socially rejected and neglected children, socially withdrawn children, and those who are likely to refuse to answer sensitive questions (Beck, 1984; Ellickson & Hawes, 1989; Frame, 1987; Josephson & Rosen, 1978; Kearney, 1983; Leuptow, Mueller, Hammes, & Lawrence, 1977; Severson & Ary, 1983; Singer, 1978; Thompson, 1984). These effects are of particular concern for research on preventing adolescent drug use and other risk-taking behaviors. Besides producing smaller, less representative samples, active consent also tends to produce fewer "at-risk" students for whom school-based prevention effects are more desirable.

Few studies have identified promising strategies that are feasible with large, multisite, and demographically diverse school samples. One notable exception is a paper published by Thompson (1984) reporting the results of four relatively inexpensive techniques used to supplement an initial mailing to parents of 500 elementary school students: (1) incentive for children (free photograph of the child); (2) incentives for parents (copy of research results and related articles); (3) communication (outreach) to the children (short fun exercise to expose children to the project goals and solicit their help in getting parents to sign consent forms); and (4) communication (outreach) to the parents (phone calls to parents to explain study background and goals). Mail methods alone yielded an overall return rate of 40%, but the response was substantially lower among minority parents (African-American parents were two times as likely not to respond by mail). Of the four additional methods tested, calling parents was the most effective procedure for both white and African-American children; however, it was also the most time consuming and costly procedure. The child incentives had the second most positive impact on response rates for both white and African-American children—both groups responded quite similarly. Promising to give parents the published results improved parental response, but more so for whites than for minority parents. Communicating with the child was less effective than the other three strategies but did help boost the overall return rate. These results suggest that the most promising strategies for contacting difficult-to-reach minority parents are child incentives coupled with follow-up phone calls to parents.

Because of active consent's potential for severely reducing sample size and increasing sample bias, it is not surprising that implicit (passive) parental consent has been the dominant method in most school-based research for decades. Most of the large, federally funded school-based studies, such as Monitoring the Future, the National Education Longitudinal Study (NELS:88), High School and Beyond, and the National Longitudinal Study of High School Class of 1972 (NLS-72), as well as numerous, smaller cross-sectional surveys sponsored by state and local governments and private foundations, have relied primarily on implicit parental consent methods.[2]

While it is true that in the past, few school districts or schools have required active parental consent, this trend has been changing in recent years, due to efforts to enact state and federal laws that require active parental consent for research on sensitive topics (drugs, sex, etc.) (Esbensen et al., 1996). In 1994, there was an aggressive campaign to enact the Grassley Amendment to the Goals 2000 Act that would have required active parental consent for sensitive research funded by the Department of Education. While this federal law has not been enacted yet, the debate over this initiative has had a major influence on how IRBs have responded to research requests for implicit consent procedures. It has also led many IRBs to require that researchers fully investigate whether there are any state laws in existence that might require active consent.[3] Informal discussions with our colleagues at other survey institutions throughout the country indicate that the recent movement to pass federal and state consent laws is one of the major factors driving more and more IRBs to push for active consent—despite the known adverse effects on response rates, sample bias, fieldwork procedures, schedules, and costs.[4]

Legitimate questions have been raised about whether implicit parental consent meets the ethical and legal standards for obtaining informed consent from parents. Some observers argue that implicit consent procedures do not fully inform parents about the research or give them adequate opportunity to refuse participation, particularly when there is a one-time distribution of consent materials just prior to the survey

[2]In an unpublished paper presented at the 1992 National Field Directors' conference, Abraham (1992) presented response rate results from four well-known federally funded school studies (High School and Beyond, Monitoring the Future, National Educational Longitudinal Studies, and NLS) that showed that under implicit consent procedures, they achieved response rates from 82% to 93%. The parental refusal rate for these studies was extremely low. Most of the nonresponse was due to factors other than parental refusals (e.g., absenteeism, tardiness). We observed a similar pattern at RAND when we used implicit parental consent for Project ALERT and Project ALERT Plus, two longitudinal school-based drug prevention research studies, one involving a panel of over 6,000 adolescents in 30 schools in California and Oregon and a recent panel of 6,000 adolescents in 60 schools in South Dakota. Both of these studies achieved nearly identical results under implicit consent: Only 8.5% to 10% of the parents refused to allow their child to participate. Both RAND studies achieved an overall baseline survey completion rate of 85%. The sample loss due to parental refusals under implicit consent procedures has been consistently low on every RAND project that has employed a three-stage parent notification procedure.

[3]To date we have identified only one such state law. A little-known state law has been in existence in California (California Education Code 60.650) since 1977; it requires active parental consent if any "test, questionnaire, or examination contains any questions about the pupil's personal beliefs or practices in sex, family life, morality, and religion, or any questions about his parents' or guardians' beliefs or practices in sex, family life, morality, and religion." South Carolina had a law similar to the California consent ruling pending in the legislature several years ago.

[4]See Esbensen et al. (1996) and Ellickson and Hawes (1986) for an excellent summary of the legal, ethical, and methodological issues raised by active versus implicit consent and results from their own research that provide new insights about how these two consent methods work in practice.

administration. Others question the underlying assumption that nonresponse under implicit consent means that the parent has granted consent. They worry that typically implicit consent methods do not give parents sufficient time to refuse; that there is no written proof of parental consent; that such procedures may not be appropriate for surveys on sensitive topics; and that some state or federal laws may restrict the use of these methods for certain types of studies.

Study Challenges

Because active parental consent procedures were mandated by California state law (California Education Code 60.650) as well as the Federal Office of Protection from Research Risks (OPRR), we faced the challenge of how best to design a cost-effective and timely plan for maximizing parental response rates from a large, multiethnic, urban sample in Los Angeles designed to evaluate a school-based prevention program, Lions-Quest Skills for Adolescence. Determining how best to use limited project resources to minimize parental nonresponse rates involved difficult cost and time tradeoffs. We needed to address several potential challenges: (1) the sheer size and racial-ethnic diversity of the student population; (2) language barriers; (3) parental indifference and lack of involvement in school activities; (4) problems gaining quick access to accurate, computerized parent name, address, and phone lists, as well as school-level and class-level student lists; (5) varying levels of school support or "buy-in" from participating principals, program coordinators, and teachers; (6) large number of teachers (n = 130+) and individual classes (n = 250+) involved in the research activities; (7) schedule constraints due to the school calendar, which imposed several limitations on the time available to obtain parental consent and complete the baseline survey before the 1997–98 school year ended;[5] and (8) cost constraints for mounting an extensive active parental consent campaign. For all these reasons, we had to devise an innovative and multifaceted consent plan and set of incentives directed at schools, coordinators, teachers, classes, and individual students to ensure the success of the consent procedures.

Methodology

Subjects

Research subjects included 6,300 sixth-grade children in 12 schools in the Los Angeles Unified School District (LAUSD).[6] The sample was composed of 52% Hispanic, 12% African-American, 10% white, 8% Asian-American, and 16% Other and combinations. In the spring of the 1997–98 school year, the RAND Survey Research Group collected data from the 6,300 sixth-graders during their regular science classes.

Procedures

Over a 2-month period, from March to April 1998, we implemented a multifaceted active parental consent plan designed to maximize response rates. First, we established a partnership with school principals and teachers to devise a customized consent plan for each school that could be implemented within a 4–6-week period. We solicited input from principals and teachers *before* the consent plan was finalized, to ensure their "buy-in" and to get their advice and recommendations. We started this process by conducting a "kick-off" planning meeting with our school district contact and our principals to present them with the survey challenges and goals and to get their ideas. Next, we visited each school to conduct "brainstorming" meetings with individual science teachers, because past educational research shows that teachers are a critical link to children and parents.

Based on input from principals and teachers and drawing upon past RAND research experience and published studies, we designed and implemented a consent plan that included the following 12 critical components:

1. Each school principal designated a school coordinator to serve as a primary liaison with the research team.
2. We established a personal relationship and partnership with principals, coordinators, and teachers via frequent phone, mail, and in-person contacts.

[5]The timeline for the consent and baseline survey activities was constrained by five major factors: (1) the long lead times to recruit the district and participating schools—all schools were not "on board" with a designated school coordinator until February 1998, which left us only 3 months to complete the parent consent and the baseline survey for over 6,000 students; (2) at least half of the 3-month window of opportunity was needed for the consent process alone, in order to maximize response rates; (3) due to previously scheduled school events (standardized test periods, special school events and programs, other school activities, etc.) that occurred at the end of the school year, we had a very limited number of days that were available to the research team for scheduling consent and/or survey work; (4) we had a maximum of 4–6 weeks to complete baseline surveys and makeup sessions with 6,000+ students in 12 schools—this generated a need for a large, experienced data collection team (16 data collectors, plus two field managers to administer 250 survey sessions in English and Spanish over a 4- to 6-week period); and, finally, (5) we could not delay the parent consent and/or survey activities to the next school year because the Lions-Quest curriculum (over 40 core lessons) was scheduled to start as soon as the 1998–99 school year started—to delay the curriculum implementation would have made it impossible for teachers to teach all of the designed SFA lessons before the school year ended. Thus, the feasibility of conducting this study was completely dependent on our ability to obtain parental consent to complete the baseline surveys with 6,000+ sixth-graders roughly 4–6 weeks before the 1997–98 school year ended.

[6]In this paper we present results from the Los Angeles site only. The national study is being conducted in three metropolitan areas (Los Angeles, Detroit, and Montgomery County, Maryland) with a total of 34 middle schools and 7,400 enrolled students who have parental consent to participate in annual surveys from grades 6–8 (and possibly beyond). Because of limited project funds, there was variation in the level of resources that were invested in each site. Ultimately, for pragmatic reasons, the bulk of the resources were placed in the Los Angeles sites and the Maryland site, where our chances for success were the greatest.

3. We provided incentives for school coordinators, teachers, classes, and individual students. Teachers and coordinators were given two free movie passes (equivalent to about $5.50 per ticket) prior to the start of the consent process. In most schools, students also received free stickers after they returned signed consent forms, whether they were marked YES or NO (equivalent to about $0.50 to $0.75 per sticker). Some students also received homework credit from their teachers if they returned a signed consent form (marked YES or NO). Classes that achieved a 100% participation rate also received a free pizza party after the survey was completed (equivalent to about $100 per class for a typical class of 25 students). This provides about two pizza slices and one soda per child, plus the same for the classroom teacher.

4. We used school channels for parent notification, in lieu of direct mailings and phone calls to parents, because of time constraints and teacher recommendations. We sent each of the 130+ classroom teachers a set of preaddressed consent packets to give to students to take home to their parents. Teachers played a critical role in distributing, collecting, and accounting for all consent forms. They also gave reminder notices to children to take home to nonresponding parents. Some teachers also tried to maximize parental response by including the parent packets with other school information that is routinely given to children to take home to their parents on designated days each week.

5. We also implemented additional methods to direct parents' attention to the consent form, including putting the parent letter on school stationery, signed by individual school principals; translating the materials into Spanish; and giving students a replacement consent packet to take home to parents who did not respond to the initial request within two weeks.[7]

6. We also devised streamlined recordkeeping procedures to minimize burden on schools and teachers. The RAND survey team handled all of the logistical support, including printing, assembling, distributing, and collecting consent materials. We devised simple, user-friendly checklists for teachers' use in keeping track of the returned consent forms. Efforts to minimize burden on teachers were essential to promote their cooperation and support in distributing and collecting parent consent forms.

7. We gave regular feedback to school principals, coordinators, and teachers via faxes and Federal Express letters regarding the parent response rates and solicited their support to boost returns where needed.

8. We involved the principal investigator in troubleshooting as needed, including making unannounced visits as needed to schools to resolve problems.

9. We devised a "work activity" for students who did not have parental consent to take part in the drug survey. For the baseline survey sixth-graders, we used an alternative, anonymous nonsensitive survey, which also gave us some basic demographic and non-drug-use-related data for the students for whom active parental consent could not be obtained. These data will be used to explore some of the basic differences between students participating and not participating in the evaluation component in the 12 study schools. For the one-year follow-up survey, we used educational puzzles in lieu of an anonymous survey for the nonparticipants. These alternative work activities were needed to address practical school concerns about what to do with large numbers of students who might not have parental permission to complete the drug prevalence survey.

10. We set aside a minimum 4- to 6-week period for the distribution, collection, and tracking of parent consent forms to ensure that parents had sufficient time to receive and respond to the consent request.

11. We scheduled makeup sessions for consented students who were absent on the day of the baseline survey or whose parent consent form was received late.

12. We increased the survey budget for the Los Angeles schools relative to Detroit and Maryland to ensure that we could implement a rigorous follow-up campaign to maximize parental response rates. Because 60% of the study's sample was clustered in the L.A. sites and a substantial portion of the minority population was found in L.A., the principal investigator made a conscious decision to redirect a larger proportion of the budget's funding into this challenging school setting.

Results

Overall Response Rates

Extensive follow-up efforts via school channels raised the overall consent rate well beyond that typically reported in the active consent literature. Overall, 77% of the 6,300 parents contacted returned a permission form. This included 66% who consented, 11% who refused, and 23% who did <u>not</u> return a consent form after repeated follow-up.

Baseline Survey and Sample Completion Rates

We successfully completed baseline surveys with 95% of those children who had parental consent to take part in the Lions-Quest surveys. This yielded a final baseline sample completion rate of 63%. Three major factors account for the 37% nonresponse: (1) parent refusal (11%); (2) parents who did not return a consent form (23%); and (3) student absen-

[7]Previous research conducted by Mobers and Piper (1990) shows that parental response rates are higher when school stationery was used instead of university letterhead. On RAND studies, we routinely use school stationery for all parent consent letters to maximize the probability that parents will read and pay attention to mail requests.

teeism and refusal, combined (3%). Thus, most of the nonresponse was associated with parents who did *not* return a consent form and who, therefore, had to be treated as "refusals."

Can we assume that nonresponse means that these parents did *not* want their child to take part in the research? Or should we assume that nonresponse *means parental apathy or lack of motivation?* If the research conducted by Ellickson and Hawes (1989) is the rule rather than the exception, we can assume that parental apathy and lack of motivation to sign and return the consent form without considerable prompting is the *more* likely explanation. With more time and follow-up, these authors would argue, most of the nonrespondents in the current study would probably ultimately approve their child's inclusion in the research. However, the study by Esbensen et al. (1996) produced mixed results, with one site's data supportive of the apathy hypothesis and another site's data suggestive of true parental opposition. Given the budget and time constraints for the current study, we were not able to investigate the true reason behind the 23% parental nonresponse rate.

Impact of Active Consent Requirement on Sample Bias

We collected basic demographic and non-drug-use-related data via an anonymous baseline survey for almost 1,600 students for whom active parental consent could not be obtained so that we could explore some of the basic differences between consenters and nonconsenters. Even though we have not yet completed our nonresponse analyses, our preliminary results suggest that there are significant demographic and personal differences between consenters and nonconsenters. The former include differences by gender, race, single- versus two-parent households, those children who are living with one or more birth parent(s) versus those living with guardians, and children's educational aspirations. The latter include differences in conformity, boredom, and goal-setting. The underrepresented groups in our final sample include males, African-American and American Indian students, children living in one-parent families, children who live with guardians such as grandparents in lieu of their own father or mother, and children who do not plan to attend college. At the personal level, those who reported that it was important to go along with friends, those who had more difficulty keeping busy, and those who had trouble setting goals were less likely to return the parent consent form.

School-Level Differences

Schools that were less committed to the study produced lower consent rates than schools with strong institutional support. As shown in Table 1, parent consent rates clustered into two groups: eight schools (schools A–H) achieved a 72–80% consent rate, while four schools (schools I–L) achieved a 51–62% parent consent rate. Three of the low-responding schools also experienced start-up problems, which delayed the parent notification and reduced the time allowed for contacting parents from four to two weeks. This schedule slippage certainly prevented us from carrying out more aggressive parental follow-up before the school year ended.

Table 1. Positive parental consent status of study participants by demographics and school

	Parental Consent = Yes*	
Variable	***n***	**% ****
Gender		
Female	2267	(73)
Male	2143	(67)
Missing	20	(—)
Race/ethnicity		
Asian-American	373	(77)
American Indian	58	(54)
African-American	445	(59)
Hispanic-American	2377	(73)
White	448	(72)
Combination (of above)	303	(67)
Other	349	(65)
Missing	79	(—)
School (in descending order based on % consenting)		
School A	600	(80)
School B	437	(79)
School C	456	(78)
School D	375	(78)
School E	419	(77)
School F	459	(74)
School G	307	(73)
School H	203	(72)
School I	288	(62)
School J	304	(58)
School K	345	(56)
School L	221	(51)
Missing	16	—

*This represents the number and percentage of parents who gave permission for their child to take part in the research.
**Total of percentages may not equal 100 due to rounding.

It is also interesting to note that there was a correlation between the parent consent rate patterns at baseline and the one-year student attrition rates. The eight schools with high parental consent rates experienced a much lower one-year student attrition rate—their student attrition rate between grades 6 and 7 was 17%, compared with a 21% attrition rate for the four schools with low parental consent rates. These results suggest that the overall school climate in schools that experience high student turnover can have an adverse effect on parental consent rates.

Teacher Support

Teachers were the key to maximizing parental response rates for the LAUSD schools. The success we experienced was due in large part to the fact that we were able to solicit

and sustain a high level of support from most of the 130+ classroom teachers who assisted us in the distribution, collection, and tracking of consent forms. We had no viable alternatives to teachers, since it was clear at the onset of the project that we did not have the time and materials to mount an aggressive mail and phone campaign to reach parents. We did not have the lead time to implement the traditional three-step parent contact method (initial mailing, postcard reminder, replacement mailing) that has been effective in past studies. Furthermore, we were not able to get quick access to computerized parent contact information soon enough to make a mail out/mail back strategy a feasible contact procedure. We also found that LAUSD schools were reluctant to release parents' phone numbers without getting official school board approval; thus, we had to abandon the possibility of contacting parents by phone. For these reasons, we had no choice but to rely heavily on teachers to reach parents.

All of the up-front measures that we took to get teachers' support and "buy-in" proved to be well worth the investment. Teachers provided many creative ideas during our brainstorming sessions about how to improve parental responses. They introduced the idea of low-cost incentives for children directed at individuals (e.g., stickers or homework credit) and classrooms (pizza party).[8] We also found that some schools had established classroom procedures for notifying parents, such as setting aside a particular day for all parent notices, and many teachers used those standard dissemination channels for the parent consent activities. We also observed that there was considerable variation in school policies and practices with respect to usual parent notification procedures. We gave each school the freedom to implement creative parent notification strategies, based on their judgments about what would or would not work in their particular school.[9]

We also found that it was important to get to know the teachers personally and whenever possible to send communications directly to them via personalized (rather than generic "Dear Teacher") letters and faxes rather than to rely exclusively on third parties (such as principals or school coordinators) to keep them informed about critical activities. Something as simple as getting a teacher name list, so that we could send them personalized letters, was an important feature of our plan to put a "face" on the project to try to maximize teacher cooperation. Also, we found that Federal Express overnight packages directed at teachers were one of the most effective ways of reaching busy teachers—far better than faxes and phone contacts.

By partnering with teachers, we were able to improve response rates considerably. The teacher cooperation rate was quite similar across schools. We observed very little within-school variation in terms of response rate patterns at the classroom teacher level. Surprisingly, we did not encounter any teachers who refused to cooperate with the consent process or whose response rate patterns differed in any noticeable way (e.g., higher refusal or nonresponse patterns) from their colleagues.

Effectiveness of Incentives

Incentives directed at school coordinators, teachers, classes, and individual students proved to be quite effective in maximizing parental responses. The personal gift of two free movie tickets to coordinators and teachers was very well received, especially by the teachers. We believe that this low-cost incentive bought us a lot of goodwill among the teachers. We received lots of unsolicited positive feedback from teachers about how much they appreciated the personal recognition. It has also set a precedent, at least in our L.A. sites; we have continued to provide low-cost personalized teacher incentives, such as movie tickets, certificates to bookstores, coffee mugs, or appreciation lunches, during each survey wave to show our appreciation for teachers' continued support. We found that personalized teacher incentives were far more appealing that providing traditional gifts of office and school supplies. Teachers appreciated having a personal gift for themselves—even if it was relatively inexpensive.

Students also responded favorably to the individual and classroom-level incentives that we offered. The stickers were offered in 10 of the 12 schools; teachers reported that this incentive was extremely popular with students. The stickers were also relatively inexpensive, since we were able to get an educational discount for schools—we paid about 50 to 75 cents per sticker for a popular assortment of stickers that would appeal to sixth-grade boys and girls. The free pizza party for classes that achieved a 100% parental response was also very appealing to students and teachers. Nine of the 12 schools had one or more classes that were eligible to receive a pizza party. A little over a third of all teachers had one or more classes that were eligible to receive the free pizza party. In the end, 25% of the classes (68 of 262 classes included in the survey) achieved a 100% parental response and received a free pizza party. Each pizza party cost us roughly $100 per class of 25 students. This translates into a per-child incentive payment of about $4 for the eligible children. Altogether, we estimated that about 1,900 of our panel of sixth-graders received free pizza parties. The logistics of organizing these pizza parties turned out to be more difficult than we had envisioned, but we now have a much better sense of the dos and don'ts of how to set up similar arrangements with vendors in the future.

Summary and Discussion

These findings suggest that acceptable parent consent rates can be achieved using active parental consent procedures, but the time and expense involved are high. Some of the most promising strategies for boosting parental response rates

[8]Esbensen et al. (1996) also offered pizza parties to classes attaining a 100% return rate.

[9]Esbensen et al. (1996) also found that cooperative schools came up with creative school-initiated incentives for children who return consent forms. This included "go early" to lunch passes, double recess passes, extra credit, or candy. They combined the school-initiative incentives with other researcher-initiated incentives like special pencils and pizza parties for classes that returned all consent forms.

(extensive mail and phone follow-up; site visits;[10] incentives for coordinators, teachers, students, and/or parents; and developing partnerships with teachers, etc.) may be impractical or prohibitively expensive for large national studies.

The results of this study support previous findings that multiple strategies, including incentives directed at school coordinators, teachers, classes, students, and parents, are needed to enhance parental response rates and to minimize sample bias. There is no single "magic bullet" to guarantee high parental response under active parental consent methods, since methods to improve parental returns can vary in their effectiveness by school, region, and student characteristics (e.g., whites versus minority students). This means that a "cookie-cutter" approach to consent methods is not desirable—creative, customized approaches are needed. The focus should be on understanding the local conditions that may operate in a particular district or school to produce low response rates and bias so that a customized consent plan can be developed to achieve the response goals.

This study also clearly demonstrated that teachers are a critical link between the researchers, parents, and students. Teachers serve as institutional gatekeepers in many ways, and they are often crucial in determining the extent to which a survey will be capable of realizing its response rate goals at both the institutional and respondent level. It is difficult to imagine that our efforts to enhance parental responses to this current study would have been as effective without strong, local support from teachers. Our results suggest that we need to reorient, or at least broaden, the way we think about securing parental cooperation so that we take into account the institutional and local environmental factors that may operate to suppress response rates and produce sample bias in a particular district or school.

Finally, this study suggests that future research is needed to examine other low-cost methods of increasing response rates among underrepresented groups, especially African-American students. Further research should also assess which methods work best with children at different grade levels (elementary versus middle school versus high school) and in different school settings (urban versus suburban versus rural).

Traditionally, schools have been an efficient and cost-effective setting in which to conduct school-based drug education research. However in recent years, school-based studies have become increasingly difficult and costly to implement when active parental consent is required, either by federal or state law or because of Human Subjects Protection Committee (HSPC) concerns. The ethical and legal standards for obtaining informed consent from parents are being passionately debated among the HSPCs throughout the country. Moreover, there are currently several major factors driving more HSPCs to require active parental consent—even though the research evidence clearly suggests that this particular consent method can have serious adverse effects on response rates, data quality, fieldwork procedures, schedules, and costs.

The high cost of active consent and the potential sample loss and sample bias inherent in the process are harsh realities that cannot be ignored. For the current study, we estimate that our per-case cost was $20 per eligible student and $30 per child who completed the baseline survey. This per-case cost includes staff time to design and implement the consent plan and the nonlabor costs associated with the parental notification process, such as incentives, materials, postage and shipping, and travel. It does not include the cost of the actual data collection.

In the end, despite the time and expense that were invested in the active consent process, we still successfully surveyed only 63% of the target population at baseline. We also have preliminary evidence showing demographic and other differences between consenters and nonconsenters. When the 37% baseline nonresponse rate is coupled with the expected 10% to 15% yearly attrition rate that we are likely to experience in future survey years as students move, transfer, or drop out, we have to face a sobering reality about how much the sample is likely to shrink over time. This study provides compelling evidence that future research debates about the ethical and legal standards for obtaining informed parent consent need to consider the costs and research implications of achieving potentially biased samples as they carefully weigh the pros and cons of active versus implicit parental consent. This is a debate that is likely to be passionately discussed for years to come.

[10]On RAND's Project ALERT Plus Study in South Dakota, based on principal recommendations, we made home visits to most of our American Indian parents, and the response was overwhelmingly positive. Our parent refusal rate among the American Indian parents was virtually nonexistent (less than 2% refused). Other researchers have also found that when schedules and budgets permit, home visits to parents can be extremely effective ways of soliciting higher parental response rates.

References

Abraham, S. (1992). *Achieving high response rates with institutional populations: Patterns of response among eighth, tenth, and twelfth grade students in a statewide school-based survey.* Unpublished paper presented at the American Association for Public Opinion Research, May 1992 Annual Conference, St. Petersburg Beach, Florida.

Beck, S. (1984). A comparison of children who receive and who do not receive permission to participate in research. *Journal of Abnormal Psychology, 12*(4), 573–580.

Ellickson, P. L., & Hawes, J. A. (1989). An assessment of active versus passive methods for obtaining parental consent. *Evaluation Review, 13*(1), 45–55.

Esbensen, F., Deschenes, E., Vogel, R., West, J., Arboit, K., & Harris, L. (1996). Active parental consent in school-based research: An examination of ethical and methological issues. *Evaluation Review, 20*(6), 737–753.

Esbensen, F., Miller, M., Taylor, T., He, N., & Freng, A. (1999). Differential attrition rates and active parental consent. *Evaluation Review, 23,* (3), 316–335.

Frame, C., & Strauss, C. C. (1987). Parental informed consent and sample bias in grade school children. *Journal of Social and Clinical Psychology, 5*(2), 227–236.

Josephson, E., & Rosen, M. A. (1978). Panel loss in a high school study. In D. Kandel (Ed.), *Longitudinal Research on Drug Use, Empirical Findings and Methodological Issues.* Washington, D.C.: Hemisphere.

Kearney, K. (1983). Sample bias resulting from a requirement for written parental consent. *Public Opinion Quarterly, 47,* 96–102.

Lueptow, L., Mueller, S., Hammes, R., & Lawrence, S. (1977). The impact of informed consent regulations on response rate and response bias. *Social Methods and Research, 6*(2), 183–204.

Mobers, D. P., & Piper, D. L. (1990). Obtaining active parental consent via telephone in adolescent substance abuse prevention research. *Evaluation Review, 14,* 315–323.

Severson, H., & Ary, D. (1983). Sampling bias due to consent procedures with adolescents. *Addictive Behaviors, 8,* 433–437.

Singer, E. (1978). Informed consent: Consequences for rate and response quality on social surveys. *American Sociological Review 43,* 144–162.

Thompson, T. (1984). A comparison of methods for increasing parental consent rates in social research. *Public Opinion Quarterly, 48,* 779–787.

Design and Methodological Issues in a National Longitudinal Study of Children in the Child Welfare System

Kathryn Dowd, Paul Biemer, and Michael Weeks

Overview

The National Survey of Child and Adolescent Well-Being (NSCAW) is designed to address crucial program, policy, and practice issues of concern to federal, state, and local governments and child welfare agencies. It is the first national study of child welfare to collect data from children and families, and the first to relate child and family well-being to family characteristics, experience with the child welfare system, community environment, and other factors. The major research questions the study will address include:

- Who are the children and families that come into contact with the child welfare system?
- What pathways and services do children and families experience while in the child welfare system?
- What are the shorter- and longer-term outcomes for these children and families?

The study is sponsored by the Administration on Children, Youth, and Families (DHHS) and is being conducted through a contract with Research Triangle Institute (RTI) and subcontracts with the University of California at Berkeley, the University of North Carolina at Chapel Hill, and Caliber Associates.

Study Design

The NSCAW data set will include 6,700 children, ages birth to 14, who have contact with the child welfare system within a one-year period that began in August, 1999. These children will be selected from two groups. Six thousand will be interviewed from those entering the system during the reference year (August 1999–July 2000), and the remaining 700 will be interviewed from among children who have been in out-of-home placement for 12 months at the time of sampling. These 6,700 children will be selected from 100 Primary Sampling Units (PSUs) in 105 counties nationwide. The children entering the system will include about 5,400 cases that enter through investigation or assessment, as well as approximately 600 cases that enter through other pathways (e.g., services provided without investigation or assessment, children on probation, persons in need of supervision, and children of families who voluntarily seek child welfare services.) The sample of investigated/assessed cases includes both cases that receive ongoing services and cases that are not receiving services, either because they were not substantiated or because it was determined that services were not required.

Four annual rounds of face-to-face interviews or assessments will be conducted with children, parents, and nonparent adult caregivers (e.g., foster parents and custodial kin caregivers). Data collection will begin in October 1999, with annual follow-up interviews in 2000, 2001, and 2002. Telephone interviews with parents or caregivers between annual assessments will be used to update information on services received. Both children who remain in the system and those who leave the system will be followed for the full study period. However, the 600 children entering the system through non-CPS pathways will be interviewed only in the first round of data collection; in each subsequent wave, we will determine from the agency whether that child has reentered the system during that 12-month period. The purpose for including non-CPS children in the first wave of data collection is to obtain nationally representative data to describe this diverse group and to relate their characteristics to those of CPS children.

The NSCAW is a longitudinal study with multiple informants associated with each sampled child, in order to get the fullest possible picture of that child. Table 1 summarizes the data collection plan for the entire study.

Instrumentation

The instruments selected and developed had to be able to answer the key research questions as well as the subquestions and the specific analytic questions identified by RTI, subcontractors, ACYF, and Technical Work Group (TWG) members. Table 2 summarizes the constructs that will be measured in NSCAW by each of the five data sources: child, caregiver, teacher, caseworker, and agency informant. The instruments have been prepared for computerization and assembled into interviews for each of the survey informants, resulting in six interviews: current caregiver, former caregiver, child, teacher, caseworker, and agency personnel.

The authors are at Research Triangle Institute, Research Triangle Park, North Carolina.

Table 1. Summary of timing of interviews, by type of respondent

	Years of Data Collection						
	1999–2000		2000–2001		2001–2002		2002–2003
Respondent	Baseline	6 Months	12 Months	18 Months	24 Months	30 Months	36 Months
Child	X		X		X		X
Current caregiver	X	X	X	X	X	X	X
Former caregiver (when applicable)	X	X	X	X	X	X	X
Caseworker	X	X	X	X	X	X	X
Teacher/day care provider	X		X		X		X
Local agency administrator	X		X		X		X
State agency administrator	X		X		X		X

Table 2. NSCAW areas of inquiry by data source

Children
All children
- Cognitive skills
- Language
- School achievement
- Behavior problems
- Mental health
- Relationship with peers and adults
- Attitudes and motivations
- Exposure to violence

Older children
- Delinquent behavior
- Sexual behavior
- Substance abuse
- Maltreatment history
- Services received

Caregiver
About child
- Health & disabilities
- Services received
- Daily living skills
- Social skills
- Temperament
- Behavior problems
- Disruptions in living environment

About themselves
- Mental health/substance abuse
- Physical health
- Services received
- Relationship with child
- Disciplinary techniques
- Social support

About family/community
- Domestic violence
- Neighborhood environment
- Parental criminal involvement
- Demographics

Teacher
About child
- School achievement
- Services received
- Attitudes and motivations
- Social skills
- Relationship with peers
- Behavior problems

Caseworker and agency representative
Risk assessment for child and family
Caseworker characteristics and attitudes
Services for child and family, including:
- Services received, including source and amount
- Reasons some services were not received
- Child placement and placement changes during time in the child welfare system

Agency information
- Structure and resources
- Policies and programs
- Organizational culture

Sampling

The target population for the NSCAW is the union of two subpopulations within the child welfare system:

(a) All children who are subjects of child abuse and neglect investigations (or assessments) conducted by Child Protective Services (CPS)

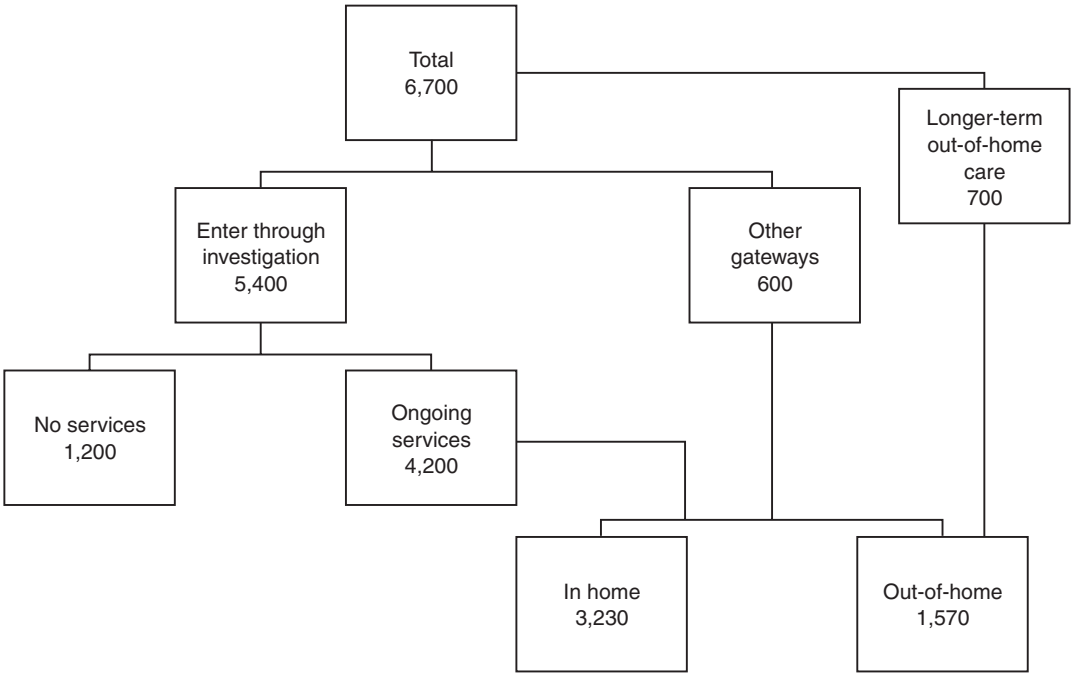
(b) All children who receive child welfare services, whether or not they were subjects of CPS investigations (or assessments)

Thus, group (a) is restricted to children who are reported to CPS and who are subjects of either an investigation or assessment of child abuse or neglect. Although some of these will go on to receive services, group (a) also includes cases that are not substantiated and cases that are substantiated but do not subsequently receive services. Subpopulation (b) includes CPS cases that subsequently receive services, as well as cases who enter the child welfare system through non-CPS pathways. Some examples of non-CPS children are children who do not have a parent because the parent is incarcerated or is in treatment for substance abuse (dependency cases); status offenders; children on probation or persons in need of supervision (PINS); and children of families who voluntarily seek services. Figure 1 shows the final sample design and the number of completed interviews targeted from each component of the sample.

First-Stage Sample Selection (PSUs)

The definition of a primary sampling unit is a fundamental aspect of any PSU sample selection method that might be considered. However, since the administrative structure of the child welfare system varies considerably across the states and even within states, it was necessary to begin with a simple definition of the PSU and then refine this definition for the sample of places that were selected. The geographic area defined by the PSU had to be large enough to support the second-stage sample selection requirement of 67 cases per year (the minimal sample size required from each PSU), but not so large as to present operational difficulties that substantially increase data collection costs. Our initial definition was simply the county or equivalent unit. For most areas of the country, the best definition of a PSU is the county, since it corresponds to a clearly defined political entity and geographic area of manageable size. In other areas, the definition of a PSU is not as straightforward. In the process of contacting the counties selected for NSCAW, we determined that further partitioning and combining of sample counties was necessary for efficient

Figure 1. Components of the NSCAW sample



within-PSU sampling. For example, in several PSUs a single child welfare agency has jurisdiction over several counties or over a special entity such as an Indian reservation, and the PSU is defined as a part of or the entire area over which the child welfare agency has jurisdiction. As part of the agency recruitment phase of the study, we inquired about these special situations and resolved any PSU definitional issues on a case-by-case basis.

Obtaining accurate information for each PSU regarding the number of children in each of the sampling domains was essential to the successful implementation of the sample design. Thus, we contacted all state agencies to request copies of readily available data items we could use for sampling. For the most part, these data are descriptive statistics of the child welfare population in the state for counties or other geographic entities. All states were able to provide at least the total number of investigations for the most recent year from data already available in the state. Some states were also able to provide detailed breakouts of these numbers by age and type of abuse. For states that are providing the Detailed Case Data Component (DCDC) to the National Child Abuse and Neglect Data System (NCANDS), the data items were obtained from NCANDS.

After collecting these data, the states were categorized into two groups, or strata:

- *Full data stratum:* States having all the data required for the sample design
- *Partial data stratum:* States with only partial data or none of the data required for the sample design

Strategies for sample design were developed to treat each stratum differently.

The county-level data that we obtained from states were used for the calculation of the composite size measure. For the Full Data Stratum, we were able to calculate the composite size measure using recent data on the domain population sizes. For the states that could provide only part of the data requested, we used the incomplete data and imputed the domain sizes from the states in the Full Data Stratum and whatever other data were available in the Partial Data Stratum.

The PSU sampling frame was divided into nine major strata. The eight states selected for state-level estimates constituted eight of the strata. The PSUs in the remaining states were grouped into one stratum, depending on whether they had sufficient data for constructing accurate composite size measures. The PSUs were allocated to the nine major strata to ensure that (1) the number of PSUs were approximately proportional to the aggregated size measure of the stratum; and (2) estimates of reasonable precision can be made for the eight states.

Finally, the PSUs were implicitly stratified by urbanicity for those in the eight state strata, and by region, state and then urbanicity for the PSUs in the remaining states. The urbanicity of a PSU was defined by whether it was part of a Metropolitan Statistical Area (MSA); Census region was used for the geographic region designator. An independent sample was then drawn from each first-stage stratum with probability proportional to size (PPS) using systematic unequal probability sampling. Because of their sizes, multiple PSUs were

selected from each of the very large counties. Thus, the selection of 100 PSUs resulted in a sample of 96 counties.

Within-PSU Sampling

As the sample agencies were recruited, we worked with them individually to refine our projections of the expected sizes of the domains of analysis for sampling. From these projected domain sizes, the initial sampling rates by domain were specified. Software was developed that applied these sampling rates to the domains during the 12-month second-stage sampling period.

Two different systems were developed for the within-PSU sampling. One, the File Transfer (FT) system, is being used for PSUs that can and are willing to transmit files and information needed for constructing the within-PSU sampling frame in electronic format. The other is being used for all remaining PSUs. The FT system (1) formats the files provided by the sites into usable form; (2) constructs the sampling frame for the current time period; (3) unduplicates records of the frame of the current time period with those from all previous months; (4) selects children according to the specified sampling rates; and (5) delivers the selected sample to the survey control system.

The other system is a computer-aided data entry (CADE) system that allows the sampling frame to be constructed in the field. With this system, the field representatives enter the information needed into a laptop computer, construct the sampling frame, then transfer the file to the RTI central office for sampling. Adequate quality control is built into the system to minimize errors during data entry and frame construction.

The second-stage sampling period began in early September 1999 and will continue until August 2000. The sample will be selected in segments on a monthly basis during this period. Sample children will be selected from those cases for which the investigation/assessment was completed in the previous month. In addition to investigated cases, cases that were not investigated but began receiving services in the previous month will also be recorded for second-stage sampling. Further, only children who are less than 15 years of age at system entry will be eligible for the study. Thus, this list of investigated children and children receiving services will constitute the second-stage sampling frame. Care will be taken to ensure that each child eligible for the study is listed once and only once on the frame and that only children who are eligible for the study are listed.

Once the relevant data on each eligible child have been recorded in the computer's database, the new entries are grouped by domain and sampled at a rate that yields the required number of sampled children in each age group. Initially, the number of children to sample in each PSU will be set at one-twelfth the annual sample size (i.e., approximately 67); however, the initial sampling rates by domain will vary from PSU to PSU and will depend upon the size measure for the PSU. Each month, RTI statisticians will review the sample yields by domain for each PSU and determine whether the sampling rates should be modified. If so, these modifications to the software will be implemented and the field systems will be updated.

Agency Recruitment

Successful recruitment of state and local agencies required significant preparation, follow-up, and calendar time. Visits were made to all 40 states and 105 counties. To facilitate a well-prepared audience, a read-ahead packet was sent to attendees two weeks prior to a recruitment visit. The packet included the Project Description, Frequently Asked Questions, and a personalized invitation from Carol Williams, Associate Commissioner, Children's Bureau. During the presentation, packets were distributed containing the above materials; copies of the color slide presentation; and materials from the National Clearinghouse on Child Abuse and Neglect Information, the National Adoption Information Clearinghouse, and the U.S. State and Local Gateway. The lead person at each site received a binder with these and other materials including the latest version of the Child Abuse Prevention and Treatment Act (CAPTA), an annotated bibliography of articles on outcomes in the child welfare field, and a guide to program evaluation. These packets served both as sources of information and tokens of appreciation for agency representatives' support.

The presentations to sites included a slide presentation and a question-and-answer period. In developing the presentation, the recruitment team, with help from the field operations and sampling teams, designed materials that specifically addressed two concerns raised by agencies: staff burden and confidentiality. The slide presentation began with the goals of the study and an introduction of the sponsors, participating organizations, and endorsing organizations. Recruiters then discussed the time frame, sampling procedure, site selection procedures, data sources and measurements, and the workload for each respondent group. The presentations concluded with a question-and-answer period.

The recruiting team asked agencies to sign a letter of agreement to participate in the study. The letter of agreement specified the duration of the study, the approximate sample size, the county from which the sample would be selected, the project team's responsibilities, and the responsibilities of the participating site. The letter also required that an agency contact person be designated, who would assist the field representative in preparing for and implementing the study at that agency.

After agencies agreed to participate, recruitment team members continued to maintain contact with the sites to inform them of the project's progress. In May 1999, a maintenance packet was sent to key participants in the sites. The packet included a thank-you note from Carol Williams, a letter from the recruitment team member, a contact sheet with NSCAW and site-specific contact names, and an executive summary of the Phase I Annual Report.

The recruitment team faced many challenges, including agency scheduling conflicts and constraints, political issues, concerns about the study's consistency with agency policies and procedures, workload and burden issues, and confidentiality concerns. These challenges were frequently interrelated. Yet, even as daunting as the task was, the effort was very successful—only six county agencies selected into the origi-

nal sample are not participating. (These six counties were replaced in the sample.) Agency concerns varied tremendously, from common questions about caseworker burden and confidentiality of client data to special problems of certain agencies. For example, one refusing county was in the midst of a lawsuit and very negative press, and one agency refused to participate because privatization of services had decimated caseworker and other agency staff. One county refused to participate because the study design could not accommodate a sample add-on to provide county-level data and more follow-up interviews with cohort members. Project staff frequently joined the recruitment team member on calls with agency staff to answer technical questions about sampling, analysis, human subject protections, and data confidentiality. The federal project officer was also involved in contacting agencies to encourage their participation in the study. Additionally, the extended project team included several researchers with existing contacts in key locations. These associations and professional networks were activated with several of the most recalcitrant agencies, and key decision makers were persuaded to participate. Other connections to child abuse advocates in the communities selected proved similarly effective.

Data Collection

Caseworker

Baseline Risk Assessment

This computer-assisted personal interview (CAPI) questionnaire focuses on the case investigation and the caseworker's assessment of risk at that point in time. Upon selection of the monthly sample, the field representative will identify the caseworker associated with each case and schedule the baseline interview. Because of the detailed nature of many of the questions, we will request that the caseworker have the child's case record available for the interview session. After the call, the field representative will mail a letter explaining the study and a project brochure developed for caseworkers, with a reminder of the day and time the interview has been scheduled. Field representatives are being trained to complete all caseworker baseline interviews within 10 days after the monthly sample has been drawn, in order to collect risk assessment data as close to the close of the investigation or assessment as possible.

Six-Month Interview

After six months, the field representative will interview the caseworker again. The midwave caseworker interview will focus on the services recommended for and received by the sampled child and the child's family, the case history before and after the report, the living environment in the household, caseworker involvement with the family, progress made by the family, and basic information about the caseworker. These interviews will be conducted with the caseworker every six months as long as the selected child is receiving any form of services and has an assigned caseworker.

Current Caregiver

Annual Interviews

This CAPI questionnaire is focused on the child's health, mental health, services received by the child and the family, the family environment, and experiences with the child welfare system. The field representative will contact the adult caregiver and administer the first in-person interview approximately 90 days after the initiation of the investigation or assessment (i.e., an average of 30 days for the investigation or assessment to be conducted and 60 days for services to begin). Annual follow-ups will be scheduled within two weeks of the anniversary of the case closing. The interview is expected to range in average administration time from 95 to 125 minutes for custodial parents and 55 to 90 minutes for foster parents for each in-person interview.

Six-Month Interviews

Approximately six months after the close of the investigation or assessment, the field representative will telephone the household where the child resided at the time of the first in-person interviews. In this 30-minute computer-assisted interview conducted by telephone, the field representative will seek to confirm that the sampled child still resides in that household and to update services utilization data for the interim period since the baseline interview. These interviews are repeated at 18 and 30 months, between in-person annual interviews.

Children

The sampled child will be interviewed during the same visit to the household in which the adult caregiver is interviewed at baseline and at 12-month intervals after the close of the investigation/assessment. Once a signed consent form has been obtained from the legal guardian and the study has been explained to the adult caregiver (who may also be the same person), the field representative will seek assent from children 7 years old or older to conduct a CAPI interview with the sampled child. (Our pretest experiences indicated that children less than 7 years of age typically do not understand the basic concepts underlying the consent process.) The timing of the adult caregiver and child interviews will vary by circumstances and the convenience of respondents; field representatives will schedule both interviews in the same visit to the household when possible.

The interview protocol varies considerably depending on the age of the child. Only physical measures (length, weight, and head circumference) and physical development assessments will be taken from the very youngest infants; older babies will be assessed with standardized measures of physical and cognitive development. Toddlers and young children

will complete several cartoon-based and other simple measures in addition to the physical measures of height and weight. The interview protocol for older children includes questions on physical health, mental health, assessments of cognitive development and academic achievement, and for 11- and 14-year-olds, questions in Audio Computer-Assisted Self Interview (A-CASI) mode about events that led to their involvement with the child welfare system. The A-CASI sections include questions on substance abuse, sexual activity, delinquency, injuries, and maltreatment. The interviews with sampled children will range from 20 to 135 minutes.

Former Caregiver

Data will be collected from the caregiver from whom the child was taken at the baseline and in subsequent follow-ups. The baseline caregiver respondents will be recontacted as long as family reunification is the goal for the case. Data from these former caregivers are critical to understanding the context in which the child lived before their out-of-home placement and to which they will return. It is also important to obtain information on the range and magnitude of services received by these caregivers during their efforts to regain custody of the child.

Teachers/Day Care Providers

The purpose of the teacher and day care provider survey is to obtain an independent measure of the child's academic performance, cognitive abilities, social skills, and relationships with other children. The teacher or day care provider will be identified in the adult caregiver interview. Note, however, that teachers will be contacted only if the signed authorization form was obtained from the legal guardian by the field representative. This will ensure that no teacher will be contacted for participation without the guardian's express approval. The survey of teachers and day care providers will be implemented through a mailed self-administered instrument, with promptings of nonrespondents by mail and telephone.

Response Rates

Given that we have no pretest of significant size to judge the adequacy of data collection procedures, we are uncertain how concerned we should be regarding the response rates we will achieve. Certainly this population has been studied before, and site-based studies of the caregivers of abused and neglected children or children at risk for abuse or neglect have achieved response rates in the mid-80s. We have incorporated the best practices contained in survey methodology literature, and those procedures that have been demonstrated to work effectively in this population in studies of a similar nature, into the data collection procedures for the NSCAW. These measures are summarized below.

- Borrowing from studies of more generalized populations, we have included advance letters, customized refusal conversion letters, specialized field representative training on refusal avoidance and conversion, field supervisor review of noninterview cases, and careful monitoring of noninterview cases by project staff in the data collection procedures.
- We have requested the use of incentives for participating children (in the form of gift certificates to toy and music/video stores), adult caregivers (in cash), and teachers (by check) to defray any costs incurred from participation in NSCAW, and as a token of our appreciation for their contributions to this important research.
- Materials to be used on the project have been assessed in focus groups of caregivers (both foster parents and permanent caregivers) and caseworkers to ensure that the appropriate questions and concerns are being addressed in language that is understandable and that the legitimacy and importance of the study is clearly communicated.
- In addition to staffing communities with significant Hispanic populations with bilingual field representatives, we have designated some more marginally Hispanic but strategically located communities for bilingual staff so that these staff may travel to communities in their region to conduct interviews in Spanish. This approach will minimize the number of interviews lost to language barriers, while minimizing data collection costs.
- Gaining the cooperation of selected children and their families is emphasized in field representative training. The training protocol includes both discussion of the decision whether or not to participate and various exercises to ensure that field representatives are very comfortable introducing the study and answering potential respondents' questions and concerns.
- Further, field supervisors will play an unusually active role in troubleshooting on cases in pending noninterview dispositions. These very experienced staff will assist the field representative in determining which strategies are most likely to result in completed interviews and will become personally involved in converting refusals, including customizing and sending refusal conversion letters and making follow-up calls to reluctant families.
- Throughout the agency recruitment process, we have concentrated on developing close, collaborative relationships with the participating agencies, because staff in these agencies will be an invaluable source of information as we approach families for participation in NSCAW.
- Adapting from procedures used on a longitudinal study of over 2,500 children and families in the child welfare system, we will send birthday and holiday greetings from the project staff to children and caregivers in the cohort established at baseline in order to maintain their commitment to the project.

Human Subject Protection and OMB Reviews

Aware that many would deem abused and neglected children to be the most vulnerable of all possible research subjects, the project went about design and development of data collection procedures with great sensitivity to what we thought would be the issues of concern. The project team established a Human Subjects Work Group, led by a psychologist who has conducted research with children and adolescents and who chairs one of three Institutional Review Board committees at RTI. On the work group was a previous chair of an RTI IRB committee and one present member. A pediatrician and the survey manager rounded out the membership. Reviewing the work of the group were no fewer than three ex-IRB committee members. The conclusions of the project were also reviewed by members of the Technical Work Group, all experienced in research with children abused and neglected.

The Human Subjects Work Group recommended a conservative balance between protecting study subjects, second-guessing just-completed investigations by professional social workers, and "doing no harm" to participating families. Consent and assent forms were very carefully constructed. Questions eliciting information about the most serious types of physical and sexual abuse, asked in A-CASI, are programmed to probe for information that will allow us to distinguish between prior abuse that generated the report and ongoing abuse. These response patterns will be transparent to the interviewer, transmitted back with all other questionnaire data, and then reported by project staff from North Carolina. Interviewers are being trained on the specific laws governing reporting of abuse and neglect in their state and will be free to follow their conscience regarding observed or unsolicited information that might indicate ongoing abuse or neglect. Well-tested procedures for handling indications of suicidal intent are also included in the data collection procedures.

Even these preparations were insufficient. As the project officer recently wrote in response to an Office of Management and Budget (OMB) suggestion to change the wording in the consent forms, "We are not aware of any other behavioral sciences project, at any institution, that has received a level of scrutiny that even approaches the attention that was given to the NSCAW project. This is, of course, a unique project with many complicated issues, and we do not feel that the attention was unwarranted." The project officer's comment described a seven-month, 13-meeting process with the full committee and a subcommittee specially created to work with the NSCAW project team.

In parallel, the review by OMB raised issues and requirements in direct conflict with the mandates of the IRB committee and subcommittee. While OMB's concerns—for reaching an acceptably high response rate and obtaining accurate self-reported data in the most sensitive portions of the questionnaires (administered by A-CASI)—were valid, the direct conflicts had to be negotiated in a way that allayed the concerns of both while not compromising the rigor and generalizability inherent in the study design. Ironically, a "privacy review" conducted by staff from OMB's Office of the Special Counsel for Privacy found objection to requests for data routinely used in longitudinal studies for locating the members of the cohort. And, with a total of seven, NSCAW must have won a prize for the project with the greatest number of conditions for clearance; one is currently being appealed.

Everyone who has even the most peripheral rights to review the project has seemed to want to leave a mark, and demand or suggest changes to consent form language or to procedures for reporting of child abuse or neglect or other types of abuse. Even the application for a federal Certificate of Confidentiality generated suggestions for changes to the language in the consent forms.

Analysis Plans

We have prepared an analysis plan that summarizes current plans for analyzing the wealth of data that will flow from NSCAW. The plan identifies the major research questions that will be addressed in the study, the data elements that will be used to answer the questions, and the types of analysis to be employed in addressing the research questions. Following each wave of data collection, data from the survey will be analyzed by the project team. Additionally, after being stripped of identifying information and analyzed for the possibility of inadvertent disclosure, data sets from NSCAW will be made available to the larger research and policy community to encourage secondary analyses that will support further research and timely policy decisions.

Our analyses will focus on the key study issues described above and summarized in Table 3. Examples of the cross-sectional and longitudinal analyses to be performed include

- Description of characteristics and risk factors for children and families at the point of entry into the child welfare system, overall and for subgroups (e.g., CPS and non-CPS cases)
- The investigation/assessment process (e.g., risk factors, decisions, family involvement)
- Children's and families' experience of child welfare and other services and of changes in services and placements during the period in the child welfare system
- The process of permanency planning and implementation for children in long-term out-of-home care
- Description of children and families who leave the system quickly and those who stay in for a longer period
- Analysis of the relationship of child, family, caseworker, agency, and other factors to child and family services and outcomes
- Analysis of how the organization, structure, and resources of agencies relate to the services provided and to whom

The primary focus of the study is on children and families; however, because data are collected from child welfare agencies,

it will also be possible to conduct some limited analyses at the agency level. Agency level data (e.g., staff turnover, use of dual tracking, budget) and caseworker data (e.g., level of experience, specialized training) will be used in analyses of child and family services and outcomes. In addition, data collected during the sampling process will be used to describe such aspects of the child welfare system as outcomes of completed cases (e.g., substantiation rates) and the disposition of substantiated cases (e.g., rates of case opening, placement rate), overall and for subgroups.

Table 3. Examples of questions NSCAW will address

Who are the children and families who come into contact with the child welfare system?
What are their backgrounds and characteristics?
What are their prior histories?
What problems and strengths do they bring?
How do the characteristics, experiences, and needs of children and families differ by the ways they come into contact with the system?
What effects do state and agency policies and programs have on the characteristics of those who enter the system?
What pathways and services do children and families experience while in the child welfare system?
What placements and services do they experience while they are in the child welfare system?
What determines the different pathways, placements, and services they experience?
How do child welfare services interact with other services and supports for children and families involved with the child welfare system?
What are the shorter- and longer-term outcomes for these children and families?
How do children and families change during the time they are in contact with the child welfare system?
How do children and families change after they leave the system?
How do child, family, system, community, and other factors influence child and family functioning?
How do these factors affect subsequent child welfare system involvement?

Comments on Sampling Issues in Collecting Data from Children and Adolescents

Sandra H. Berry

Introduction

Research on children and adolescents is very important for health promotion. Children and adolescents have health issues that affect their own lives and the lives of their families. In addition, their health behaviors put them on track for better or worse health in the present and in the future. For example, care of asthma and diabetes is an important concern for many children while they are young, and their participation in risk behaviors, such as smoking, alcohol use, drug use, and sexual behavior are also concerns. It is entirely appropriate that we ask children about their own health and health-related attitudes and behaviors, and this session is devoted to papers that report on studies that do exactly that.

First, let me congratulate the authors in this session on their fine papers. Each represented very careful and thoughtful work, and each was well written. I learned from all of them. My fellow discussant and I conferred in advance about what aspects of these papers to discuss. This turned out to be easy; I was interested in sampling and access to respondents, and he was interested in how they responded once the researcher got to them, so this discussion virtually ignores what respondents said—I leave that to him. I am going to talk about each of these papers, then draw some common themes. Since you have just heard the papers, I will not tell you what they said. Rather, I will focus on some comparisons from the sampling and access perspective.

First a quick overview of study goals to put things in perspective. Gallagher, Fowler, and Elliot conducted a randomized trial of contact and interviewing procedures on a probability sample of teens. Klein, Graff, Santelli, Allan, and Elster conducted a study of validity and reliability of self-reported data vs. record data about medical visits among teens. Riley, Rebok, Forrest, Robertson, Green, and Starfield conducted a series of cognitive interviews to evaluate appropriateness of measures for different age children. Hawes-Dawson, Zellman, Cotton, and Eisen report on the results of contact procedures for obtaining consent from parents of school children. Dowd, Biemer, and Weeks are describing the design of a planned national probability survey of children and related others who were involved with the child welfare system.

Sandra H. Berry is at RAND, Santa Monica, California.

Sampling Frames

The goal of sampling is to provide a reasonable representation of the population of interest so that generalizations can be made from the sample to the population. Each one of these papers is based on a survey sample, and they differ in how they were developed and used. However, each is seeking to inform us about methodological issues based on their samples, so it is of interest to compare their approaches. Let's start with the sample frames. The goal of a sample frame is to include an unbiased represention of the population of interest. The Gallagher et al. paper is about the 20% of the Medicaid population that are teenagers, aged 13–17. They use as their frame Medicaid data in Massachusetts, adjusted by the availability of contact information. Klein et al. are interested in teens aged 14–21 who visit primary care practices. Their sample frame was teens who visited 15 practices in upstate New York. Riley et al. were interested in children aged 14–21; they developed their frame from day care and medical clinic settings, and I am guessing these were around Baltimore, MD. Hawes-Dawson et al. were interested in 6th-graders attending public schools and based their study in 12 schools in Los Angeles. In terms of representativeness, Dowd et al. are clearly the most ambitious, attempting to develop a nationally representative sample of children from birth to age 14 who have had contact with the child welfare system.

I want to be careful to point out that none of these papers argued beyond their data, and they had very different levels of interest in the representativeness of their samples. Gallagher et al. actually used two frames, one that included all cases and one that included only those with adequate contact information. They accounted for the disposition of all of those cases in each frame, but it would have been interesting to see how many cases were excluded from the frame. Also, since they had a more and a less inclusive frame, would it be possible to identify cases in the more inclusive frame that would not have been in the less inclusive frame and to obtain a picture of those who were excluded because of incomplete contact data? Klein et al. provided some information about their frame, but it would have been useful to have even an anecdotal perspective on what kinds of patients these populations served in terms of income and education levels as well as insurance status. Once they were in the practices, they did a careful job of accounting for the numbers approached, the

number who agreed, and the number who participated or were lost at each step. Riley et al. were less interested in sampling issues. It would have been useful to know more about the populations served by these centers or clinics, how many children there were, how many were asked to participate, how many agreed, and how many were actually interviewed. Even though the focus is on carefully conducted cognitive interviews, it is useful to be able to make an informed judgment about how the results of this study might generalize to other populations. Hawes-Dawson et al. seemed to have a prespecified sampling frame of 6th-graders in selected schools. It would be useful in the context of this paper to know a bit more about these schools (e.g., test scores vs. the range in the LA unified or racial/ethnic distributions). Also, what had gone on before? Were these schools part of an ongoing demonstration program? As a user of the results, I want to understand what they represent. One complication that they allude to but do not develop is the difficulty of establishing a frame for school-based samples. Computer systems are very imperfect; things change; and from a sampling perspective, schools are often not really able to tell you who should be there at any point in time. This is also true of Medicaid files, where eligibility changes and files are often out of date.

Dowd et al. are vitally interested in their sampling frame—that is the main focus of their paper. They are implementing a difficult two-stage design, first sampling PSUs, then building lists of eligibles within PSUs and selecting them according to a stratified design, then attempting to locate and contact selected eligibles. The attention to the sampling frame is considerable, and building it will be very difficult. Once the PSUs are selected and explored (and six counties had to be replaced), the list building is a formidable challenge. They are using both existing computer lists and lists they are building in the field, and they are bound to be of inconsistent quality. A part of describing the frame will be accounting for the quality of the lists and describing how they were constructed—good nighttime reading for persons with sleep impairments, but important in assessing the representativeness of the sample.

Access

Once you have a frame, the next critical issue is whether you can actually get to the children or teens that are included, which brings us to the role of gatekeepers. There are two kinds of gatekeepers who have an influence on the outcomes: formal and informal. Formal gatekeepers are the Medicaid departments, practices, after-school care programs, clinics, schools, or child welfare agencies that must cooperate, as well as parents, who normally must give permission for individuals under the age of 18 to participate in research. Informal gatekeepers are the lawyers and review committees that generally specify the formal conditions under which you can obtain access to respondents, as well as the on-the-ground staff of agencies, schools, and so forth, whose cooperation you must have on a day-to-day basis in order to get the task done.

Gallagher et al. make little reference to these factors, so we can assume that review committees were friendly and that agency cooperation was adequate. Parents acted as specific gatekeepers for their teens in only 2% of cases, but you might infer that nonresponse by parents probably contained some component of gatekeeping. Klein et al. obtained two IRB clearances, one from the CDC and one from the University of Rochester. It appears that they obtained written consent from parents and directly from mature teens and 75% agreed to participation. Riley et al. make little reference to these issues. Their study was apparently cleared by one IRB, and if they encountered other issues, they are not described. Hawes-Dawson et al. focus on gatekeepers, specifically schools and parents, and how to work with them in an active consent situation. Ultimately, they obtained a response to the request for study participation from 77% of parents and 61% agreed to cooperate. The school record systems and the reluctance to release phone numbers dictated that contact with parents had to be through the students, instead of by mail or phone, so some of the 23% who did not respond may simply have never received the materials. Dowd et al. are also very concerned with formal and informal gatekeepers. They have encountered numerous review committees and have worked the problems of informal gatekeepers very hard, with an ambitious program of outreach and incentives. As we speak, they may be encountering some of the other constraints on how respondents may be tracked and located, how the study must be introduced and consent obtained, and how the data collection must take place.

Coverage of the Frame

So where does this lead us? Once you establish how well the sample frame represents the population, the next question is how well the completed sample represents the frame. Let us compare these studies. The Gallagher et al. study obtained interviews with about 33–40% of teen respondents whose parents were also interviewed. Klein et al. obtained responses from 59–61% of respondents, taking into account all sources of sample loss. Riley et al. do not report any response rates. Hawes-Dawson et al. ended up with surveys from 63% of the students and Dowd et al. project an 80% response rate, however this is easier to project than attain and does not take into account any sample loss due to problems with list building. None of these rates approach complete coverage, so there is work to be done both on improving response and also on characterizing nonresponse and how it may introduce bias into the results.

Formal Review Processes

Having summarized the papers from a sampling perspective, I would now like to turn to some of the general issues they present, starting with the formal review processes that

are important in research on children. Normally, such studies are subject to various kinds of legal reviews, since there is state and national legislation governing research on children. Depending on the institutions funding them, they may be subject to one or more IRB reviews. This may include review by the funder, by the organization carrying out the research, and by other organizations involved in the research as sites. In addition, there may be other reviews, such as OMB review. Each of these reviewing bodies will feel a special need to protect the rights of children, and in the absence of clear guidance about how to translate these concerns into practice, there is the potential for conflict among them, such as experienced by the RTI study. In addition, the roles of reviewers are at odds with each other. Legal review is often focused on protection of the institution from liability as well as protection of data confidentiality, IRB review on ensuring the rights of research subjects, and OMB review on scientific quality as a justification for respondent burden. There are tensions. For example, there is a tension between the need for high response rates and the constraints on the contact procedures and the incentives required by the IRB or legal counsel. The Hawes-Dawson et al. paper raises the issue in the form of implicit vesus active consent to conduct research in school populations and the effects on response rates and operational feasibility of an active consent process. The same kinds of issues may come up in the Dowd et al. study.

Obtaining Cooperation

Another issue raised by these papers is the need to build buy-in to engage the institutions that provide access to child respondents. All of the studies had to invest time and resources in gaining access to a sample. In some cases this effort was considerable. Obtaining samples of recipients of public program support is growing increasingly difficult as confidentiality concerns grow more prominent. Other institutions, such as schools, clinics, and public agencies, often receive many requests for research participation and take seriously their responsibilities to parents and children who use their services. Working with them takes time for their various levels of review and response and usually requires the participation of project leaders to represent projects and to negotiate the thorny issues at the intersection of research, human subjects, and institutional needs. Obtaining good response rates at this level is important. The sites that see the value of research and are willing to participate often are quite different from the ones that refuse. Often the refusals seem to be the places that are struggling to survive, and excluding them may produce bias in the resulting sample.

Participation of Agencies

I would also like to mention the issues about building participation once initial buy-in is obtained at a cooperating agency or institution. This generally needs to occur at all levels in order for a study to work. For example, in the Hawes-Dawson et al. study, principals and teachers were key to obtaining response from children. In the Dowd et al. study, child welfare agency staff are needed to identify and sample the children. The other studies required cooperation from after-school program staff and staff of medical practices. Obtaining such buy-in is delicate, especially with staff who are in direct contact with children for whom they feel a responsibility. It's tempting to emphasize the positive outcomes the study might produce for children, but such outcomes are not guaranteed. Moreover, relying on the study demonstrating some specific result may tend to reduce participation from staff who don't view that result as positive. I'd like to point out that it may not be the researchers who introduce these ideas. Staff may generate them on their own, and it is up to the researchers to clarify goals and objectives.

Another motivation to cooperate is incentives—from money, to supplies and resources, to pizza parties. These certainly work, and there are few who would argue that they have no place in research on children, but they are always somewhat controversial to implement. At what level does an appropriate incentive become coercive? How does it vary with the age of the child or other factors, such as economic resources of the families? To whom should incentives be directed: parents or children, groups or individuals? And what exactly should the incentive be for? In the Hawes-Dawson et al. study, for example, the incentive was for a response to the informed consent request—either positive or negative—rather than for research participation.

Where Do the Children Enter the Picture?

For the most part, from a sampling and access perspective, children present few problems as respondents. Once you get to them, they generally are cooperative and interested in research as long as you keep their perspectives and needs in mind. Typically, few children refuse to provide information once you've gotten past the barriers to asking them. The Gallagher et al. study had more problems in this respect than some of the other studies, but that may have been related to the dual participation of parents and teens in the study. Parents may feel the need to oversee the research process with children. The Dowd et al. study describes some of the procedures they are using to ensure privacy for child and adolescent respondents.

Conclusions

These papers point to the dedication and creativity with which researchers are approaching research on children and adolescents. There are many legitimate concerns about how we work with children as research subjects, and there are substantial barriers to conducting work with children. Some that I have discussed are those that affect the representativeness of samples of children. The work presented here was carefully

done but still presents problems in terms of coverage of the population. We need to work on both how to improve coverage of children and adolescents in research and, recognizing that results will be imperfect, how to handle the problems of noncoverage. This argues that research on children and adolescents needs to be approached with the same rigor that we bring to any other population, taking into account the special problems that arise in working with them.

Advantages and Limitations of Using Children and Adolescents as Survey Respondents

Nicholas Zill

The topic of this morning's session is "Collecting Data from Children and Adolescents." We have had five interesting presentations on topics ranging from young children's ability to report on their own health to strategies for increasing parental consent and protection of human subjects in surveys dealing with sensitive topics such as drug use and child neglect. My comments will focus on the strengths and limitations of using children and adolescents as survey respondents. Thus, the comments are most relevant to the papers by Gallagher, Fowler, and Elliott; Klein and his colleagues from the University of Rochester; and Riley and her colleagues from The Johns Hopkins University. My co-commentator Sandra Berry of RAND has focused on issues germane to the other two papers.

Let me begin with some definitions of age groups. For developmental reasons, it is useful to group young people into four age groups: infants and toddlers (0–2 years old), preschoolers and kindergarteners (3–5 years old), elementary-school children (6–11 years old), and adolescents (12–17 years old).

Because language development is in the very early stages for infants, toddlers, and preschoolers, child researchers have not considered using young people in these age groups as survey respondents. It is possible, however, to do developmental assessments with children even this young as part of a large-scale survey, using specially trained survey interviewers. In the Head Start Family and Child Experiences Survey (FACES), which Westat and Abt are doing for the Administration for Children, Youth, and Families (1998), we have successfully done assessments of children's emergent literacy and numeracy at the beginning and end of the Head Start year, and at the end of the kindergarten year. In the Early Childhood Longitudinal Survey of a Birth Cohort (ECLS-B), which Westat is carrying out for the National Center for Education Statistics (NCES) and a number of other federal agencies, we plan on doing assessments of motor and mental development in infants and toddlers as young as 9 to 18 months, again using specially trained survey interviewers. That study has just entered its field test stage. The main survey will study a national probability sample of more than 12,000 children born in the year 2000, who will be followed into at least first grade.

Nicholas Zill is at Westat.

Many more survey studies have gathered information about national samples of elementary-school children and adolescents, and a number of these studies have administered questionnaires or interviewed the young people themselves (Zill, Sigal, & Brim, 1983; Zill & Daly, 1993). In 1976–77 I directed a National Survey of Children, sponsored by the Foundation for Child Development, that conducted in-person interviews with 2,301 children aged 7–11, and I also collected data about the children and their families from parents and teachers (Zill & Daly 1993, pp. 286–295). A similar study was subsequently done for the National Commission on Children (1991). NORC and Ohio State University have for some years been conducting a Mother and Child Supplement to the Bureau of Labor Statistics' National Longitudinal Survey of Youth (NLSY), doing developmental assessments of children born to the female respondents in the original study, and collecting direct-report information from the older children and adolescents in the sample (Baker, Keck, Mott, & Quinlan, 1993).

Many survey studies of adolescents have been done, such as "Monitoring The Future," the annual in-school questionnaire survey of high-school seniors that has been conducted by the University of Michigan Institute for Social Research since the mid-1970s (Johnston, Bachman, & O'Malley, 1995). There is currently an ongoing survey of Adolescent Health being funded by the National Institute of Child Health and Human Development. Several rounds of the National Household Education Survey (NHES), which Westat has conducted for NCES, have included complementary telephone interviews with parents and their adolescent children, focusing on the topics of school safety and discipline, and citizenship and service learning (Nolin, Collins, & Brick, 1997).

Most child health surveys, however, have used an informed parent, usually the mother, as the proxy respondent. Examples are the Child Health Supplements to the 1981 and 1988 National Health Interview Survey (Coiro, Zill, & Bloom, 1994; Zill & Daly, 1993, pp. 159–189), and the child health portion of the redesigned National Health Interview Survey itself. I would argue that the mother is still the single best informant about the health and health-related experiences and behavior of her child, at least until sometime in adolescence (Zill & Coiro, 1992). This remains true even in an era when a majority of mothers are employed outside the home and a majority of children spend significant amounts of time on weekdays in the care of someone other than the

mother. And it remains true in an era when there is more emphasis on equalizing parental responsibilities and at least some fraction of fathers are participating quite actively in the rearing of their children. Mothers tend to be more interested in the details of their child's development and health, and are more willing and able informants about such things as their child's weight at birth, immunization history, and experiences of illnesses and accidents. If you doubt the truth of this assertion, I suggest you conduct some experiments yourself. Ask both mothers and fathers what the child's current height, weight, and shoe size are, or the names of the child's current teachers, and see who knows.

Clearly, though, there are good reasons for wanting to get information directly from children. Some of these reasons have been mentioned in the papers presented this morning. There are others as well, including the following:

1. The child is the one best informed about his or her own subjective sensations, perceptions, and thoughts. These include such things as feelings of discomfort or pain, hunger, fears and worries, likes and dislikes. It is one thing for the child to know these things, though, and another to get him or her to report them reliably to a survey interviewer. I shall have more to say about that in a moment.

2. The child has the best information about health-related experiences and behaviors that parents are likely not to know about, often because the child was someplace he or she was not supposed to be or doing something he or she was not supposed to do. The types of events that children are likely to keep from their parents include fights or accidents, substance use, and early sexual experiences.

3. The child may be able to supply information about areas of family life about which parents tend to have especially biased views or qualms about reporting. Young people may be more unguarded or frank in their reports about parental arguments, the extent of supervision their parents exert over their television viewing, and similar topics. A variable that has proven to be important in accounting for differences across youths in health-related behavior is the quality of the relationship between the child and each parent—whether the child feels close to the parent, whether he or she can talk with the parent about things that really matter, whether he or she wants to be like the parent when he or she is an adult (Peterson & Zill, 1986; Zill, Morrison, & Coiro, 1993). Although the parent can certainly be asked these kinds of relationship questions, it is painful for a parent to admit it when the relationship is far from ideal. Thus, there are good reasons for believing that the young person is likely to be the better informant about his or her relationship with each parent.

4. The child has better information about peers and peer influences, which are very important in the initiation and maintenance of health-related habits such as smoking, drinking, and drug use, and activities such as unsafe driving, delinquent acts, sexual activity, and teen violence (Zill & Nord, 1994). Children and youth also know more than parents about environments such as school, which young people are exposed to on a regular basis, unlike parents (Chandler, Nolin, & Zill, 1993).

5. It is valuable to allow children and adolescents to be respondents in order to give young people a voice—a chance to speak for themselves (Zill et al., 1983). This is a central aspect of public opinion polling and survey research, and one that ought to be shared with that part of our population below voting age. When young people speak for themselves, they often have perspectives on things that adults find surprising, illuminating, charming, or especially telling. In the words of Art Linkletter, "Kids say the darnedest things." For example, in the National Survey of Children, children were asked to give "the name of a famous person you want to be like." The frequency of TV characters or actors and rock stars in their responses clearly demonstrated the influence of television, popular music, and other mass media on young people's thinking. On the other hand, many children named their mothers or fathers in response to this question. Though this was technically an incorrect response, because the parents were not usually famous, it was obviously a heart-warming one.

Despite these good reasons for asking young people to be survey respondents in health surveys, there are serious constraints and limitations on doing this, particularly as far as elementary-school children are concerned. Despite their considerable virtues, I feel that this morning's papers have underemphasized the difficulties involved in working with survey data from children. These difficulties include the following:

1. There is the need to get parental consent as well as the young person's agreement to be interviewed. As has been mentioned, this often means that you take a hit in response rates, compared with interviewing only the parent. It is also necessary to have sensitivity about the subject matter covered in the youth interview, avoiding topics that are potentially upsetting to the child or offensive to the parent, or handling such topics in a discreet manner.

2. The shyness of some children makes them reluctant to be interviewed by survey interviewers whom they do not know well. Even when such children agree to be interviewed, they are often very reticent in their responses. For similar reasons, young children often give very short or inadequate responses to open-ended questions. Thus, though children's responses to open-ended questions may be charming or especially illuminating, the use of these questions often means you do not hear from a significant segment of the child population.

3. Children and adolescents have limited attention spans, and tire under sustained interviewing or when asked to complete lengthy questionnaires. Once they become restive, they are likely to fool around or to give stereotypic

or random responses to survey questions. This means that special care must be taken to keep instruments relatively short, to take breaks, or to spread data collection over several sessions.

4. The limited language comprehension of children means that more complex terms and concepts simply cannot be covered in interviews with them. This was well illustrated in the Riley paper.

5. Children have cognitive limitations in placing events within temporal reference periods, such as "last month," or spatial areas, like "your neighborhood" or "within a mile of your home." This means that questions involving such a frame of reference cannot be used with children without being greatly modified.

6. Survey responses from young people are less reliable than those from adults. The younger the child, the greater the unreliability. In the National Survey of Children, in which we interviewed a national probability sample of 7–11-year-olds, we found that the average inter-item correlation between responses to pairs of questions that were logically related increased in a linear manner with the age of the child. Test-retest reliabilities were also higher in older children than in younger ones. The increase in reliability with age extends right through the adolescent years. Although Klein did not find a significant relationship between report reliability and adolescents' ages, I believe this is because his respondents were all within a narrow age range. Had the range been broader, I think he would have seen such a relationship.

   Because of this unreliability, it is harder to find significant relationships between independent variables of interest to social scientists and dependent variables based on child responses than is the case with dependent variables based on adult responses. Correlations tend to be low, and you need to have very strong effects or very large sample sizes in order to obtain relationships that meet standards of statistical reliability, let alone account for meaningful proportions of variance.

7. Survey responses from young people are subject to some of the same biases that responses from parents exhibit. One of these is a positivity bias in describing qualities of the child or the childrearing environment in the home. For example, when parents are asked to rate their child's academic standing in comparison to others in the class, using one of five categories ranging from "one of the best students in the class" to "near the bottom of the class," the large majority of parent responses fall in the top two categories. A similar skewness is seen in the response distribution of children when they are asked the same question. By contrast, the responses of teachers display a more symmetrical and normal distribution. Differential responding across racial groups and education groups to questions about negative behavior, such as drug use or delinquency, also seems to occur among adolescents as well as their parents.

8. Although the notion of obtaining perspectives from multiple informants in order to triangulate on the true state of affairs is appealing, we have no well-established method for combining reports when they correlate weakly with one another. The evidence is that parent and teacher, parent and child, and teacher and child reports on the same topic are, indeed, weakly related. Do we take the average, or pay more attention to negative information, no matter what its source? More work is needed to identify fruitful ways of combining survey reports from young people with reports from their parents and teachers.

What I conclude from this enumeration of the strengths and limitations of survey responses from children and adolescents is that there is no universal answer to the question of whether we should include interviews with young people as a regular part of child health surveys. The answer very much depends on the specific focus of the study, and whether it includes variables—such as those mentioned above—for which the young person is a uniquely appropriate respondent. We certainly need more of the high-quality methodological research on child and youth responses that these papers exemplify.

I should also note that many of the considerations that go into the process of making survey questionnaires more suitable for use with youthful respondents are quite appropriate for respondents of any age. These include making sure that the language in which questions are couched is sufficiently clear and can be understood by respondents at varying stages of language development, and ascertaining whether respondents are capable of using reference periods (e.g., "in the last month," "since the start of the school year") in the way the researcher intends them to be used.

Studies like the NICHD Adolescent Health study and the National Household Education Survey show that it is possible to mount large-scale studies involving responses from young people as well as their parents and get acceptable completion rates. The public attention that the findings of these studies have generated shows that the product of such efforts can be of substantial public interest and policy relevance. More research like the studies presented in this session will help put the conduct of future studies on a firmer footing and enable researchers to interpret the findings with greater confidence.

I close with a few comments specific to three of the papers:

With respect to the Gallagher paper, I believe the finding that more negative evaluations of the health plan were obtained by telephone than by the mailed questionnaire may not be as surprising as the author indicated. It may be that parents and youth from the low-income population that was eligible for the plan were more reluctant to write down negative evaluative information that they perceived might be linked to their names and might affect their eligibility status than they were to give such evaluations over the phone.

One thought I had, inasmuch as the authors seem to be leaning toward using the parent as the sole respondent in future questionnaires: For those few items where the youth clearly seemed to be the better respondent, such as in describing the

quality of the relationship between the doctor and the youth, perhaps the parent could be asked to question the youth directly before filling out the answers to these questions. That is, the parent could serve as intermediary between the researcher and the young person. This might be a suitable compromise procedure that could lead to lower costs and better data quality.

With respect to the Klein paper, I reiterate my belief that there would have been a positive relationship between adolescent age and response reliability had a broader range of ages been studied. Also, although the reliabilities obtained by Klein are certainly reasonable, they are far from perfect. Such reliabilities will significantly restrict the degree of relationship that can be obtained with independent variables.

With regard to the Riley paper, I think more attention should be paid to the issue of the unreliability of responses from young children. The authors deal quite fully and skillfully with other difficulties of youthful respondents, but the problem of unreliability is not dealt with sufficiently. It is also important to specify exactly what additional benefit will be obtained from getting health status reports from young children that will not be obtained by talking to their parents. More needs to be said about what the authors perceive to be the potential payoff of their methodological efforts. It would also be good to replicate the studies with a more representative sample of young people, rather than simply relying on convenience samples.

These small points aside, I thought this was an excellent set of papers that contributes much to our understanding of why and how to do studies on child and adolescent health that collect information directly from young people.

References

Administration on Children, Youth, and Families. (1998). *Head Start program performance measures: Second progress report.* Washington, DC: U.S. Department of Health and Human Services.

Baker, P. C., Keck, C. K., Mott, F. L., & Quinlan, S. V. (1993). *NLSY child handbook: Revised edition.* Columbus, OH: Ohio State University, Center for Human Resource Research.

Chandler, K., Nolin, M., & Zill, N. (1993). *Parent and student perceptions of the learning environment at school.* National Center for Education Statistics, Statistics in Brief (NCES 93-281). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Coiro, M. J., Zill, N., & Bloom, B. (1994). Health of our nation's children. *Vital and Health Statistics*, *10*(191), DHHS Publication No. (PHS) 95-1519.

Johnston, L. D., Bachman, J. G., & O'Malley, P. M. (1995). *Monitoring the future: Questionnaire responses from the nation's high school seniors: 1993.* Ann Arbor: University of Michigan, Institute for Social Research.

National Commission on Children. (1991). *Speaking of kids: A national survey of children and parents.* Washington, DC: Author.

Nolin, M. J., Collins, M., & Brick, J. M. (1997). *An overview of the National Household Education Survey: 1991, 1993, 1995, and 1996.* Technical Report NCES 97-448. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavioral problems in children. *Journal of Marriage and the Family, 48,* 295–307.

Zill, N., & Coiro, M. J. (1992). Assessing the condition of children. *Children and Youth Services Review, 14,* 7–27.

Zill, N., & Daly, M. (Eds.). (1993). *Researching the family: A guide to survey and statistical data on U.S. families.* Washington, DC: Child Trends.

Zill, N., Morrison, D. R., & Coiro, M. J. (1993). Long-term effects of parental divorce on parent-child relationships, adjustment, and achievement in young adulthood. *Journal of Family Psychology, 7,* 91–103.

Zill, N., & Nord, C. W. (1994). *Running in place: How American families are faring in a changing economy and an individualistic society.* Washington, DC: Child Trends.

Zill, N., Sigal, H., & Brim, O. G. (1983). Development of childhood social indicators. In E. Zigler, S. L. Kagan, & E. Klugman (Eds.), *Children, families, and government: Perspectives on American social policy* (pp. 188–222). New York: Cambridge University Press.

Discussion Notes, Session 1

David Maglott and Elsie Palmuk, Rapporteurs

The discussion centered on four topics: respondent protections versus data quality, nonresponse issues, child-unique measurement issues, and medical record quality.

Respondent Protections and Data Quality

There are increasing—and conflicting—pressures being exerted upon survey researchers. The Federal Office of Management and Budget (OMB) stresses improved data quality and adequate study response rates. However, more restrictive approaches taken by the Information Review Board (IRB) to protect respondent rights may reduce response rates, and pressures from stakeholders may conflict with either OMB or IRB guidance. The Institute of Medicine (IOM) recently issued a report stressing the representation of ethnic minorities in research, while Congress is considering legislation that restricts data collection. Discussants suggested that these issues need to be confronted and discussed, perhaps by asking for an IOM report that would reconcile the conflicting demands of survey inclusion and respondent protection. Although this issue was raised in the context of youth studies, it has broader implications for the field of health survey research.

Nonresponse Issues

Increasing or high nonresponse rates in adolescent surveys are a concern. Incentives and novel approaches to increasing responses can be expensive, and may not be built into survey budgets. It was noted that sampling error may be only a small part of the effects of nonresponse. Collecting some information from or about nonrespondents was discussed. If there is a master list from which people are queried, it might be possible to learn about nonrespondents. Discussants also wondered about the feasibility of including information on the characteristics of nonrespondents on public-use files so that analysts could make appropriate adjustments. In the national longitudinal study of children in the child welfare system, information on nonrespondents could theoretically be obtained from the caseworker. While the IRB prohibited use of these data because they were obtained without parental consent, it was willing to let the researchers use information that already existed in administrative databases. Another commentator noted that the other side of this issue is the risk to nonrespondents of having data about them being accessed, especially if the data are obtained from other sources.

Regarding the presentation on strategies for increasing active parental consent, the comment was made that nonresponse to parental consent forms could result primarily from parents never seeing the form, judging from the low return rate of consent forms for activities as benign as school field trips. In this instance, it is difficult to see how incentives could help improve survey returns. This raised the question about how the investigators were able to conduct a survey, albeit anonymously, of nonrespondent children. The author explained that the IRB approved the survey, for one year, on the condition that the anonymous questionnaire was outlined in the consent form. Under the approval, consenting children were given the long form of the questionnaire while nonrespondents were given a shorter one, lacking identifier information. The IRB approval took into consideration the fact that nonrespondent surveys allowed all children to complete a survey, rather than stigmatizing children with nonrespondent parents. The use of parent respondent proxies to increase response rates also was discussed. There is a substantial body of research in the area of child psychiatric epidemiology that compares children, adults, and teachers as informants of child mental health status. This research has been summarized in several publications. The studies suggest modest agreement across informants. An area for further study is the influence of a mother's health service experience on her proxy reporting of child behaviors.

Child-Unique Measurement Issues

A potential disconnection in the visual analog scales between the anchoring images and the scalar line was discussed. In the example given, the pictures were of a nonsymptomatic and a symptomatic child, but the line referred to the number of days the child felt that way. Do young children really understand this concept? Dr. Anne Riley noted that the slide showed a very early version of the device, and one that is no longer used. The measurement has switched to discrete circles rather than the continuous line. Discussion ensued on the usefulness of recall periods with children. In several larger studies (approximately 500 minority children), researchers further questioned children who endorsed a particular symptom, to make sure they understood when it occurred and its frequency. While age range is relevant here, the majority were able to voluntarily explain how they knew the event fell within the last four weeks.

Had individuals studied interviewer effect on young children and, if so, how had it influenced responses? Dr. Jonathan Klein noted that they had detected interviewer effect on the consent/participation rate, with some interviewers clearly more able to gain participation, so participant bias did occur. It was noted that any time interviewers talk with children, there is tremendous interviewer "rapport," and that this tends to impact responses. This may especially be seen when interviewers "coach" children in responding. Some researchers use only female interviewers to minimize interviewer effects. A strategy to negate this bias is a highly scripted telephone technique, which may diminish interviewer effects. Dr. Riley agreed that interviewer effect can be a concern with children, and that her team felt the use of closed-ended responses was helpful in this respect. As a general principle, young children don't do well with the "think aloud" format; they really need to be restricted to a limited number of choices.

Dr. Klein addressed comments made by Dr. Nick Zill in his discussion by stating that the well-documented change in cognitive ability around age 14 is not usually incorporated in considering children's abilities to respond to various questions. In his discussion, Dr. Zill combined children 12–17 years of age. Dr. Zill agreed that this was an important point and that developmental information should be taken into account in designing surveys rather than simply relying on "standard" age breakdowns. He pointed out that risky health behaviors all tend to accelerate around ages 15 to 16. Including 11- and 12-year old children with older youths may not be appropriate.

It was suggested that young children would have better recall if dates were tagged to important events rather than having the interviewer impose a uniform reference period such as "last week" or "in the last 4 weeks." Dr. Riley responded that they had used a calendar to show the reference period, identified the reference period by the number of Saturdays, and picked out one or two anchoring events (such as the start of school), but that they were unwilling to give up a standard reference period. The usefulness of a standard reference period was further discussed; one recommendation was that the period should just be "ever."

It was emphasized that the above topics also apply to adults (especially the elderly); that is, we need to pay attention to the level of difficulty of the cognitive tasks we are asking respondents to perform and aim for simplicity.

Medical Record Quality

In regard to the Klein paper on adolescent health care, it was noted that an incidental finding worth publication was the poor quality of the studied medical records, as assessed against the tape of provider/patient interaction. Given the widespread use of medical records as a data source, this was noteworthy.

Policy Challenges for the Future: International, National, and State Surveys

A fundamental assumption underlying the planning for this particular Health Survey Research Methods conference was that a conference held at this particular time would ideally focus on whether and how our current and developing methods and research foci can advance and contribute to the implementation and assessment of a dynamic, ever-evolving health policy agenda. In fact, this assumption is a direct extension of one of the basic objectives articulated by the "charter" members of the initial planning group for this series of conferences—"to identify policy issues that can be addressed by survey scientists"—and pursued in the original conference held at Airle House in 1975 (see the Foreword). While that theme was clearly played out in part in most of the individual sessions held in Williamsburg, a special lunchtime panel discussion was also organized and convened to address this theme more directly. The observations, remarks, and insights of two key experts on these issues from a health policy (as opposed to a survey research or methods) perspective are summarized here.

Policy Challenges for the Future: International, National, and State Surveys

Chair: Lu Ann Aday
Panel Members: Lu Ann Aday, Cathy Schoen

Remarks by Lu Ann Aday

Welcome to the special panel on Policy Challenges for the Future: International, National, and State Surveys. I would like to provide some introductory remarks regarding what I see to be the major transformations in conceptualizing and measuring access to health care in the context of the growth of managed care in both the public and private sectors, as well as the new challenges presented to health services and health survey researchers as a result. Cathy Schoen will then present the lessons and implications for the field of health survey research emerging from the major national policy-oriented surveys that the Commonwealth Fund has either conducted or supported. What, from the foundation's point of view, are the ways that we might best approach the survey research and dissemination process? Then I'd like to briefly outline the design and implementation issues that surfaced in a session I chaired at the Association for Health Services Research meetings this past June, concerning state surveys of health and family insurance coverage.

In terms of the policy changes and research challenges in measuring and monitoring access, one of the principal dynamics, of course, is the changing health care system itself, and the burgeoning blending of the organization and financing of care into increasingly complex arrangements. A framework developed by Elizabeth Docteur, David Colby, and Marsha Gold (1996) as the theoretical underpinnings for a survey of Medicare managed care access, provides instructive guidance for measuring and modeling these changes (Figure 1). As we look at the array of factors considered in that framework, we might view it as pointing to a new direction for access studies—that of turning inward, beyond initial entry to identifying the dynamics and the structure that influence people's choice of plan, their experience with the system, and their willingness or ability to stay in that plan or to move out of it, as well as identifying the intermediate and ultimate outcomes of the resulting care-seeking process.

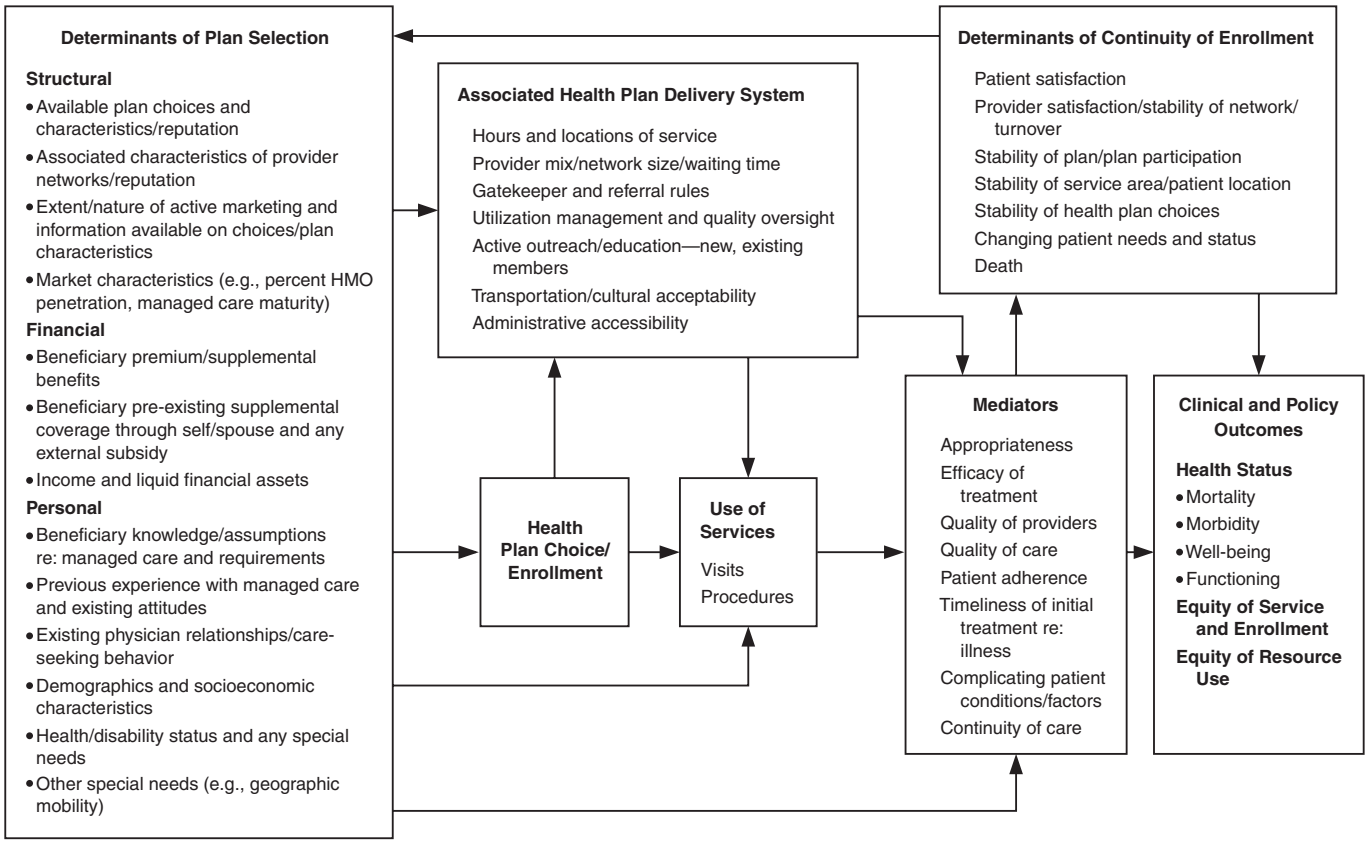
There are a variety of factors influencing plan choice, related to the structure of the plan itself, its reputation, the characteristics of the providers, and the extent and nature of information available. There are financial issues related to beneficiary premiums and supplemental benefits, income, and liquid financial assets that may be available to the individual. And there are personal issues related to the beneficiaries' knowledge of plan attributes or operations. Many types of information may be useful in ultimately predicting who chooses a plan and the rate of use of services governed by plan characteristics. Plan delivery system issues related to hours and location of services, provider mix/networks, waiting time, and gatekeeper referral rules are all presumably determinative ultimately of the outcomes of care, increasingly related to the effectiveness and efficiency of services. The various boxes in Figure 1 profile fundamental plan characteristics in a descriptive sense. The arrows define hypothesized relationships of various plan characteristics to outcomes. We can also envision how the framework could be used for evaluative research to compare the dimensions and performance of various plans.

The question posed by the framework from a research point of view is, Why is it important to know this information? Indeed, who are the audiences for the study? Why would it be important to capture such data in terms of monitoring plan performance or comparing plans? What evaluative agendas might the data inform? What is the meaning, in effect, in the message that we might be trying to capture? Can we use the perspective that such frameworks provide in guiding and shaping more directly the import of data systems such as HEDIS (Health Employer Data Information System) and others that are being developed by the industry and applied in a variety of settings? To what uses will they be put? And who or what are the best sources to provide the required information?

As we look across these indicators, probably one of the first thoughts that comes to mind is how well can individual plan members themselves accurately report that information? They would, of course, be able to report their perceptions, based on their own experiences, but what about some of the structural characteristics of the plan? Those of you who have worked with the Consumer Assessment of Health Plans Survey (CAHPS®) and other plan surveys have had to wrestle very directly with these kinds of questions. How do we best capture the patient's own experience? Are patients always the best sources for providing that information, or does it entail a look at a provider survey or other organizational sources? The answers are ones that we can provide some guidance toward as we consider, as we did this morning, who might be the best informant for certain types of questions.

Lu Ann Aday is at the University of Texas School of Public Health.
Cathy Schoen is Vice President for Research and Evaluation at the Commonwealth Fund.

Figure 1. Framework for monitoring access in managed care



Source: Docteur, E. R., Colby, D. C., & Gold, M. (1996). Shifting the paradigm: Monitoring access in Medicare managed care. *Health Care Financing Review, 17*(4): 5–21.

The first perspective was one of turning inward, trying to look at the dynamics of particular health care plans, and the structure, process, and impact of those plans. A second important dynamic within the health care system is one of turning outward. As we look at the emergence of HMOs and managed care organizations in different markets, what do these changes portend for developing integrated systems of care, in which we try to link the prevention-oriented, treatment-oriented, and long-term care functions served by the health care system? One of the guiding perspectives might be the care needs of vulnerable populations—the mentally ill, homeless, chronically ill and disabled, persons with AIDS, refugee populations, the elderly, and others. These groups require an extended continuum of care that moves outside the bulge that essentially defines the health care system in this country—the acute medical care system—to consider the community resources, the role of public health, and the role of long-term institutional care and home- and community-based services. From the health services research point of view, that continuum encompasses a broader range of services than what many managed care entities and markets are attempting to develop. To what extent has the managed care industry or the evolution of medical care systems in different markets bridged those various components of the system necessary to create a comprehensive, coordinated continuum of care?

What are the research issues of this trend toward turning outward that confront us in health services research and accompanying survey research on this topic? What are the services or programs included in such a continuum? When we think about the role that managed care might play, we also must consider the role of public health and its traditional community focus, as well as the tension within the field of public health over continuing to provide medical care versus moving to more of a broker or bridging role.

A related research question is, What are the possibilities and problems in developing integrated databases to capture these changes? If we think of an integrated delivery system as the model for developing a continuum of care, what is the accompanying data system that might of necessity emerge to evaluate and describe, in effect, the evolution of that system? As we think about the various components that are implied in the development of such a continuum—increasingly, managed care proprietary organizations and resource-constrained public health environments—what are the databases that may be required to monitor and measure what's going on in the

system, and what possibilities and challenges exist for developing and integrating those databases? Who may best provide answers to the given questions, and what data sources might we need to link, to provide the fullest perspective on these transformations?

A third important trend, which has been compelled to address identified weaknesses in the emergence and evolution of managed care–dominated systems of care, is the formation of partnerships with a variety of sectors and providers, particularly to meet the needs of the most vulnerable. A market-oriented perspective on health care reform concentrates on the management of and competition between discrete providers of services. A community-oriented focus seeks to illuminate the distribution of and linkages between providers along a continuum of prevention-oriented, treatment-oriented, and long-term care for all social groups or strata within a community. Evidence that such partnerships are being forged is manifest in attempts by the public health sector to redefine its role in the managed care–dominated marketplace, as well as increasing awareness on the part of managed care entities that some of the problems they encounter as they penetrate selected markets (e.g, victims of violent crimes, child abuse, high-risk pregnancies) are best addressed by broader partnerships with community agencies better equipped to deal with them upstream.

The question posed by the emerging trend of building bridges or partnerships is, Who is served and who is not, in the context of stratification within the community with respect to insurance coverage or socioeconomic status and related (and relative) access to services? A corresponding health survey research issue is how to deal meaningfully and soundly with developing research designs and how to aggregate, analyze, and interpret data gathered at a variety of levels. How do we capture the denominator population in a community? Health plan membership in a given community is unlikely to encompass all of those who are potentially at risk, particularly the uninsured. Further, provider groups may be nested within plans, and within those provider groups there are enrollees, and within those enrollees there are people who are active patients. What types of insights do the respective levels provide, and what sort of comprehensive perspective is required to assess overall system performance and impact?

These emerging trends and issues challenge us to move beyond traditional access studies, which looked at the barriers to entry on the part of individuals, to turning inward to understand the dynamics and processes that affect individuals as they move through selected managed care environments, and then turning outward to identify the system that lies beyond medical care, and the bridges and partnerships being built to extend it. Finally, how do we mirror and model these developments from a survey research and health services research point of view to most informatively illuminate the nature of these emerging dynamics?

Remarks by Cathy Schoen

I'm delighted to be here and to share thoughts about current health policy issues and opportunities for strategic use of health survey research. The topic itself seems risky. I'm the sole survivor out of three, and I was wondering as the other panelists were falling sick, one by one, if there was a health hazard in trying to make survey research policy relevant.

My remarks this afternoon focus on four health survey topics of current and likely future policy concern:

1. Uninsured and underinsured
2. Access to health care
3. Health and socioeconomic status
4. Violence and abuse

To illustrate the potential and challenges of policy concerns for survey research, I've selected findings from several recent Commonwealth Fund surveys and one Robert Wood Johnson Foundation (RWJF) survey.

Each issue, in its own way, presents a common set of challenges.

- Identification of issues and unanswered questions
- Creative questions (often untested) to explore relevant policy issues or concerns
- Capturing multidimensional public experiences and concerns
- Translating research findings in a manner that reaches and resonates with a broad policy and public audience

Communicating results can be more than half the challenge. Whether the audience is the media or the policymakers themselves, insights gained from survey research can be lost if not targeted. Policymakers often have fairly short attention spans, but when they take up an issue, they follow it and look for new information that addresses the concern. Communicating results may mean telling a story in a way that is memorable or that presents a new way of looking which resonates or plants a seed or train of thought that builds along with the policy concerns.

Sometimes it's the personal subtext of the survey—a subgroup of the survey or a cluster—that gives the findings a more human face. Sometimes it's the single statistic or comparison that startles and draws attention.

All four policy issues—the uninsured, access, violence and abuse, and socioeconomic status and health—illustrate the opportunities and challenges to policy-relevant survey research.

The Uninsured

The uninsured remain a central, enduring issue of primary concern to U.S. health policy. With numbers continuing to rise despite a strong economy—now more than 44 million total uninsured—survey research on access, financial distress, and health consequences of being uninsured remains essential to national, state, and local debates.

Myths abound about the uninsured. Despite careful research in the past by many of those attending this conference, public opinion polls and columnist and policymaker comments indicate persistent beliefs either that the uninsured don't need insurance—they are all healthy and manage to stay so while uninsured—or that the uninsured receive an open welcome in the health care system and all get appropriate care when sick. Or that they are all unemployed.

Based on recent polls supported by the Kaiser Family Foundation, the public still sees the uninsured as a top policy concern and would support using budget surpluses to improve coverage (rather than provide tax breaks). Yet linked to the demise of public debate and survey research on negative consequences, public opinion polls today reveal a decline in the percentage of people who believe that the uninsured suffer from lack of access to care. A greater proportion of the population in 1999 think that the uninsured get the care they need than did in 1993.

Surveys addressing policy issues related to the uninsured face the simultaneous task of defining the dimensions of the problem, addressing the myths, and sparking public interest by finding new ways to look at the consequences of not having health insurance.

Defining the Uninsured

- What is the measure of the uninsured? What about spells of uninsuredness?
- What about the "underinsured"? What about policies that omit essential health care services or leave families exposed to unaffordable costs?

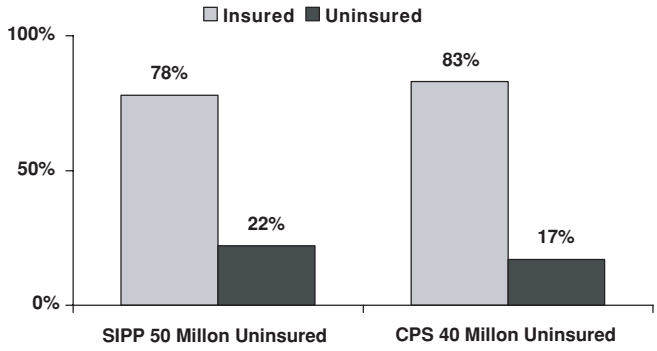
Thanks to longitudinal surveys supported by the federal government, we've known for decades that the number of people counted as uninsured varies widely depending on whether the surveys measure uninsured at a point in time or over a period of time. Yet cross-sectional surveys and national statistics on the uninsured drawn from the Current Population Survey (CPS) typically categorized people as uninsured based on point-in-time estimates. Such estimates ignore those with insurance now who have recently been uninsured.

The definition matters. Defining "uninsured" to include any time without insurance increases the estimate of the uninsured by at least 5%—or 10 to 11 million people—based on recent comparisons of 1995 CPS and Survey of Income and Program Participation (SIPP) surveys (Figure 2).

Definitions also matter for public identification with the problem. In today's economy, with frequent changes in jobs and welfare reform pressures to move families off public insurance, an increasing share of the population is likely to have experienced unstable coverage and gaps in coverage.

Yet survey research on access has had relatively little to say about spells of uninsurance or underinsurance. In part, the silence reflects the expense of longitudinal surveys and the cost of efforts to profile contents of insurance policies. Although the Medical Expenditure Panel Survey (MEPS) will now help provide more frequent estimates of experiences over time, we also need to explore less expensive survey strategies if surveys are to respond in a timely manner to local, state, and regional concerns of the uninsured and underinsured and give the statistics a more human face.

Figure 2



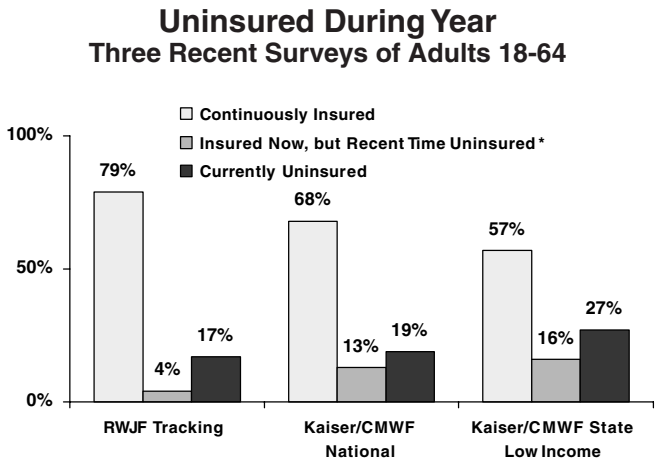
Source: Copeland, C., *EBRI*, June 1998.

Should we be looking at experiences of those who were recently uninsured but are now insured? Can we address this issue with relatively inexpensive cross-sectional surveys? The next three charts draw from some recent Commonwealth Fund surveys and one RWJF cross-sectional household survey to explore these questions.

Several recent Commonwealth Fund surveys and the Robert Wood Johnson Foundation community tracking survey include a question asking those currently insured whether they spent a recent time uninsured. The RWFJ survey asks about a time in the past year; two surveys supported by the Fund and the Kaiser Family Foundation ask about a time uninsured in the past two years—with a follow-up question about the length of time uninsured (Figure 3). Comparing responses for adults ages 18 to 64, the RWJF found about 4% of those currently insured had been uninsured during the year—a rate similar to the difference between the SIPP and CPS estimates. The two Fund surveys both found higher rates as a result of the two-year referent period, with rates notably highest among low-income adults.

Using this "gap" group as a proxy for those with unstable coverage in addition to those currently uninsured, the next few charts compare the access experiences of the two types of "uninsured" with adults who have been continuously insured—with no time uninsured. The findings reveal a striking pattern—on almost every measure of access (except for any visit to a doctor), the "gap" group experiences closely resemble the experiences reported by adults currently uninsured (Figures 4 and 5). The pattern indicates that spells uninsured—even very short periods—can result in access difficulties or struggles to pay for medical care due to lack of ability to pay for care.

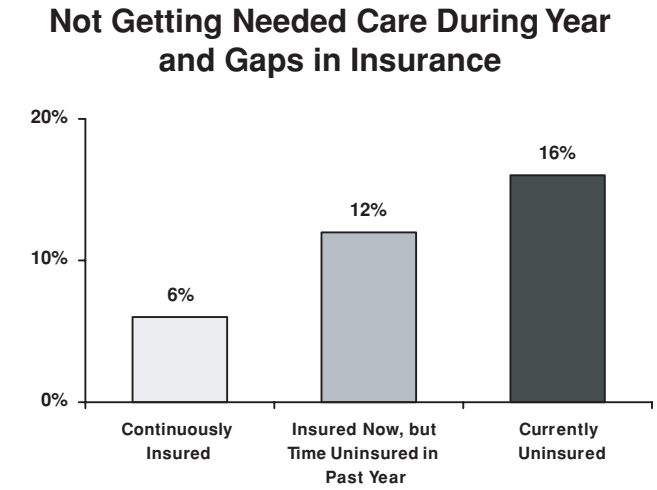
Figure 3



*Recent time uninsured = 1 year for RWJF and 2 years for Kaiser/Commonwealth Surveys

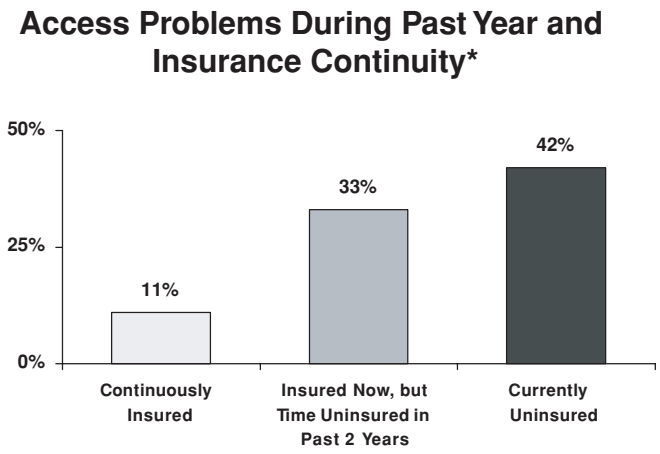
Source: The RWJF 1997 Community Tracking Survey, The Kaiser/Commonwealth 1997 National Survey of Health Insurance, and The Kaiser/Commonwealth 1996-1997 State Low Income Surveys.

Figure 4



Source: The RWJF 1997 Community Tracking Survey.

Figure 5



*Did not get needed care or did not fill a prescription due to cost in past year

Source: The Kaiser/Commonwealth 1997 National Survey of Health Insurance.

The findings indicate the potential for cross-sectional surveys to look at spells uninsured and insurance instability, both for population estimates and to explore the consequences. By separating out the "gap" group, survey researchers can also more clearly illustrate the effect of continuous coverage on access to care.

Indeed, when we listened in on pilot tests of a recent survey of older workers and discussed the issue of spells uninsured with other survey researchers, we heard frequent stories about what happened to them during the one or two weeks between jobs when they were uninsured. In one case, a diagnostic test revealed a severe problem that resulted in surgery for a preexisting condition when the woman finally returned to work and once again gained coverage. Another woman remembered one time during the past 10 years or so when she was uninsured, and went on to say "And let me tell you what happened during that short time . . . ." Having a spell uninsured, at a minimum, appears to heighten anxiety and insecurity and resonates with a broad public audience. Compared with longitudinal surveys, inclusion of a question in cross-sectional surveys that asks those with insurance about a time uninsured has the potential of more timely results as well as addressing a frequent public concern.

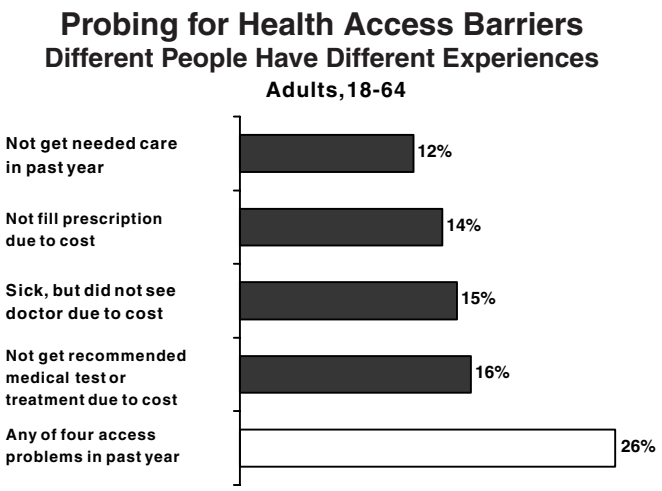
In discussing these findings with the press, we've found that an ability to capture some of these personal stories is often critical to communicating results. Reporters repeatedly ask us for a personal story that brings to life a statistic or survey finding.

Access to Health Care: Multidimensional Concerns

Inadequate health insurance, as well as the advent of managed care and changing health insurance rules and restrictions, raise a host of access concerns that move beyond more simple questions of having a regular doctor or recent physician visits. With barriers at multiple entry points in the health care system, access is a multidimensional concern. Some people may have had one problem and not the other. To capture these personal experiences, surveys exploring access typically need to include an array of questions.

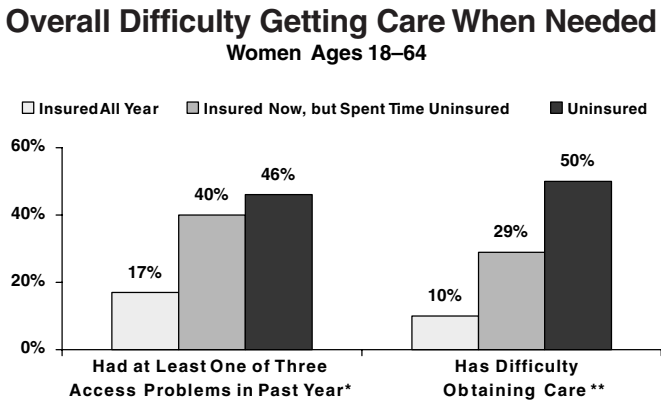
Responses to questions about access "problems" also indicate that access perceptions are often subjective, with expectations conditioned by past experiences. As a result, low-income and minority populations, and others who have little expectation that they'll get anything out of the health care system, often answer "no" if you ask the question, "Was there a time you

Figure 6



Source: The Commonwealth Fund 1999 Workers' Health Insurance Survey.

Figure 7



*Had a time they did not get needed care, specialist care or did not fill a prescription because of costs

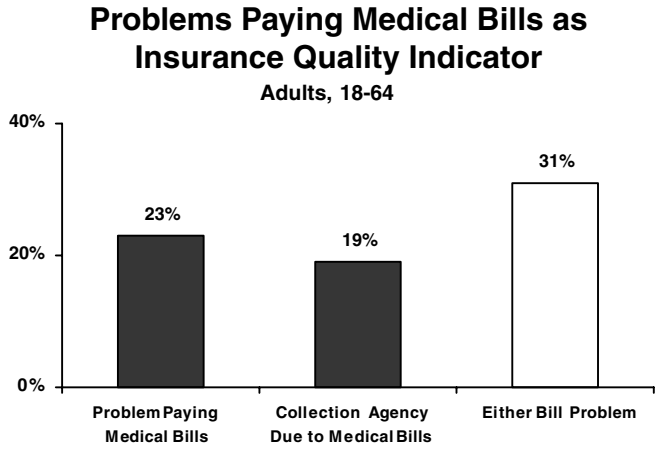
**Getting care when needed is "extremely," "very," or "somewhat" difficult.

Source: The Commonwealth Fund 1998 Survey of Women's Health.

didn't get the care you thought you needed?"—although the same respondent may list health problems and no contact with health care providers, or other indicators of access barriers.

Figure 6 helps illustrate the need to probe for experiences. In this 1999 survey of working-age adults, 12% of men and women said they had a time they did not get care when it was needed. On each of several other questions the percentage going without the specified service varied, ranging as high as 16% for not following up on a test or treatment due to costs. Altogether more than one of four—26% of the sample—had a time when they had gone without some type of health care in the year—double the rate on the single question. If we add dental care, the rate jumps up noticeably.

Figure 8



Source: The Commonwealth Fund 1999 Workers' Health Insurance Survey.

The National Health Interview Survey, MEPS, and other federal databases have begun to include a broader array of access probes. Although they are not designed to produce a scale or composite measure, looking at the cumulative results helps to capture the diversity of population experience.

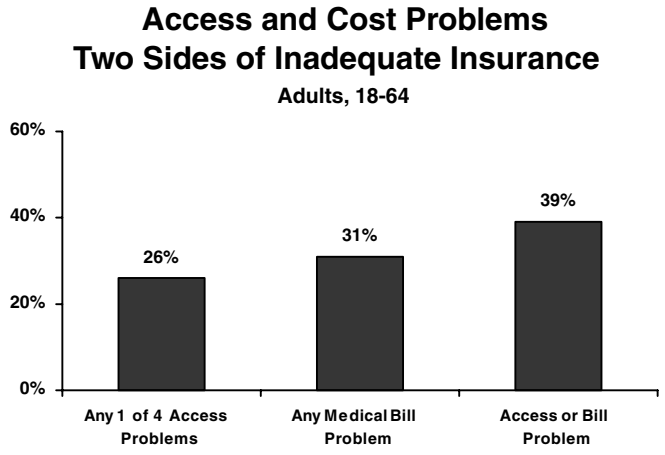
The access probes also indicate a frequent concern about follow-up care. Although uninsured or underinsured patients may succeed in getting to a clinic for a physician visit, gaps in coverage may undermine a patient's ability to get to the next stage of treatment or to follow up with appropriate treatment. The uninsured and underinsured are particularly at risk for prescription drugs and diagnostic tests and follow-up specialty care.

With not all survey respondents likely to have needed care in the referent time period, asking about difficulty in getting care when needed further explores access concerns. As recommended and tested by Andy Bindman, asking "How difficult is it to get care when you need it?" can pick up groups at risk for not seeking or receiving care when needed, although they had no recent access problem (Figure 7).

Difficulties paying for medical care provide another dimension of insurance concerns that resonate with the public. Since financial protection is the goal of insurance, findings that indicate struggles to pay when sick point to policy concerns.

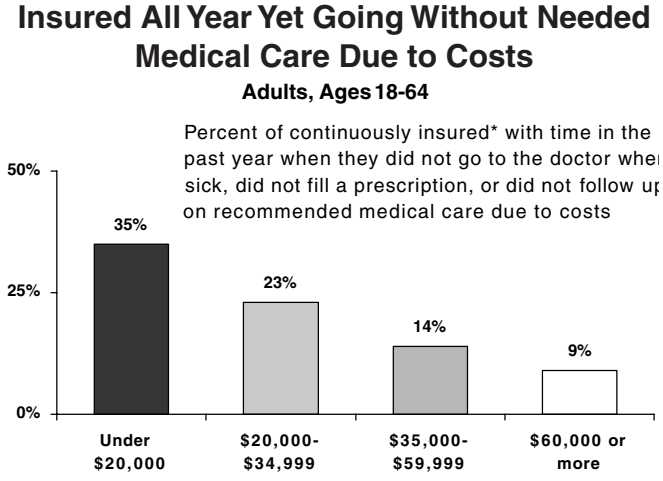
Most recently, Kaiser Family Foundation and Commonwealth Fund jointly sponsored surveys, as well as Fund-sponsored surveys, have included a general question about problems with paying medical bills or a more specific question that looked at severity by asking about dealings with collection agencies as a result of medical bills (Figure 8). In a 1999 Fund survey, nearly one-third of adults aged 18–64 report at least one of the two problems, and the problems extend well into the middle class (Figure 9). In comparison, in recent international surveys we find that the United States is unique

Figure 9



Source: The Commonwealth Fund 1999 Workers' Health Insurance Survey.

Figure 10



* No time uninsured in past year.

Source: The Commonwealth Fund 1999 Workers' Health Insurance Survey.

in the extent of financial insecurity. Including the cost dimension helps to highlight issues of "underinsurance" and gaps in coverage—uncovered benefits, as well as spells uninsured. Bill-paying problems also appear to resonate with the public and policymakers.

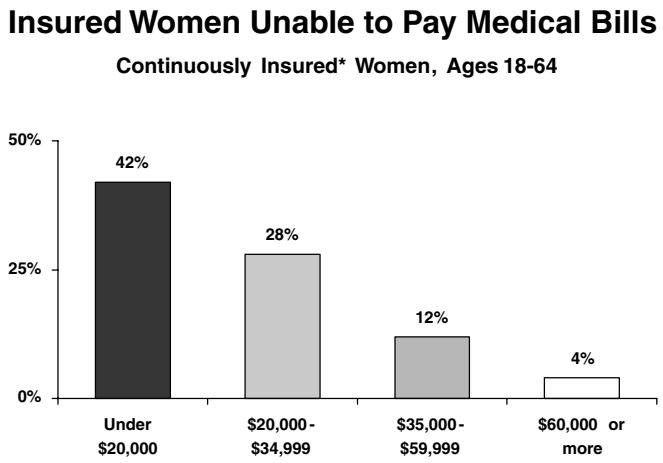
Inadequate Coverage

Questions about access and bill problems can also address policy concerns about those with inadequate insurance—the underinsured. Figures 10 and 11 provide examples from recent Fund surveys of women's health.

Restricting the analysis to women who have had no time uninsured during the year, we find that one-third had gone without needed care and 40% had been unable to pay medical bills in the past year.

Such questions also work well to distinguish among managed care plans. Patient access and cost experiences vary significantly in plans with complex in- and out-of-network arrangements or sharp restrictions on specialized services that expose patients to paying on their own.

Figure 11



* No time uninsured in past year.

Source: The Commonwealth Fund 1998 Survey of Women's Health.

Health and Socioeconomic Status

The strong relationship between socioeconomic status and health has long been noted within the United States. Internationally, various studies in industrialized countries find that the relationship persists and is remarkably similar in countries with universal health insurance coverage.

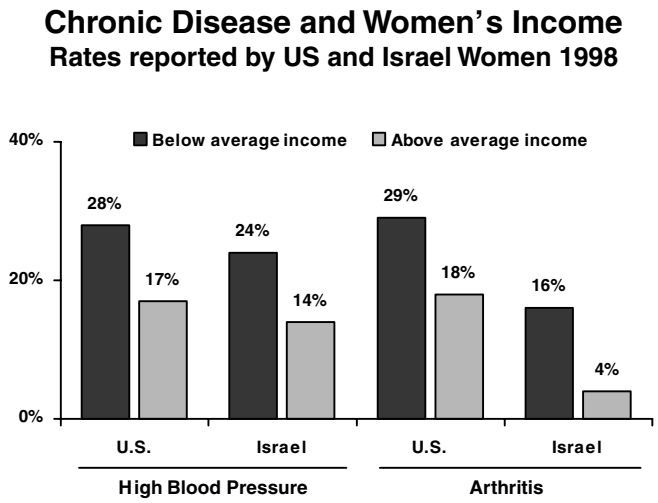
Surveys offer the potential of providing a standard metric across countries and of exploring underlying differences in access as well as health that persist even when financial barriers are removed. Recent Fund surveys of women's health in the United States and Israel, for example, raise a host of issues of common concern (Figures 12–15).

In both countries, less-educated and lower-income women are notably more likely to report physician diagnosis of chronic disease, more likely to report access problems when seeking care, and less likely to receive counseling or preventive health services. Yet Israel has universal coverage with a strong emphasis on primary care and "managed" care.

Violence and Abuse

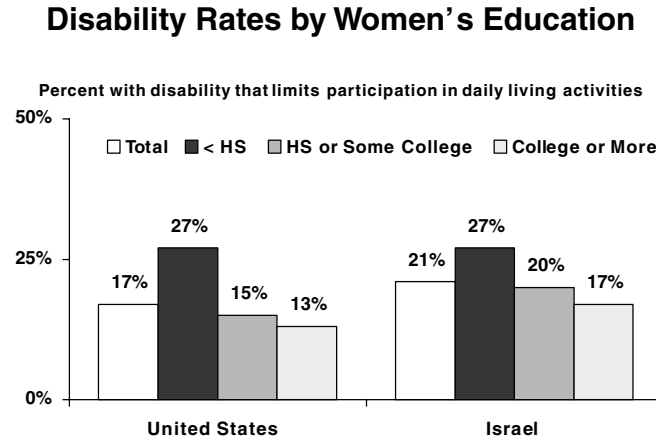
Emerging clinical research provides strong evidence of both short-term and long-term negative health effects of violence

Figure 12



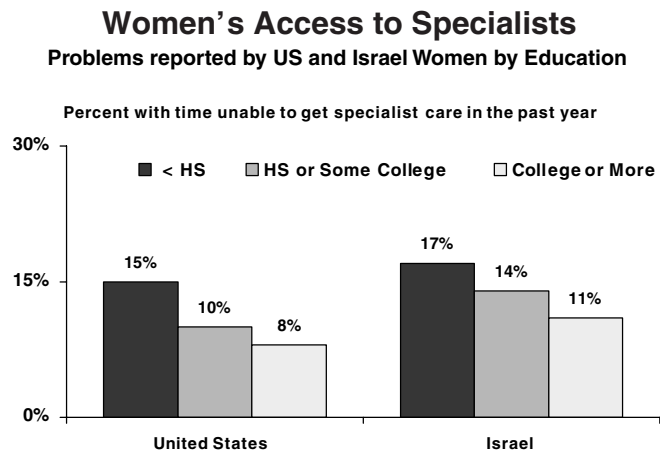
Source: The Commonwealth Fund 1998 U.S. and Israeli Women's Health Surveys.

Figure 13



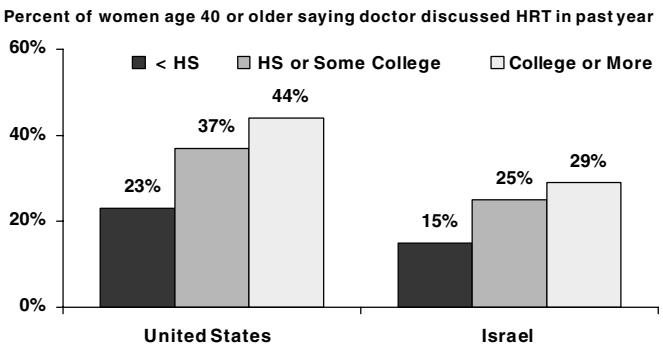
Source: The Commonwealth Fund 1998 U.S. and Israeli Women's Health Surveys.

Figure 14



Source: The Commonwealth Fund 1998 U.S. and Israeli Women's Health Surveys.

Figure 15



Source: The Commonwealth Fund 1998 U.S. and Israeli Women's Health Surveys.

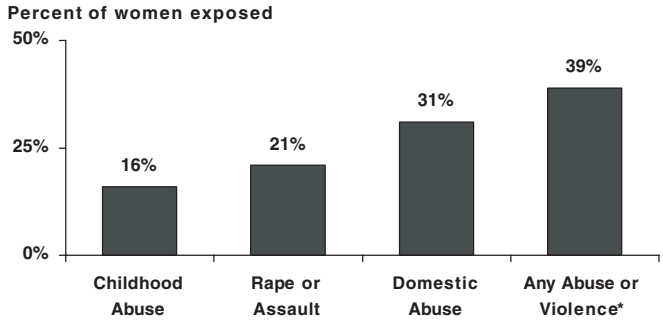
and abuse. Yet the topic appears but rarely in survey research. As a result, we have few national or regional estimates of the population at risk.

Recent Fund surveys on adolescent and women's health that included questions on violence and abuse found a media and policy audience eager for more. Although each survey included only a short section on the issue, both surveys were publicized as surveys of "abuse"— filling a void.

When designing surveys of adolescent and women's health, we encountered the dilemma of how to include violence or abuse along with questions about health and other issues. Typically, scales or topics include a lengthy question series—beyond the scope of interview time. Yet a more narrow focus on violence and abuse could have reduced participation in surveys and undermined the ability to analyze relationships with other experiences and health. To allow time to ask about health and mental health violence and behaviors, along with other issues, we thus had to develop new questions and select short versions of existing depression survey series. Both surveys had to depart from well-tested survey measures of health or violence.

The result was a rich source of new information on the interaction between violence, health, and behavior that stimulated policy discussions at national and regional levels and provided a national estimate of the population at risk (see Figures 16–19).

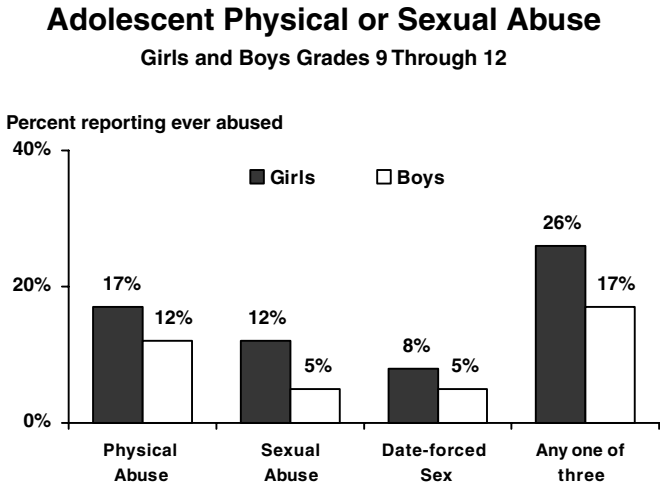
Figure 16



*Includes assault, battery, or rape by a spouse or partner, or physical/sexual assault or rape by anyone else, or physical or sexual abuse that occurred in childhood.

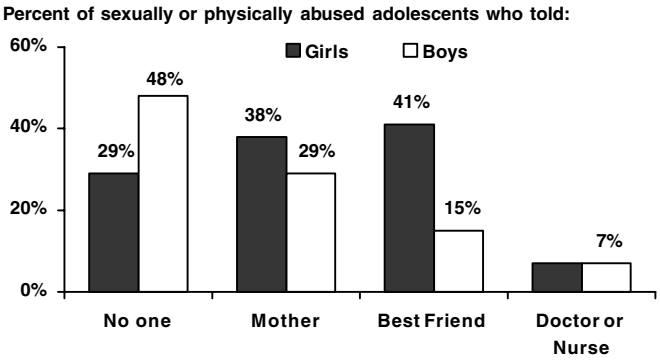
Source: The Commonwealth Fund 1998 Survey of Women's Health.

Figure 17



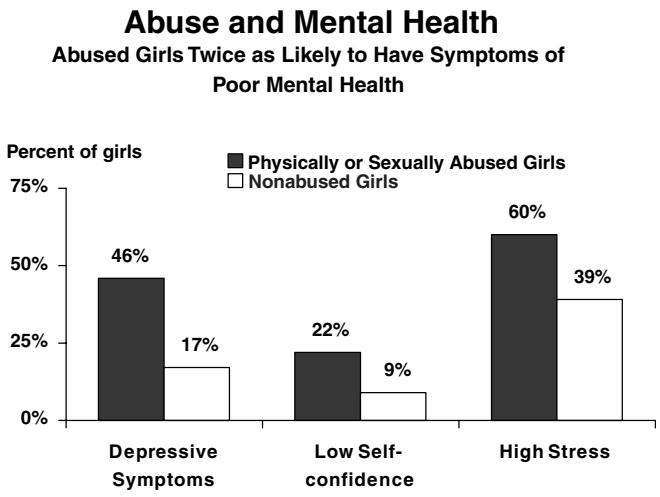
Source: The Commonwealth Fund Survey of the Health of Adolescent Girls, 1997.

Figure 18



Source: The Commonwealth Fund Survey of the Health of Adolescent Girls, 1997.

Figure 19



Source: The Commonwealth Fund Survey of the Health of Adolescent Girls, 1997.

Concluding Comments

In an ever more complex and dynamic health system, survey researchers face a wealth of enduring and emerging population health concerns including the uninsured, access, health and income, and violence and abuse. The four issues illustrate the potential and challenge of conducting policy-relevant survey research. Typically, tapping into the changing policy debates in a timely manner requires the use of new, untested questions or the creative use of question series. Making room for new questions while enabling analysis of interactive effects may require shortened versions of tested survey scales or series as well as new questions.

Working on the enduring issues—such as the uninsured—is often the most difficult. For the "old" issues, communication strategies may be key. Having a survey "story" from pilot tests or the single statistic that captures a larger experience in new ways can help stimulate as well as inform the public debate. The effort to address public concerns may also require the inclusion of questions that put a more "human" face on results.

Last, but not least, effective communication of results to a public and policy audience may require finding new ways to combine results or focus findings on particular issues in order to resonate with and speak to public concerns.

Further Remarks by Lu Ann Aday

I would like to present some of the methodological issues identified by designers of statewide health surveys, based on a session I chaired at this year's Association for Health Services Research Conference on State-Initiated Surveys of Family Health and Insurance Coverage. At that session, presenters from Ohio, Utah, Vermont, and Wisconsin reviewed their state data collection activities and the design and implementation issues they confronted in carrying out these efforts.

State surveys are being used to provide baseline data on the number and characteristics of the uninsured to guide state health care reform; to generate estimates of health and health care needs to inform statewide or local health planning and program development; to evaluate the impact of specific policy initiatives, such as state Medicaid managed care reform; and either to anticipate or measure the impact of the major state-federal program to expand coverage to uninsured children and families through the Child Health Insurance Program (CHIP).

My intent in highlighting the health survey design and implementation issues identified by state and local surveyors is to point to the decisions they confront, which the state of the art we are attempting to advance through our discussions today and throughout the conference might help to guide and inform.

State survey developers confront a cafeteria of design choices with respect to who is the focus of the survey; what issues are addressed; the universe that is targeted for sampling; and when, or how often, data are gathered.

Who? Some state surveys collect data from adult respondents about themselves, some for a sampled child, and others on all family or household members by an identified proxy respondent.

What? Most of the surveys ask questions on health status and disease prevalence, health risks, health care utilization, health insurance coverage, access to care, and respondent or household demographics. Some also address special topics identified by project staff or stakeholders.

Where? A number of state surveys, in addition to producing state-level estimates, have employed complex and costly sampling schemes to yield a sufficient number of cases to generate estimates for substate regions, districts, counties, cities, or special subgroups of interest (by race/ethnicity, for example).

When? In some instances, the survey represents a one-time effort—to date. In other cases, there are plans for or past experiences with conducting surveys at selected intervals to trace changes over time. The respective states identified pros and cons for each data collection strategy.

They also highlighted an array of issues involved in designing and conducting policy-relevant state health surveys, related to their meaning and applications to national and state policy debates, the best methods for minimizing survey errors and costs, and how to effectively manage and monitor such studies.

Meaning. The design and conduct of policy-relevant state surveys require an attunement to state health policy issues, an ability to craft and tailor the survey to address them, soliciting and garnering the support of key policy stakeholders for what is likely to be perceived as a costly research effort, and generating reports that are timely and interpretable to interested stakeholders and the public. No mean feat!

Methods. Most statewide surveys use computer-assisted telephone interview methods. Issues of phone noncoverage and nonresponse are of particular concern, especially as state data collectors attempt to compare their estimates with state-level estimates available from national surveys, such as the Current Population Survey.

Instrument development decisions compel considerations of whether to develop new questions or draw upon those from existing studies, as well as whether to designate core questions as well as supplementary modules asked only of selected respondents or on designated waves of data collection.

Many state surveys employ complex sampling designs requiring oversampling of selected areas or subgroups. These types of designs present special challenges in minimizing both systematic and variable survey errors associated with the noncoverage of special populations or persons without phones, low overall response rates, high nonresponse rates on selected questions, and variance estimation adjustments required by the complex nature of the sample design.

Management. Deciding upon relevant data collection subcontractors involves considerations of how best to (a) identify and evaluate potential bidders; (b) monitor the quality of data collectors' activities, and (c) deal with what often seem inevitable time delays and cost overruns.

Surveys that entail substantial oversampling are costly to conduct and require major external or departmental resources to carry out successfully. Such studies also either intermittently or on a sustained basis place extra demands on the staff and administration of the organizations charged with conducting them.

Given the resource demands of such studies, in addition to the commitment in many arenas to develop integrated data systems for monitoring and evaluating federal- or state-level health policy, survey developers must often forge new collaborative interagency arrangements for carrying out the survey, as well as sharing study results.

Obligations to make the data available for public use also raise knotty questions regarding the confidentiality of the data and the timeliness of release in relationship to optimal data cleaning, imputation, weighting, and data documentation priorities.

I would hope that in our own methodological research on health surveys, we attend to how evolving survey tools and technologies can directly serve the needs of this important cadre of state-level health survey researchers.

Discussion Notes, Panel Session

D. E. B. Potter and Richard Strouse

Comments focused on the rapidly changing health care system and the need for survey methods to accurately measure access to care. At the state level, where much of the change in health policy is taking place, particular concern was expressed regarding the difficulty in obtaining accurate information about insurance coverage. State-level surveys sometimes find it difficult to select national surveys as models in developing standardized measures for insurance coverage, as various national surveys differ in time frames (current versus last 12 months), questions, and other design features, depending on their objectives. State surveys also often lack the resources to replicate more costly national designs. It also is becoming more difficult to accurately discriminate between private and public insurance coverage, as states shift beneficiaries from Medicaid to private managed care plans and enroll the uninsured in subsidized plans. For example, a recent Washington state survey observed that respondents experienced considerable difficulty identifying their plan names, as former Medicaid beneficiaries had recently been assigned to various HMO plans.

A suggested response was to develop a vehicle to standardize elements of questionnaire and sample design and mode of administration. The Consumer Assessment of Health Plans Survey (CAHPS®), which may be used for both mail and telephone modes of data collection and for various population subgroups, was cited as a useful vehicle for standardizing wording for questions on plan characteristics and satisfaction.

However, standardization per se does not necessarily meet the needs of policymakers responding to emerging issues. Researchers also need to be creative in developing and validating new measures and in providing timely data to inform policy. A potential approach, which was illustrated in Schoen's presentation and reiterated during the discussion, is for researchers to use large national health surveys as benchmarks to understand and to further develop new measures. In addition, it is essential that national surveys continue efforts to understand and explain differences among their respective estimates of insurance coverage and other measures of access to health care.

The discussion closed with a recommendation to convene a collaborative consortium of foundations and government statistical agencies to systematically assess surveys and survey methods used to track changes in access to care and to develop more standardized measures and designs. A similar need to share insights and future priorities was expressed in a recent conference convened by the Robert Wood Johnson Foundation (see A. Bindman and M. Gold, *Measuring Access to Care through Population-Based Surveys in a Managed Care Environment*: *Articles from an Invitational Symposium*, Washington, D.C., March 26–27, 1997, and published in *Health Services Research, 33* (3), August 1998, Part II). Sharing information on methods and developing more standardized measures suggests a useful future direction to assist policymakers and researchers.

The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services and the Agency for Health Care Policy and Research is intended or should be inferred.

Racial and Ethnic Populations: Cross-Cultural Considerations

Surveys of special populations have become increasingly more common and important as health planners and policymakers require more and better data to address the health care needs of specific populations and population subgroups. Thus, while these issues have drawn both substantive and methodological interest among the health research community for some time, the urgency to address these issues—and to do so in a manner that fully recognizes their complexity, diversity, and uniqueness—has increased dramatically in recent years. And nowhere is this more critical and pressing than in health research with racial and ethnic populations.

Our health statistics clearly show that, relative to the white majority population, racial and ethnic minority groups in the United States generally have less access to care, lower levels of health care utilization, and poorer health status. These and other disparities in health between minorities and whites are of sufficient magnitude and concern that reducing these discrepancies has become a major target of the federal health policy agenda over the next decade, a key strategic goal that permeates virtually every component of the recently published *Healthy People 2010*. To address these objectives, it is obviously essential that our health survey methods be of sufficient sensitivity, flexibility, diversity, and rigor to provide the appropriate, critical data required to better understand why these discrepancies exist and to accurately measure and monitor our progress toward meeting these goals.

In combination, the five papers featured in this session illustrate quite well some of the key challenges and complexities associated with gathering accurate, meaningful, and appropriate data from multicultural populations, especially those based on race and ethnicity. The first four papers, in particular, focus on the need for and challenges associated with developing sound, culturally appropriate survey measures and instruments, as well as the potentially deleterious effects of *not* doing so. The final paper is significantly more far-reaching, providing a provocative description and set of examples that illustrate how important and pervasive the impact of racial, ethnic, and cultural factors can be on virtually every component of the survey research process in collecting data from racial and ethnic minorities.

Culture and Item Nonresponse in Health Surveys

Linda Owens, Timothy P. Johnson, and Diane O'Rourke

Introduction

Patterns of variability in responses to health survey questionnaires across racial and ethnic groups have been documented within many nations (Polednak, 1989). What is often unclear is the degree to which these differences are a consequence of cross-cultural variability in the concepts being investigated or of culture-based methodological artifacts. Cultural differences in self-reports of health conditions and behaviors, for example, may be at least in part a consequence of group variations in question interpretation (Johnson et al., 1996) and/or response styles such as the social desirability trait (Ross & Mirowsky, 1984). One indicator available to all analysts of health surveys that may be useful in identifying cultural variations in these processes is item nonresponse. Typical sources of item nonresponse, commonly referred to as "missing data," include respondent failure to answer questions, interviewer failure to ask questions, and researcher failure to design appropriate instruments and surveys. Item nonresponse may also be a consequence of respondent-interviewer miscommunication. These errors may be manifested as either inability (i.e., "don't know") or unwillingness (i.e., refusal) to answer specific survey questions. In this paper, we present a systematic analysis of patterns of item nonresponse across several cultural groups in the United States using four national health survey data sets collected during the past decade. Before doing so, we briefly review previous studies that have investigated this topic.

About half of the available research concerned with item nonresponse in health-related surveys has reported assessments across cultural groups. Ten of 21 studies identified failed to do so (Bradburn, Sudman, Blair, & Stocking, 1978; Brock, Lemke, & Woolson, 1986; Catania, McDermott, & Pollack, 1986; Colsher & Wallace, 1989; Dengler, Roberts, & Rushton, 1997; Garrard, Skay, Tratner, Kane, & Chan, 1989; Guadagnoli & Cleary, 1992; Ingles, 1987; Kimberlin, Pendergast, Berardo, & McKenzie, 1998; Sherbourne & Meredith, 1992). Of the 11 studies that have examined racial and ethnic variations in item nonresponse, 7 reported finding differences (Aday, Chiu, & Anderson, 1980; Kupek, 1998; Peterson & Catania, 1997; Sabogal, Binson, & Catania, 1997; Smith, 1992; Witt, Pantula, Folsom, & Cox, 1992; Ying, 1989) and four others did not (Aquilino, 1992; Johnson & DeLamater, 1976; Michael, Laumann, Gagnon, & Smith, 1988; Stueve & O'Donnell, 1997). We note that only 3 of these 11 studies applied multivariate methods that were able to control for other factors also known to be associated with item nonresponse (Aquilino, 1992; Johnson & DeLamater, 1976; Kupek, 1998).

This small body of research suggests that minority group respondents and members of less acculturated immigrant groups may have greater difficulties comprehending survey items that in most cases are developed by middle-class representatives of a nation's dominant cultural group. In addition, they may be less willing to reveal sensitive information during survey interviews. Based upon this research, we hypothesize more broadly that minority cultural groups in general will have higher nonresponse to individual survey items than non-Hispanic white respondents.

Methods

The analysis focused on four large health-related surveys: the 1992 Behavioral Risk Factor Surveillance System (BRFSS), the 1992 National Household Survey on Drug Abuse (NHSDA), the 1991 National Health Interview Drug Use Supplement (NHIS), and the 1990–91 National Comorbidity Survey (NCS). We selected these four data sets because they contained questions reflecting several health dimensions and because they represent a variety of data collection methods. The BRFSS contains information on health behavior, the NHSDA and NHIS focus primarily on drug use, and the NCS contains questions concerned with psychological health. The basic characteristics of each survey are presented in Table 1.

In each data file, we chose several items that we felt reflected different components of health. For the sake of simplicity, we chose items that were asked of everyone and avoided items that were based on skip patterns.

For two of the four data sets, we were able to develop three summary missing-data indicators for each set of health questions of interest: (1) respondent provided a "don't know" answer to one or more questions, (2) respondent provided a "refusal" answer to one or more questions, and (3) respondent provided either a "don't know" or "refusal" answer to one or more questions. In the NHIS drug supplement, all missing data, whether "don't know," "refusal," or blank, were simply coded as "missing." In the NHSDA, refusals and blanks were grouped together while "don't know" responses were not analyzed. Using these procedures, a total of 10 sets of health survey items

The authors are at the Survey Research Laboratory, University of Illinois at Chicago.

Table 1. Summary of data sources

Sample Characteristics	BRFSS	NHIS Drug Use Supplement	NHSDA	NCS
Year of data collection	1992	1991	1992	1990–91
Total sample size	96,213	21,174	28,832	8,098
Sample analyzed	96,213	12,825	21,578	7,411
Geographic coverage	49 states (Arkansas excluded)	50 states plus D.C.	50 states plus D.C.	48 states plus D.C.
Respondent ages	18+	18–44	12+	15–55
Ages analyzed	18+	18–44	18+	18–55
Data collection method	Telephone	Self-administered	Self-administered	Face-to-face
Number of items analyzed	16	23	29	16
Missing data	DK + NA, refused	Blank + DK + refused	Blank + refused	DK, NA

were developed and examined. Each of the 10 items is a dichotomous variable measuring whether or not any of the component questions contains missing data.

Because income is typically considered the most problematic when it comes to item nonresponse, we also report analyses of this variable for each data file as a benchmark for comparisons with the health measures of interest.

Each measure was initially examined using simple cross-tabulations, followed by logistic regression models in which we controlled for several sociodemographic variables associated with item nonresponse, including age (Ferber, 1966), gender (Aquilino, 1992), education (Kupek, 1998), and marital status (Witt et al., 1992).

In the BRFSS and NHSDA, age is a categorical variable with the three categories being 18–34, 35–54, and 55 or older. In the NHIS and NCS, there were no respondents 55 or older, so age is a continuous variable.

In all four data files, education is continuous while marital status, race, and sex are categorical. The four categories of marital status are married, widowed, separated or divorced, and single. The race categories for the BRFSS, NHIS, and NHSDA are white, African American, Hispanic, and other. In the NCS, the race categories are white, African American, Hispanic, and Native American. In the logistic regressions, the reference categories for the independent variables are married, white, and male. The only exception is the NCS, where the reference category for gender is female.

The items analyzed in the BRFSS include seven health behavior items and eight AIDS knowledge/attitude questions. For the AIDS questions, we analyzed whether or not respondents refused to answer the questions. For the behavior items, we analyzed whether or not the respondents refused, said "don't know," or either. We did not analyze the "don't knows" for the AIDS questions because we believe that "don't know" represents a valid response to questions about general knowledge or attitudes. Like the health behavior questions, we analyzed three income measures—"don't know," refused, and any missing.

The NHIS survey is unique in that the person being interviewed, or reference person, also serves as a proxy respondent for other members of the household. The records in the data file refer to the reference person, the reference person's spouse, the reference person's children, and so on. To eliminate any confusion arising from answering questions for other household members, we limited the analysis to those records pertaining to the reference person; records referring to other household members were deleted. The resulting sample size is 12,825. For all the NHIS questions, missing data are indicated by a value of 9, meaning "unknown." Therefore, it is not possible to distinguish between refusals, blanks, and "don't knows."

The NHIS analysis includes three sets of items. The first (EVERUSE) contains eight questions that ask if the respondent ever used particular drugs. The second set (PAST12MO) asks if the respondent used those substances in the last 12 months. The third set (MJALCOKE) includes seven items asking about the use of alcohol, marijuana, and cocaine. In addition, we created a summary measure—ANYMSG—that indicates missing data on any of the other three items.

The NHSDA questionnaire asks respondents about their use of several different substances, including alcohol, sedatives, tranquilizers, stimulants, analgesics, marijuana, inhalants, cocaine, hallucinogens, and heroin. For the recreational drugs—alcohol, marijuana, inhalants, cocaine, hallucinogens, and heroin—respondents were asked all questions in the relevant section, even if they stated that they had never used the substance. The available response categories for each question always include "never used [drug in question]."

Because all the questions about recreational drugs were asked of everyone, we were able to analyze four different constructs—age at first use (AGE_REC), frequency of use (FREQ_REC), quantity of use (QNTY_REC), and recency of use (REC_REC).

For the prescription drugs, the respondents were given a list of specific drugs and asked which types they had ever used. If they indicated none, they skipped the remainder of the section. The skip patterns in the prescription drug questions limited us to only one index. For stimulants, tranquilizers, sedatives, and analgesics, there is a summary measure indicating whether the respondent has used any of these types of drugs. We combined the information from these four variables into a single dichotomous variables called NEVER_RX, which indicates whether there is missing data on any of the items. For all items, the missing category includes refusals and blanks, because there are too few refusals to analyze separately. "Don't know" responses are not included.

The National Comorbidity Survey was conducted via face-to-face interviews in 1990–91. We limited our analyses to adults aged 18–55 (n = 7,411). Sixteen items concerned with various aspects of the respondent's social relationships were examined, along with the survey item concerned with family income. We developed three missing data indicators for each: "don't know," not ascertained, and total missing data. In analyzing these data, we were also able to examine Native Americans as a separate racial and ethnic group.

Results

The results of the logistic regressions are presented in Tables 2 through 5. In the NHSDA, two of the regressions had no significant results. None of the independent variables had an effect on AGE_REC or REC_REC.

Regarding the sets of health items examined, we found race/ethnic differences in 13 of 18 regression models. In each case of significant differences, higher item nonresponse rates were found among one or more of the minority groups examined when contrasted with non-Hispanic white respondents. African American respondents most commonly had higher item nonresponse rates to health questions (in 7 of 18 equations); Hispanics and respondents from other racial and ethnic groups had higher rates in 2 and 4 of the 18 equations, respectively.

Cultural group differences in nonresponse to income questions were also identified in 8 of the 10 income models examined. Comparisons between white and minority group respondents, however, were not consistent in their direction. Seven contrasts found African American, Hispanic, and other racial and ethnic groups had higher nonresponse rates to income questions than whites, while five other contrasts found whites had higher nonresponse.

As with previous research, education was consistently associated with item nonresponse. In general, health question item nonresponse was greater among less educated respondents. More educated respondents were more likely to refuse to answer income questions and less likely to answer "don't know" to them. Male respondents generally showed higher item nonresponse rates to health questions and lower nonresponse rates to income questions. The association between age and item nonresponse was more complex. At least one nonlinear effect was identified (Table 2) and only one trend was identified across the models examined: Older respondents were more likely to refuse to answer income questions. Marital status also presented no clear pattern of findings. Among the health survey items examined, though, currently unmarried groups—those divorced or separated in particular—frequently had higher item nonresponse rates than did married respondents.

Discussion

Although item nonresponse rates to sets of health survey questions appear in general to be low, this research suggests that item nonresponse may vary systematically across cultural groups. Consistent with the small number of other studies available, we found higher item nonresponse rates among each of the minority racial and ethnic groups examined compared to non-Hispanic white respondents in all four data sets. Separate examination of trends in refusals and "don't know" answers further suggests that both information processing and social desirability considerations may contribute to these

Table 2. Results of BRFSS regression analyses (odds ratios)

Independent Variables	Refused AIDS	Don't Know Behavior	Refused Behavior	Missing Behavior	Any Refused (AIDS + Behavior)	Don't Know Income	Refused Income	Missing Income
Race								
African American	1.1833	.9094	.9306	.9110	1.1640	1.1731**	.8146***	.9982
Hispanic	1.1526	.9534	.8613	.9496	1.1267	1.4707***	.4918***	.9843
Other	1.6602***	1.3822***	.7530	1.3609***	1.5516***	1.8491***	.6663***	1.2156***
Age								
35–54	1.6368***	.6641***	1.6892*	.6868***	1.6398***	.8882**	1.7237***	1.2533***
55+	5.6036***	.6565***	2.3811***	.6945***	5.0999***	1.7947***	3.2929***	2.5626***
Marital status								
Widowed	1.6237***	1.6211***	1.1967	1.5881***	1.6191***	1.0150	.8140***	.8991**
Sep/Div	.9841	1.1647**	1.3645	1.1728**	1.0269	.6759***	.6835***	.6624***
Single	1.2029	1.4734***	1.1049	1.4606***	1.1879	1.8824***	.7359***	1.2721***
Education	.8392***	.7730***	.9153	.7801***	.8466***	.7371***	1.0593***	.8814***
Female	.8895*	.8798***	.4550***	.8584***	.8389***	1.6407***	1.2159***	1.4216***
Model N	95,249	95,249	95,249	95,249	95,249	95,249	95,249	95,249
Model χ^2	1199.8***	568.2***	41.8***	545.1***	1166.3***	2036.3***	1444.89***	2291.3***
R^2	.013	.006	.000	.000	.012	.021	.015	.024

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 3. Results of NHIS regression analyses (odds ratios)

Independent Variables	EVERUSE	PAST12MO	MJALCOKE	ANYMSG	MSGINC
Race					
African American	1.4197**	1.7858***	1.5007***	1.5737***	1.5048***
Hispanic	1.0916	1.1603	1.1453	1.0599	.9086
Other	1.1379	1.1099	1.2528	1.1265	.9349
Age	1.0024	1.0144*	1.0037	1.0069	1.0045
Marital status					
Widowed	1.6372	1.0693	.9495	1.0169	.9044
Sep./Div.	1.2134	1.3379**	1.0400	1.0619	.9237
Single	1.1429	1.1592	1.0102	1.0641	.8346*
Education	.9384***	.9115***	.9237***	.9171***	.9348***
Female	.7488**	.7306***	.9125	.9060	1.1256
Model *N*	12,774	12,774	12,774	12,774	12,71374
Model χ^2	35.611***	98.985***	96.959***	165.039***	85.661***
R^2	.003	.008	.008	.013	.007

* 1 ≤ .05, ** $p \leq .01$, *** $p \leq .001$

differentials. Given that survey questionnaires are most commonly constructed by white middle-class researchers, it should not be surprising that non-white respondents might be somewhat more likely to experience difficulties interpreting survey questions, even after careful pretesting. Likewise, members of minority groups may understandably be more reluctant to report about sensitive, particularly illegal, health-related behaviors.

In addition to these general interpretations of majority versus minority group differences in item nonresponse, some group-specific cultural differences should be noted. First, the largest odds ratio associated with respondent culture reflected Hispanic refusals to answer one or more questions concerned with their social relationships (OR = 3.15, Table 5). Given the central importance of family ties documented among Hispanic populations (Locke, 1998), we speculate that higher item nonresponse to these questions may be a consequence of a culturally driven unwillingness to report anything other than positive and harmonious interactions with family and friends. Other research has documented a similar pattern in which survey respondents preferred not to respond to evaluative survey questions rather than to report negative information (Johanson, Gips, & Rich, 1993). Second, greater reluctance to report substance use information (Tables 3 & 4) by African American respondents can also be appreciated given that group's long history of discrimination and persecution in the

Table 4. Results of NHSDA regression analyses

Independent Variables	NEVER_RX	FREQ_REC	QNTY_REC	ANY_REC	REF_FINC	BLK_FINC	MSG_FINC
Race							
African American	1.727*	1.3555	1.8502*	1.3796*	1.8982***	1.2464**	1.5715***
Hispanic	1.1425	.9123	.8530	.7459	.6128***	1.0249	.8200**
Other	.5115	1.3187	1.4363	.8288	1.4589*	.7720	1.1219
Age							
35–54	1.4621	.6486	1.0530	.9372	1.3971***	.6681***	1.0264
55+	1.2892	.4604	.3227	.3841*	1.2863	.4109***	.8223
Marital status							
Widowed	1.3570	1.6935	2.9721	1.9552	.3541***	1.8931**	.6487**
Sep./Div.	.7579	2.0598*	.8709	1.5076*	.2364***	4.3510***	1.1812*
Single	1.1766	1.5564	1.4386	1.3510	.4571***	2.2193***	.9171
Education	.9879	.9999	.9590	.9541	1.0433***	.9706*	1.0091
Female	.8031	.8802	.8892	.7366*	1.3722***	1.5548***	1.4752***
Model *N*	21,578	21,578	21,578	21,578	21,578	21,578	21,578
Model χ^2	15.052	13.454	18.536*	32.046***	456.895***	431.099***	235.158***
R^2	.001	.001	.001	.001	.021	.020	.011

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table 5. Results of NCS regression analyses (odds ratios)

Independent Variables	Don't Know Social Relation	Social Relations NA	Missing Social Relations	Don't Know Income	Income NA	Missing Income
Race						
African American	.5730	1.1284	.9616	.5314*	1.0737	.7866
Hispanic	1.2064	3.1513**	2.4562**	.2546*	1.3326	.7356
Native American	1.4770	1.0561	1.0790	1.7706	.5219	1.0929
Age	1.0442*	1.0184	1.0248*	.9510***	1.0234**	.9940
Marital status						
Widowed	1.6662	1.4457	.8144	3.8900	0.5150	1.2341
Sep./Div.	.4998	1.4184	1.0677	1.2113	.8547	.9692
Single	.6300	1.0603	.9083	5.0856***	1.6385**	2.7092***
Education	.8433**	.9856	.9238*	.8013***	1.0740*	.9481*
Female	2.3192*	1.3373	1.5591*	.5329***	.9155	.7325**
Model N	7,401	7,401	7,401	7,390	7,390	7,390
Model χ^2	25.09**	10.942	20.867**	186.451***	19.643*	86.662***
R^2	.061	.015	.022	.149	.013	.040

$^*p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$

United States and the not-unrealistic belief that drug use laws have been selectively enforced against African Americans.

The lack of patterns of effect across the four data files may be due, in part, to the fact that missing data were categorized differently. For example, in the BRFSS data, "don't know" and "not ascertained" were grouped together, while refusals were a separate category. In the NHSDA data, the blanks and refusals were grouped together, and in the NHIS all missing data were grouped into one category. Perhaps if all four data files had separate categories for "don't know," "refused," and "blank,"we may have seen more similarities in the results.

One must question the substantive significance of our findings, given the generally low prevalence of item nonresponse in these data. In general, it is probably correct to conclude that the differential rates of nonresponse are of insufficient magnitude to seriously bias survey findings. Yet our data also suggest that cultural differences in item nonresponse may become much more problematic under certain conditions. For example, nearly one-quarter (23.4%) of African American respondents to the 1991 NHIS supplement left unanswered at least one of the self-administered questions concerned with drug use (data not shown). The overall nonresponse rate to this block of questions was also very high: 16.8%. Although not conclusive, these data suggest that minority group respondents may be more likely to leave sensitive questions unanswered when given the opportunity to do so as part of a self-administered questionnaire. As the collection of sensitive survey information continues to shift toward self-administered modes, the effects of minority group status on item nonresponse rates should continue to be monitored.

References

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

Aquilino, W. S. (1992). Telephone versus face-to-face interviewing for household drug use surveys. *International Journal of the Addictions, 27,* 71–91.

Bradburn, N. M., Sudman, S., Blair, E., & Stocking, C. (1978.) Question threat and response bias. *Public Opinion Quarterly, 42,* 221–234.

Brock, D. B., Lemke, J. H., & Woolson, R. F. (1986). Identification of nonrandom item response in an epidemiologic survey of the elderly. In *Proceedings of the Section on Survey Research Methods* (pp. 430–434). Alexandria, VA: American Statistical Association.

Catania, J. A., McDermott, L. J., & Pollack, L. M. (1986). Questionnaire response bias and face-to-face interview sample bias in sexuality research. *Journal of Sex Research, 22,* 52–72.

Colsher, P. L., & Wallace, R. B. (1989). Data quality and age: Health and psychobehavioral correlated of item nonresponse and inconsistent responses. *Journal of Gerontology: Psychological Sciences, 44,* P45–P52.

Dengler, R., Roberts, H., & Rushton, L. (1997). Lifestyle surveys—the complete answer? *Journal of Epidemiology and Community Health, 51,* 46–51.

Ferber, R. (1966). Item nonresponse in a consumer survey. *Public Opinion Quarterly, 30,* 399–415.

Garrard, J., Skay, C., Tratner, E. R., Kane, R. L., & Chan, H. C. W. (1989). Nonresponse to survey questions by elderly in nursing homes. In F. J. Fowler (Ed.), *Conference proceedings: Health survey research methods* (pp. 129–137). DHHS Publication No. (PHS) 89-3447. Washington, DC: National Center for Health Services Research and Health Care Technology Assessment.

Guadagnoli, E., & Cleary, P. D. (1992). Age-related item nonresponse in surveys of recently discharged patients. *Journal of Gerontology: Psychological Sciences, 47,* P206–P212.

Ingles, S. (1987). *Evaluation of item nonresponse in the National Medical Care Utilization and Expenditure Study.* Washington, DC: National Center for Health Statistics.

Johanson G. A., Gips, C. J., & Rich, C. E. (1993). "If you can't say something nice": A variation on the social desirability response set. *Evaluation Review, 17,* 116–122.

Johnson, T. P., O'Rourke, D., Chavez, N., Sudman, S., Warnecke, R. B., Lacey, L., & Horm, J. (1996). Cultural variations in the interpretation of health survey questions. In R. B. Warnecke (Ed.), *Health Survey Research Methods Conference Proceedings* (pp. 57–62). DHHS Publication No. (PHS) 96-1013. Hyattsville, MD: National Center for Health Statistics.

Johnson, W. T., & DeLamater, J. D. (1976). Response effects in sex surveys. *Public Opinion Quarterly, 40,* 165–181.

Kimberlin, C. K., Pendergast, J. F., Berardo, D. H., & McKenzie, L. C. (1998). Issues related to using a short-form of the Center for Epidemiological Studies-Depression Scale. *Psychological Reports, 83,* 411–421.

Kupek, E. (1998). Determinants of item nonresponse in a large national sex survey. *Archives of Sexual Behavior, 27,* 581–594.

Locke, D. C. (1998). *Increasing multicultural understanding: A comprehensive model.* Thousand Oaks, CA: Sage.

Michael, R. T., Laumann, E., Gagnon, J. H., & Smith, T. W. (1988). Number of sex partners and potential risk of sexual exposure to human immunodeficiency virus. *MMWR, 37,* 565–568.

Peterson, J., & Catania, J. A. (1997). Item nonresponse in the National AIDS Behavioral Surveys among African American and White Respondents. In J. Bancroft (Ed.), *Researching sexual behavior: Methodological issues* (pp. 106–109). Bloomington: Indiana University Press.

Polednak, A. P. (1989). *Racial and ethnic differences in disease.* New York: Oxford University Press.

Ross, C. E., & Mirowsky, J. (1984). Socially desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior, 25,* 189–197.

Sabogal, F., Binson, D., & Catania, J. A. (1997). Researching sexual behavior: Methodological issues for Hispanics. In J. Bancroft (Ed.), *Researching sexual behavior: Methodological issues* (pp. 114–133). Bloomington: Indiana University Press.

Sherbourne, C. D., & Meredith, L. S. (1992). Quality of self-report data: A comparison of older and younger chronically ill patients. *Journal of Gerontology: Social Sciences, 47,* S204–S211.

Smith, T. W. (1992). A methodological analysis of the sexual behavior questions on the General Social Surveys. *Journal of Official Statistics, 8,* 309–325.

Stueve, A., & O'Donnell, K. N. (1997). Item nonresponse to questions about sex, substance use, and school. In J. Bancroft (Ed.), *Researching sexual behavior: Methodological issues* (pp. 376–389). Bloomington: Indiana University Press.

Webster, C. (1996). Hispanic and Anglo interviewer and respondent ethnicity and gender: The impact on survey response quality. *Journal of Marketing Research, 33,* 62–72.

Witt, M. B., Pantula, J., Folsom, R. E., & Cox, B. G. (1992). Item nonresponse in 1988. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 85–208). Rockville, MD: National Institute on Drug Abuse.

Ying, Y. W. (1989). Nonresponse on the Center for Epidemiological Studies-Depression Scale in Chinese Americans. *International Journal of Social Psychiatry, 35,* 156–163.

Cross-Cultural Adaptation of Survey Instruments: The CAHPS® Experience

Robert Weech-Maldonado, Beverly O. Weidmer, Leo S. Morales, and Ron D. Hays

Background

Collecting accurate health data on the growing number of ethnic minorities in the United States has increased in policy relevance in recent years. Today, most general-population sample surveys require translation into at least one language (usually Spanish), and often other languages as well. However, cross-cultural research is threatened by the failure to produce culturally and linguistically appropriate survey instruments for minority populations. Guillemin, Bombardier, and Beaton consider that cross-cultural adaptation of instruments is a "prerequisite for the investigation of cross-cultural differences" (1993, p. 1425). A survey conducted with an inadequate instrument may lead to erroneous conclusions that are difficult to detect during analyses. Conclusions drawn from such research may be mistakenly attributed to differences between the source and target populations. These risks, and the increasing importance of cross-cultural research, have led to a reexamination of the prevalent techniques for developing survey instruments that will be used in different languages and for assessing the cultural appropriateness of survey instruments that are utilized for this type of research.

In this paper we define culturally appropriate translated survey instruments as conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population. This paper provides recommendations for the cross-cultural adaptation of survey instruments and illustrates with examples of what is being done in the Consumer Assessment of Health Plans Study (CAHPS®).

The CAHPS® Surveys

CAHPS® is a 5-year initiative that aims to produce a set of standardized survey instruments that can be used to collect reliable information from health plan enrollees about the care they have received. CAHPS® items include both evaluations (ratings) and reports of specific experiences with health plans. CAHPS® surveys are constructed from two pools of items: "core" items that apply across the spectrum of health plan enrollees and "supplemental" items that are used in conjunction with core items to address issues pertinent to specific populations, such as Medicaid fee-for-service and Medicare managed care. The results of these surveys are then used to prepare reports that provide information to consumers who are trying to select a health plan.

CAHPS® recognizes the need to translate its instruments into several languages in order for its users to adequately collect data on consumers. The CAHPS® survey instruments were translated into Spanish because it is the second most widely used language in the United States (Weidmer, Brown, & Garcia, 1999). As CAHPS® has expanded, several states and users have expressed the need to translate the CAHPS® instruments into other languages as well. The principal goal of the translation process of the CAHPS® surveys and protocols is to produce instruments that are culturally appropriate for the different groups in the selected languages. The main challenge is to produce such instruments while maintaining equivalency with the English-language version.

Cultural Adaptation of Survey Instruments

Guillemin et al. have described the process of cross-cultural adaptation of surveys as "oriented towards measuring a similar phenomenon in different cultures; it is essentially the production of an equivalent instrument adapted to another culture" (1993, p. 1425). We define culturally appropriate translated survey instruments as conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population.

In translating, it is important to distinguish between technical and conceptual equivalence. Technical equivalence refers to equivalence in grammar and syntax, while conceptual equivalence refers to the absence of differences in meaning and content between two versions of an instrument. A technically equivalent instrument is a literal translation using the "equivalent denotative meaning" of the words in the original survey. However, different terms may have a different connotative, or implied, meaning in different cultures, requiring an assessment of conceptual equivalence in the translation of instruments (Marin & Marin, 1991).

Conceptual equivalence includes item and scalar equivalence of the source and translated surveys. Item equivalence

Robert Weech-Maldonado is at Pennsylvania State University, University Park.

Beverly O. Weidmer is at RAND, Santa Monica, CA.

Leo S. Morales and Ron D. Hays are at the University of California at Los Angeles and RAND, Santa Monica, CA.

signifies that each item has the same meaning for subjects in the target culture. Scalar equivalence is achieved when the construct is measured on the same metric in two cultures (Hui & Triandis, 1985). Health surveys generally use categorical rating scales where response choices are ordered along a hypothesized response continuum (e.g., *excellent* to *poor*). It is important to determine if there is equivalence in the distances between the response choices in the two cultures (Keller et al., 1998).[1]

Cultural competence refers to the requirement that the translated instrument adequately reflect the cultural assumptions, norms, values, and expectations of the target population (Marin & Marin, 1991). Cross-cultural researchers differentiate between universal or common meaning across cultures ("etic") and group-specific ("emic") constructs or ideas. The source survey reflects the assumptions and values of the researcher's culture, and in translating surveys it is generally assumed that the constructs of the source survey are etic. Translated surveys should include both etic and emic items in order to reflect properly the reality being studied. This implies the development of new items that reflect the emic aspects of a concept in the target culture (Brislin, 1986).

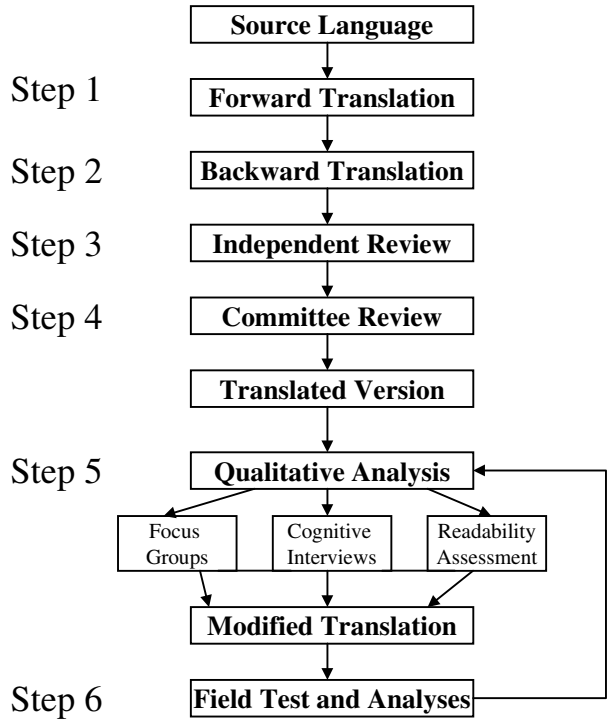
Linguistic appropriateness refers to the language readability and comprehension of the translated instrument. The goal is to develop instruments using wording at a level easily understood by the majority of potential respondents. An instrument developed in the source language at an eighth-grade reading level does not automatically maintain the same reading and comprehension level upon translation. The problem of equivalence in reading level is further compounded if the target population is at a lower average reading level than the source language population.

In order to cross-culturally adapt survey instruments, we propose a framework (Figure 1) that comprises the following activities:

- Translation (steps 1 to 4)
- Qualitative analysis (step 5)
- Field test and analyses (step 6)

Based on the results of the field test, additional qualitative analysis may be necessary. The International Quality of Life Assessment (IQOLA) project group has used a similar protocol in translating the SF-36 Health Survey into different languages (Bullinger et al., 1998; Gandek & Ware, 1998).

Figure 1. Cultural adaptation of survey instruments



Translation (Steps 1 to 4)

Most researchers today agree that it is no longer acceptable to use a direct-translation technique (or one-way translation) for translating survey instruments. A review of the literature indicates that the most accepted approach to translation is one in which a variety of techniques are used to ensure the reliability and validity of the translated survey instrument (Brislin, 1986; Bullinger et al., 1998; Marin & Marin, 1991). The rationale behind this approach is that no single technique adequately demonstrates and improves the equivalence of an instrument, and that only a multistrategy approach that provides and evaluates different types of equivalence can produce an adequate translation. We recommend a process for translating surveys that includes translation, back-translation, independent review, and review by committee.

1. Forward-translation

Professional translators (two or more) experienced in translating similar survey instruments, preferably native speakers of the target language, are retained to translate the survey instrument. The translators used for this task should have familiarity with the target population and with data collection procedures. Before starting the translation, the translators should be briefed on the objectives of the study, the demographic characteristics of the sample, the interviewing mode to be used, and the targeted reading level of the translation.

2. Back-translation

Once the instruments are translated they go through a process of back-translation. In this process the translated instrument is given to two translators, native English speakers, who are instructed to translate the questionnaire back into English. It is important that these translators not have access to the

[1]For a discussion of the Thurstone scaling exercise applied to the SF-36, see Keller et al. (1998).

original English language versions of the instrument and that they do not consult with the forward-translators.

3. Independent Review and Comparison

The third step in the translation process it to give the translated versions of the survey instruments to one or more bilingual reviewers. The reviewers are provided with the original English versions and the back-translated versions and are instructed to compare the two, highlighting any discrepancies in meaning or equivalence.

4. Review by Committee

Once the review process is completed, the forward-translators, the back-translators, and the reviewer(s) hold a series of meetings to discuss problems found during the review process, to correct errors in grammar and syntax, and to resolve problems of equivalence found among the versions. Decisions on wording and corrections are made by consensus. The rationale is that a translator or back-translator can introduce his or her own bias or error into a translation. The review-by-committee approach is useful in neutralizing the cultural, social, and ethnic bias that can be introduced when using only one translator and one back-translator.

CAHPS® Translation

Rather than produce multiple, population-specific Spanish translations, CAHPS® sought to produce an instrument that would be understood by most respondents by using "broadcast Spanish" and that maintained a reading and comprehension level accessible to most respondents. "Broadcast Spanish" refers to a type of Spanish that is understood by most Spanish speakers regardless of their country of origin or ethnic background (Marin & Marin, 1991).

A professional translator experienced in translating survey instruments similar to the CAHPS® instrument was retained. The translated instrument was then given to a bilingual reviewer experienced in designing and translating survey instruments for cross-cultural research. The reviewer focused on identifying syntax and typographic errors, identifying questions or terms that sounded awkward, and identifying terms that were conceptually problematic. Once this process was complete, the reviewer was provided with the English version and was asked to compare the two instruments, highlighting any discrepancies in meaning or equivalence.

In an effort to adhere as closely as possible to the English version, the translator produced an initial Spanish version of the survey instruments that was technically equivalent to the English version, but in many instances was not conceptually equivalent, and in some cases was not linguistically appropriate for the target population (by using terms that are seldom used in Spanish, anglicisms, or words that are too sophisticated for the target population). The translator had been instructed to aim for a translation that would be appropriate for a Spanish-speaking Medicaid population likely to have less than 6 years of formal education. However, this proved to be difficult to accomplish while maintaining equivalence to the English version.

A member of the RAND CAHPS® team met with the translator and the reviewer to go over discrepancies related to equivalence. The reviewer and the translator back-translated problem areas in the Spanish version to further distinguish the source of the problems before decisions were made about addressing them. A final review of the original English version, the translation, and the back-translation was conducted by the committee—the translator, the reviewer, and CAHPS® team member—and alternative wording for problematic terms was implemented. Table 1 shows terms that were problematic because they were not conceptually equivalent, were too sophisticated for the target population, or were too infrequently used by most Spanish speakers. The alternative wording in the final version comes closer to the conceptual meaning in the English version and is easier for the respondents to understand.

Table 1. Terms that presented difficulty in translation

Original English	Alternative Wording Used in the Final Spanish Version	Back-translation
health insurance plan	plan de seguro médico	medical insurance plan
health provider	profesional de salud	health professional
rating/rate	calificación/califica	grade/grade
usually	normalmente	normally
preventive health steps	medidas de salud preventiva	preventive health measures
listen carefully	escucharon atentamente	listen attentively
health care	atención médica	medical attention
prescription medicine	medicamentos recetados	prescribed medications
male or female	niño o niña/hombre o mujer	boy or girl/man or woman
background	ascendencia	ascendancy
grade	año	year
school	estudios	studies
highest	avanzado	advanced

Qualitative Analysis (Step 5)

Qualitative research consists of "research methods employed to find out what people do, know, think, and feel by observing, interviewing, and analyzing documents" (Shi, 1997, p. 398). These methods should be viewed as complementary to quantitative methods. Qualitative methods are particularly useful in assessing the cultural competence or content validity of the translated survey instrument.[2] It is important to evaluate whether the survey measures the group-specific domains of the phenomenon under study for the target population. Qualitative methods assist in identifying the "etic" (universal) and "emic" (culture-specific) constructs or behaviors of a group. This constitutes an evaluation of the "subjective" culture whereby patterns in responses by members of a group are used to identify the group's cognitive structure (Marin & Marin, 1991). The assumption is that the group's norms, values, and expectancies influence the observed consistencies or similarities in responses of a given cultural group. Qualitative methods can also be used to assess the conceptual equivalence and linguistic appropriateness of the translated survey.

We are using qualitative methods to investigate the appropriateness of the CAHPS® survey content for Spanish-speaking Latino patients enrolled in Medicaid. First, we want to determine whether the items and scales currently contained within CAHPS® address the key concerns and expectations of Latino patients with respect to their health care providers and health plans. Second, we want to verify that the translated survey items, initially developed in English, have similar meaning in Spanish. Finally, we want to determine the readability level of the Spanish language survey instruments and determine whether it is appropriate for the Spanish-speaking Medicaid population.

There are three types of qualitative research pertinent to cross-cultural research: focus groups, cognitive interviews, and readability assessments. In this section we discuss the use of focus groups and cognitive interviews. For a discussion on readability assessments and its application to the CAHPS® surveys, see Morales, Weidmer, & Hays (1999) in the conference proceedings.

Focus Groups

Focus groups are a research tool that relies on group discussions to collect data on a given topic (Morgan, 1996). Participant interactions help to reveal experiences, values, beliefs, and feelings. In addition, group discussion helps to uncover the extent of consensus or diversity, and its sources. Focus groups have been used extensively in marketing research to obtain customer input on new products (Burns & Bush, 1995); however, their use in cross-cultural research has been more limited. The primary objective of focus groups in cross-cultural research is to assess whether the domains currently covered in the survey adequately address the needs and expectations of the target population and to assess the need for developing new domains or expanding current domains. The focus group process usually starts with a literature review and analysis of health surveys that focus on the target population, to aid in the identification of issues and concepts particular to the cultural group.

Stewart & Shamdasani (1990) have identified eight steps in the design and conduct of focus groups:

- Formulation of the research question
- Identification of sampling frame
- Identification of moderator
- Generation and pre-testing of structured protocol
- Recruiting the sample
- Conducting the focus group
- Analysis and interpretation of data
- Writing the report

A group size of 8 to 12 respondents per focus group is recommended (Burns & Bush, 1995). Homogeneous groups based on demographics or other relevant characteristics are also recommended. This is important to elicit conversation among participants. Focus groups in cross-cultural research generally involve culturally homogeneous groups. However, the researcher may consider additional relevant demographic characteristics in forming the groups—for example, Hispanic elderly versus Hispanic teenagers.

The moderator is the most crucial factor to ensure the effectiveness of the focus group. The focus group moderator conducts the entire session and guides the flow of group discussion across specific topics. According to Burns and Bush, the moderator "must strive for a very delicate balance between stimulating natural discussion among all of the group members while at the same time ensuring that the focus of the discussion does not stray too far from the topic" (1995, p. 200).

In analyzing the data, the qualitative statements of the participants are translated into categories or themes and an indication is given of the degree of consensus apparent in the focus groups. The results of the focus groups inform the development of new items for the survey and the modification of existing measures as needed.

CAHPS® Focus Group

A focus group study was conducted on November 7, 1998, at one of the clinics of a local health plan. The participants were recruited from among the Latino patient population of the health plan's clinics in two Los Angeles County communities with high concentrations of Latino people. In order to be considered for participation in the focus group, patients had to be adults (18 and over) and primarily Spanish speaking.

A member of the RAND CAHPS® team moderated the focus group using a scripted discussion guide. The focus

[2]Herdman, Fox-Rushby, and Badia (1997) recommend that qualitative methods of instrument evaluation precede the translation of survey instruments.

group was conducted entirely in Spanish and lasted for approximately two hours. Twelve women, ranging in age from 24 to 73 years old, attended the focus group. Eleven of the participants were from Mexico and one was from Nicaragua. All of the women had been in the United States for many years, ranging from 10 to 23 years.

The specific objectives of the focus group included:

- Determining Latino patients' perceptions about health providers
- Collecting information on communication issues between Latino patients and their providers
- Gathering information on the use of interpreters by Latino patients
- Seeking information on the role of the family in health seeking behavior and in making decisions about healthcare
- Collecting information on Latino patients' satisfaction with their health care
- Determining the most important aspects related to health care for Latino respondents

Briefly, the results of this focus group raised interesting points:

- The provider's communication is highly valued by Latino people. They prefer that a doctor spend enough time with them, that he or she ask them questions, and that he or she provide sufficient information about the patient's illness and medications. Participants were less concerned with the doctor's Spanish-speaking ability (although they do value it) or the doctor's race or gender.
- Participants reported some dissatisfaction with the care that they received from their health plan. Their chief complaints related to issues regarding promptness of care. Specifically, patients complained of difficulty obtaining timely appointments and of long delays in seeing the doctor once they had arrived at the clinic.
- Most of the participants reported problems in using interpreters. They complained about the quality of the translation. In addition, patients reported not discussing certain personal health problems because of being ashamed to speak in front of their interpreter.
- Some participants reported going to Mexico to receive health care, and the rest reported that they too would seek health care in Mexico if they could afford it financially. Among the reasons for preferring the care received in Mexico were the promptness of care, continuity of care, and the provider's communication and approach to care.

The findings from the focus group suggested that the substantive issues covered in version 2.0 of the CAHPS® Survey Instrument are culturally and substantively appropriate. Two of the findings from the focus group are not addressed as part of the survey and require further exploration. The first of these findings centers on the use and quality of interpreters and how this affects provider-patient communication. Although the CAHPS® supplemental item set contains items that ask about the need and availability of interpreters, it does not cover the issue of interpreter quality and the effect of interpreters on communication between a provider and his/her patient. The second of these findings relates to patients who travel to Mexico to seek health care in spite of the fact that they can receive health care from their health plan. This information is being used to field-test additional CAHPS® survey items that address care in Mexico.

Cognitive Interviews

Cognitive-testing techniques are often used in the process of questionnaire development to investigate, assess, and refine a survey instrument (Berkanovic, 1980). Cognitive testing can detect and minimize some sources of measurement error by identifying question items or terms that are difficult to comprehend, questions that are misinterpreted by the respondents, and response options that are inappropriate for the question or that fail to capture a respondent's experience (Jobe & Mingay, 1991).

One of the most common forms of cognitive testing is the cognitive interview to examine the thought processes of the interviewee. There are two forms of cognitive interviews: the concurrent and retrospective approaches. With the concurrent technique, the respondent goes through a process of "thinking aloud" or articulating the thought processes as he or she answers a survey item. In the retrospective or "debriefing" technique, the interviewer asks questions about the survey process after the respondent completes the survey (Harris-Kojetin, Fowler, Brown, Schanaier, & Sweeney, 1999). Verbal probes or follow-up questions may be used in either type of cognitive interview. One common probe is to ask the respondent to paraphrase the survey question. This helps to determine whether the respondent understands the question and gives it the intended interpretation. This may also suggest more appropriate wording for the survey item.

Prior to conducting the cognitive interviews, a structured protocol is developed to ensure that all participants receive similar prompts from the facilitators. The structured protocol is translated. Interviewers are bilingual in both English and the target language and are trained in cognitive interview techniques. Using notes taken during the cognitive interviews and audiotapes of each of the interviews, each interviewer writes up a summary for each interview in English. These summaries are then combined into one report outlining the results of the cognitive testing.

CAHPS® Cognitive Testing

The CAHPS® team completed 150 cognitive interviews in different geographic locations (Harris-Kojetin et al., 1999). Seven cognitive interviews were completed in Spanish in

California during June–July 1996. A concurrent think-aloud technique with scripted probes was used in this case. The Spanish-language interviews were completed with adult women on Medicaid who were receiving AFDC benefits and were enrolled in either an HMO or a fee for service plan through Medicaid.

The primary objectives of the cognitive interviews were:

- To assess whether respondents understood the CAHPS® survey instruments
- To determine the optimal response categories for ratings and reports of care
- To identify the source of problems in comprehension: translation, reading level, survey content, and cognitive task involved

The results of each cognitive interview were summarized in reports and analyzed for points of convergence. In addition, the interviewers were debriefed and asked to provide general feedback on how well the instruments were working and to discuss content areas or issues that were problematic.

For the overall ratings, an adjectival scale (*excellent, very good, good, fair, poor*) was compared with a numeric scale (0–10). Translation was less difficult with the numerical catagories than it was with the adjectival categories. It was particularly difficult to translate "fair" and "poor" into Spanish (Harris-Kojetin et al., 1999).

The cognitive tests were also used to explore whether key words and concepts worked equally well in Spanish and English. Specific wording and terms that were particularly problematic for Spanish-speaking respondents were modified based on the results of the cognitive testing and used to produce instruments that were ready for pretesting.

The interviewers reported that the survey instruments worked better with the respondents who seemed to be more educated or acculturated. Another issue identified by interviewers as problematic was that the instrument presumed that all prospective respondents were reasonably familiar with the terminology and landscape of the health care system in the United States. Familiarity with the system may be common for most Medicare and Medicaid recipients, but it also is related to length of time in the United States and to levels of acculturation, usually lower for non-English-speaking respondents.

Field Test and Analyses (Step 6)

A field test of the translated survey instrument is also recommended. Psychometric analysis can then be used to assess the reliability and validity of the translated survey instruments. Psychometric testing can also be used to test for measurement equivalence across cultural groups. Three types of analysis commonly used are:

1. Reliability estimates, such as Cronbach's (1951) alpha coefficients, to measure the internal consistency of the instrument. Cronbach's alpha is based on the number of items in the scale and the homogeneity of the items. The homogeneity of the items represents an average of the inter-item correlations in a scale and measures to what extent items share common variance.
2. Factor analysis to examine the internal structure of the instrument or construct validity of the scales. In addition, factor analysis can be used to test measurement invariance across groups (Reise, Widaman, & Pugh, 1993).
3. Item Response Theory (IRT) methods provide an ideal framework for assessing differential item functioning (DIF), defined as different probabilities of endorsing an item by respondents from two groups who are equal on a latent trait. When DIF is present, trait estimates may be too high or too low for those in one group relative to another (Thissen, Steinberg, & Wainer, 1993).

CAHPS® Field Test

A pretest of preliminary drafts of the CAHPS® 1.0 survey instruments was conducted as part of the Medicaid field-test data collection conducted by RAND in 1996 (Brown, Nederand, Hays, Short, & Farley, 1999). Only 23 respondents completed the interview in Spanish. All 23 Spanish-speaking respondents completed the interview by telephone. The total number of interviews in Spanish was insufficient to conduct sensitivity analyses to determine whether the Spanish-language instruments were performing like the English-language instruments.

Conclusion

Adept translation of a survey instrument is an integral part of the instrument development process, but it alone does not ensure that a culturally appropriate survey instrument will result. Cross-cultural adaptation of survey instruments requires that the translated instruments be conceptually and technically equivalent to the source language, culturally competent, and linguistically appropriate for the target population. Producing a survey instrument that is culturally appropriate for Latino people in the United States requires subjecting the Spanish-language instruments to rigorous testing. That testing must include conducting focus groups and cognitive interviews that evaluate the cultural appropriateness of the survey content as well as the cognitive task required in the survey instrument, determining the reading level of survey instruments in Spanish, and field-testing the survey instrument to ensure that the survey measures perform equally well in Spanish and English.

The results of the cognitive interviews, focus groups, and readability assessments may require modifying the English version of the survey instruments by adding domains to capture the experiences of Latino consumers, modifying the construction of items in English to make them more "translatable" into Spanish, modifying the Spanish version to accommodate ethnic and regional variations in Spanish lan-

guage use, and simplifying the translation to make the reading level of the document appropriate for the target population.

In order to assess the cultural appropriateness of the CAHPS® 2.0 survey instruments among different Latino ethnic groups and to account for regional variations in care, focus groups and cognitive interviews will be conducted in San Diego, New York, and Miami. By conducting focus groups across these sites, we will incorporate Latino people of Mexican, Puerto Rican, and Cuban origins in our focus groups. The qualitative component of CAHPS® is being done later than we would like. Ideally, this phase would have taken place before finalization of the English-language instrument. Currently, we are also conducting a field study of the CAHPS® surveys among a Medicaid managed care population in the San Diego area. Our goal is to obtain 50% of completed surveys in Spanish.

References

Berkanovic, E. (1980). The effect of inadequate language translation on Hispanics' responses to health surveys. *American Journal of Public Health, 70*, 1273–1276.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research*. Beverly Hills, CA: Sage Publications.

Brown, J. A., Nederand, M. A., Hays, R. D., Short, P. F., & Farley, D. O. (1999). Special issues in assessing care of Medicaid recipients. *Medical Care, 37*, MS79–MS88.

Bullinger, M., Alonso, J., Apalone, G., Leplege, A., Sullivan, M., Wood-Dauphinee, S., Gandek, B., Wagner, A., Aaronson, N., Bush, P., Fukuhara, S., Kaasa, S., & Ware J. E. (1998). Translating health status questionnaires and evaluating their quality: The IQOLA project approach. *Journal of Clinical Epidemiology, 51*, 913–923.

Burns, A. C., & Bush, R. F. (1995). *Marketing research*. Englewood Cliffs, NJ: Prentice Hall.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Gandek, B., & Ware, J. E. (1998). Methods for validating and norming translations of health status questionnaires: The IQOLA project approach. *Journal of Clinical Epidemiology, 51*, 953–959.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology, 46*, 1417–1432.

Harris-Kojetin, L. D., Fowler, F. J., Brown, J. A., Schnaier, J. A., & Sweeny, S. F. (1999). The use of cognitive testing to develop and evaluate CAHPS 1.0 core survey items. *Medical Care, 37*, MS10–MS21.

Herdman, M., Fox-Rushby, J., & Badia, X. (1997). "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. *Quality of Life Research, 6*, 237–247.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16*, 131–152.

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

Keller, S. D., Ware, J. F., Gandek, B., Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J. B., Bullinger, M., Fukuhara, S., Kaasa, S., Leplege, A., Sanson-Wisher, R. W., Sullivan, M., & Wood-Dauphinee, S. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51*, 933–944.

Marin, G., & Marin, B. V. (1991). *Research with Hispanic populations*. Newbury Park, CA: Sage Publications.

Morales, L. S., Weidmer, B., & Hays R. D. (1999). Readability of CAHPS 2.0 child and adult surveys. *Proceedings of the 7th Conference on Health Survey Research Methods*.

Morgan, D. L. (1996). Focus groups. *Annual Review of Sociology, 22*, 129–152.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

Shi, L. (1997). *Health services research methods*. Albany, NY: Delmar Publishers Inc.

Stewart, D. W. & Shamdasani, P. M. (1990). *Focus groups*. Newbury Park, CA: Sage Publications.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holand & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Weidmer, B., Brown, J., & Garcia, L. (1999). Translating the CAHPS 1.0 survey instruments into Spanish. *Medical Care, 37*, MS89–MS96.

Readability of CAHPS® 2.0 Child and Adult Core Surveys

Leo S. Morales, Beverly O. Weidmer, and Ron D. Hays

Purpose

To assess the readability of the Spanish and English language CAHPS® 2.0 surveys.

Background

In recent years, the emergence of managed care has prompted interest in collecting survey information from health care consumers. Many public and private purchasers of care either already administer patient surveys to their beneficiaries or plan to in the near future. However, the growing diversity of the U.S. population poses major challenges for developing such survey instruments. First, the cultural and linguistic diversity of many beneficiary groups requires that surveys be appropriately translated into various languages and adapted for different groups. Second, because patient surveys are often self-administered, attention must be given to survey readability.

Research studies from many sources, including national literacy data, tell us that a large share of U.S. adults can only read at very basic levels. This problem is particularly striking among Medicaid beneficiaries. According to the 1993 National Adult literacy survey (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993), 75% of welfare recipients read at or below the eighth-grade level and 50% read at or below the fifth-grade level.

Moreover, low reading skills may be more concentrated among certain Medicaid beneficiary subgroups than others. For instance, immigrants and refugees from less-developed countries may be more likely than other, U.S.-born Medicaid beneficiaries to have low educational attainment and, as a result, low reading skills. Among recent Central American immigrants and refugees entering the United States from El Salvador and Guatemala, nearly 80% reported less than a high school education (Lopez, 1996). Among foreign-born Hispanic people living in the Los Angeles region, 10% reported no schooling, 38% reported elementary school only, and 21% reported some high school education (Cheng & Yang, 1996).

The mismatch between an intended respondent's reading ability and the survey instrument may have important implications for the validity of patient satisfaction research, particularly for self-administered surveys. Some of the consequences of this mismatch may include low response rates, especially in vulnerable populations, and unreliable responses because of poor item comprehension.

This study assesses the readability of the English and Spanish versions of the Consumer Assessments of Health Plans Study (CAHPS®) 2.0 adult and child core surveys. The linguistic and cultural adaptation of these surveys is discussed in a separate paper (Weech-Maldonado, Weidmer, Morales, & Hays, 1999).

The CAHPS® Surveys

CAHPS® is a 5-year initiative that aims to produce a set of standardized survey instruments that can be used to collect reliable information from health plan enrollees about the care they have received and their experiences with their health plan. The results of the surveys are turned into reports that provide decision support to other consumers selecting a health plan.

To date, several instruments have been developed as part of this study, each targeting a specific population served by health plans throughout the U.S. CAHPS® has also developed surveys for children, designed for a proxy respondent. Although variations exist between the different versions of these instruments depending on the target population and the age of the respondent, a core set of survey questions is common to all versions of the survey. Five specific domains of care (getting needed care, getting care quickly, communication with providers, office staff courtesy and respect, and health plan customer service) and global ratings (care overall, personal doctor or nurse, specialist care, and health plan) are assessed in the CAHPS® 2.0 surveys.

The CAHPS® investigators recognized the need to translate its instruments into other languages. Indeed, the CAHPS® survey instruments were translated into Spanish (Weidmer, Brown, & Garcia, 1999) because many participating health plans are located in states that have large numbers of Spanish speakers, including Texas, California, New Jersey, and Florida. Hence, we evaluate the Spanish versions of the adult and child core surveys along with the English survey.

Assessing Readability

Two major approaches are available for assessing the readability of documents—measurement and prediction. Measuring

The authors are at the University of California at Los Angeles (LSM and RDH), and RAND, Santa Monica, CA (LSM, BOW, and RDH).

readability, by judgment or comprehension tests, involves using readers. Readability by judgment is usually obtained by asking literacy experts to determine the readability level of a document based on their experience or on use of an algorithm. Readability by comprehension test is obtained by administering a reading comprehension test based on the written material to readers of known ability. A test score criterion is chosen that defines comprehension of the material. When a proportion of readers of similar ability achieve that score, the reading ability of the test takers corresponds to the readability level of the document.

In the second approach, mathematical formulas predict the readability of a document. Unlike judgments or comprehension tests, readability formulas do not rely on readers to establish the readability level of written materials. Because no measurements are made, readability formulas are strictly predictive tools.

The selection of readability technique depends upon time, availability of subjects, level of resources available to conduct the assessment, and the degree of accuracy required in assessing the materials for the target groups (Klare, 1974). Predicting readability by formulas does not involve readers and is therefore much less expensive, but it only provides an approximate indication of the readability of a document. Measurements obtained by tests and by experts require greater resources but provide more accurate assessment of readability. We chose to use the former approach for this study for two reasons. First, prior research had addressed the readability of the CAHPS® surveys through cognitive interviews (Harris-Kojetin, Fowler, Brown, Schnaier, & Sweeny, 1999) and expert judgments (Brown, Nederend, Hays, Short, & Farley, 1999). Second, available resources restricted us to using readability formulas.

To identify appropriate readability formulas for our study, we conducted a literature search. Our goal was to identify formulas appropriate for survey instruments in Spanish and English. Although we found references to numerous readability formulas, we did not identify any formulas appropriate for evaluating survey instruments in English or Spanish. The principal problem with applying readability formulas to survey instruments is that the formulas become unreliable when applied to passages of fewer than 100 words (Fry, 1990). Because the CAHPS® surveys are composed of multiple closed-ended questions followed by a set of response options, passages of fewer than 100 words are common. Furthermore, the vast majority of formulas we identified were appropriate for English written materials but not Spanish.

Most readability formulas typically use two factors in their calculations: a sentence, or syntactic factor; and a word, or semantic factor (Rush, 1985). Formulas using these two factors include the Fry Readability Graph (Fry, 1965), Dale-Chall (Dale & Chall, 1948), FOG (Gunning, 1968), Flesch (Flesch, 1948), and Flesch-Kincaid (Kincaid, Fishburne, Rodgers, & Chissom, 1975). The SMOG (McLaughlin, 1969) is an exception because it has only a syntactic factor. The syntactic factor frequently estimates the grammatical complexity of the writing by using sentence length. The semantic factor purports to measure the degree of difficulty of the vocabulary in a piece of writing. Readability formulas usually estimate semantic load either with a measure of word length (such as number of syllables) or with a count of unusual words. Thus, the assumption that word and sentence length are reasonable correlates of syntactic complexity and semantic load underlies readability formulas (Rush, 1985).

Readability formulas are typically validated against performance criterion passages of varying but known levels of difficulty. Two common sources of criterion passages are the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1961) and the Gates-MacGinite reading tests (Gates & MacGinite, 1965). The validity of a particular readability formula is determined by how accurately it predicts the grade level of a criterion passage. In addition, the validity of more recent formulas is established in part through correlation with older formulas.

In addition to using a readability formula, some investigators have chosen to describe readability using a variety of counts of syntactic and semantic factors (Leadbetter, 1990). Fry recommends the use of word counts and sentence length to assess the readability of passages having fewer than 100 words (Fry, 1990). Because readability formulas were not originally intended for survey instruments, we have supplemented the readability formula results with counts of a variety of syntactic and semantic factors (see Table 1).

Adapting Survey Instruments for Readability Assessments with Formulas

Using readability formulas to assess the CAHPS® surveys required us to exclude the question response scales, leaving only the instructions, question preambles, and the survey questions themselves. The question response scales were deleted from the text of the surveys because they do not have a sentence structure, which readability formulas assume.[1]

Fry Readability Graph

The Fry Readability Graph (FRG) is the principal readability assessment tool used in this study because it has been validated for Spanish- and English-language documents. Like most readability formulas, the FRG has syntactic and semantic factors—sentence length and syllables. To implement the FRG, one first randomly selects three sample passages of exactly 100 words—from the beginning, middle, and end of the source document (our source documents consisted of the CAHPS® surveys, stripped of all response options). After the total number of sentences and syllables for each of the 100-word passages has been recorded, the average number of sentences and syllables is computed. The resulting figures are plotted on a graph, and the resulting coordinate point is associated with an established grade-level designation. An illustration of the FRG is shown in Figure 1. The FRG is

[1]Other researchers have turned response options into sentences and included them in their readability analysis (Lewis, Merz, Hays, & Nicholas, 1995).

Table 1. Synatic and semantic factor counts in used in readability assessment

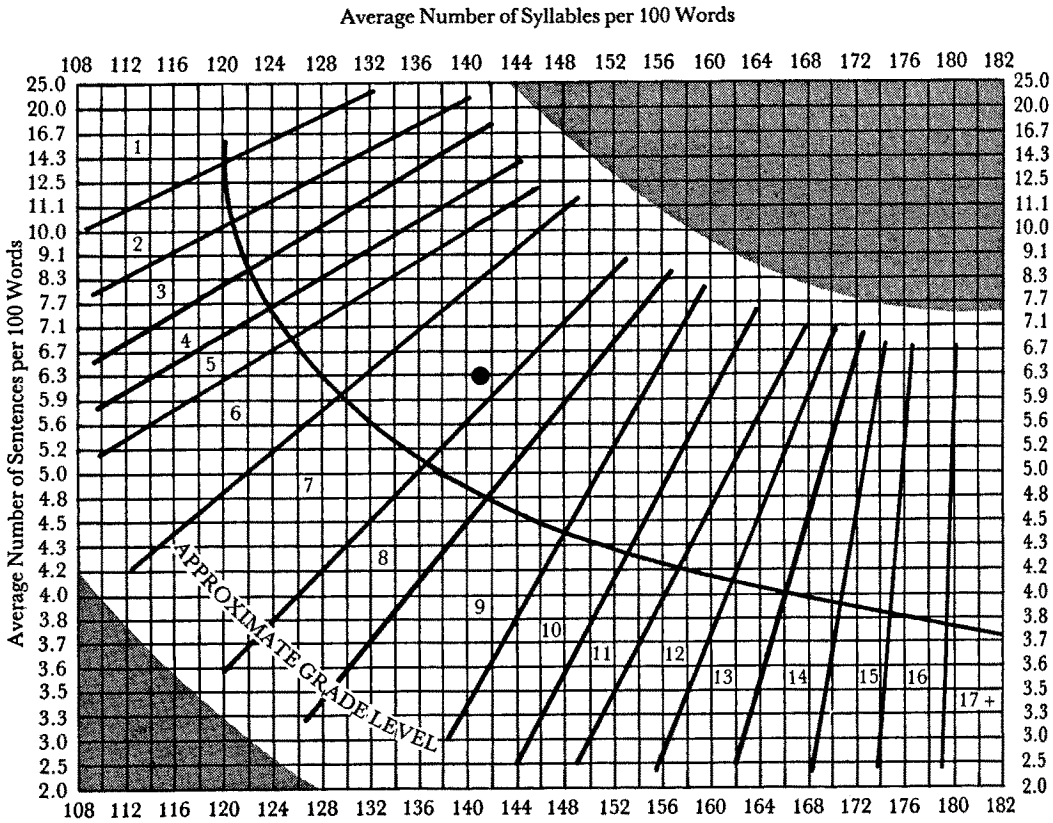
	Syntactic (sentence) Factor	Semantic (word) Factor
Average number of sentences	√	
Average number of words per sentence	√	
Average number of syllables per sentence	√	
Number of characters per sentence	√	
Average number of syllables		√
Average number of one-syllable words		√
Average number of two-syllable words		√
Average number of characters per word		√
Average number of syllables per word		√

appropriate for assessing materials from the first grade through the college level (Fry, 1969, 1977).

The FRG is one of the few readability assessment tools that is adapted for Spanish-language documents (Gilliam, Peña, & Moutain, 1980). Spanish-language application of the FRG is similar to its English-language application, with the exception of syllable counting. Because of differences in the structure of words in the two languages, the syllable counts for 100-word passages in Spanish tend to be much higher than for the same passage in English. To correct for this discrepancy, 67 is subtracted from the total syllable count for each 100-word passage in Spanish (Gilliam et al., 1980).

The comparability of the FRG applied to Spanish-language documents (with the adaptation) and English-language documents has been assessed. Using Spanish primary textbooks, the readability level of the FRG and the publisher's grade level were compared. In 10 of 12 cases, the FRG grade level and publisher's grade level were the same (Gilliam et al., 1980). Unfortunately, a similar comparability study has not been conducted using the FRG for documents at higher reading levels.

Figure 1. Graph for estimating readability–extended (by Edward Fry, Rutgers University Center, New Brunswick, Reading NJ 08904).



DIRECTIONS: Randomly select 3 one hundred word passages from a book or an article. Plot average number of syllables and average number of sentences per 100 words on graph to determine the grade level of the material. Choose more passages per book if great variability is observed and conclude that the book has uneven readability. Few books will fall in gray area but when they do grade level scores are invalid.

Count proper nouns, numerals and initializations as words. Count a syllable for each symbol. For example, "1945" is 1 word and 4 syllables and "IRA" is 1 word and 3 syllables.

FRASE Graph

The FRASE graph (Fry Readability Graph Adapted for Spanish Evalulation) is a readability assessment tool specifically developed for Spanish written materials (Vari-Cartier, 1981). The FRASE graph addresses two limitations of the FRG by increasing the syllable count range beyond 182 per 100 words and altering the readability designations from grade levels to the reading difficulty designations used in English as a Second Language instruction (beginning, intermediate, advanced intermediate, and advanced).

The FRASE graph is derived from the FRG, also basing its readability assessment on a syllable and sentence count. However, the FRASE graph uses five 100-word samples rather than three.

The FRASE graph has been extensively validated using subjective teacher judgments, Spaulding formula scores, cloze test scores, and informal multiple-choice test scores. Correlation coefficients between the FRASE graph readability designations and the alternative readability estimates ranged from 0.91 to 0.97, indicating that the FRASE graph is equivalent to other established methods for estimating readability (Vari-Cartier, 1981).

FOG Index

The FOG Index was developed by Gunning (1968) and uses as few as 100 successive words to determine both sentence length and the number of words with three or more syllables. The counts are then substituted into a formula,[2] and the reading difficulty is calculated according to formal grade level in school. For longer written works, the author recommends selecting several 100-word samples from various parts of the material and averaging the results to determine the reading level. This formula is appropriate for assessing materials from the fourth grade through the college level.

The FOG Index has not been adapted for Spanish-language materials.

SMOG Grading Formula

The SMOG grading formula is based solely on syllables. It was developed by G. Harry McLaughlin as a fast and accurate test of readability (McLaughlin, 1969). The SMOG grading formula estimates the grade level of a document by counting the number of polysyllabic words (words with three or more syllables) in three chains of 10 consecutive sentences taken from the beginning, middle, and end of the assessed document.

An advantage of the SMOG is that the standard error of the readability prediction has been estimated (SE = 1.5 grades) based on validation studies using the McCall-Crabbs passages. A standard error of 1.5 grade levels means that the material being tested will be fully comprehended[3] by 68% of its readers who have reached a reading skill level within 1.5 grades of the SMOG score.

The SMOG grading formula has been adopted by the National Cancer Institute as the preferred method for assessing the readability of cancer communications after a comprehensive review of advantages and disadvantages, including how well alternative formulas predict readability (Romano, 1979).

The SMOG grading formula has not been adapted for Spanish-language materials.

Flesch Reading Ease Score

The Flesch Reading Ease score is one of the most widely used readability assessment formulas. Rudolf Flesch published his first reading formula in 1945, based on the number of affixes, the average sentence length, and the number of personal references. He subsequently introduced the Reading Ease formula, which is based on number of syllables per 100 words and average number of words per sentence. When applied to a document, the Flesch Reading Ease formula results in a number ranging from 0 to 100. The lower the score, the more difficult the material is to read and comprehend. The Flesch Reading Ease score has been validated against the McCall-Crabbs passages (Klare, 1974).

Studies have shown that scores of 90–100 characterize most comic books, scores of 60–90 characterize articles from the popular press (e.g., *Better Homes and Gardens, Newsweek*), and scores of 20–30 characterize reports from medical journals (e.g., *Journal of the American Medical Association, New England Journal of Medicine*) (Morrow, 1980).

In the computer adaptation of the Flesch Reading Ease formula, the syllable count is replaced by a vowel count, something computers can do more easily. Research by Coke and Rothkopf (1970) has shown that counting vowels provides estimates very similar to counting syllables.

The Flesch Reading Ease formula has not been adapted for Spanish-language materials.

Findings

English-Language Adult and Child CAHPS® Surveys

Table 2 shows the readability formula and word and sentence difficulty results for the CAHPS® 2.0 adult and child core English-language surveys. The average number of sentences and the average number of syllables are the main indicators of syntactic and semantic complexity used in all readability formulas except the SMOG. The average number of sentences per 100-word sample was 5.1 for the adult survey and 7.9 for the child survey. The average number of syllables per sentence per 100-word sample was 134.0 for the adult survey and 124.3 for the child surveys. In general, lower

[2]FOG readability formula: Grade level = 0.4(average sentence length + percentage of words with three or more syllables).

[3]This corresponds to the reading ability, indicated by the grade placement score, needed to answer 100% of test questions on the McCall-Crabbs passage for that grade level (Klare, 1974).

Table 2. Readability levels of English-language CAHPS® 2.0 surveys

	CAHPS® 2.0 Adult Core	CAHPS® 2.0 Child Core	*New York Times* Article	*Cricket* Reader (ages 9 and up)
Fry Readability Graph score	7th grade	7th grade	12th grade	5th grade
Average number of sentences per 100-word sample	5.1	7.9	3.4	9.0
Average number of syllables per 100-word sample	134.0	124.3	153.3	133.0
Flesch reading ease score	71.3	89.6	45.8	81.3
FOG readability score	8th grade	6th grade	12th grade	5th grade
SMOG readability score	7th grade	7th grade	12th grade	7th grade
Syntax indexes				
Average number of words per sentence	19.8	15.2	30.7	11.4
Average number of syllables per sentence	26.5	18.9	46.4	15.1
Average number of characters per sentence	81.2	61.3	141.7	49.9
Semantic indexes				
Average number of one-syllable words per 100 words	76.3	83.7	65.3	75.7
Average number of two- or more syllable words per 100 words	23.7	16.3	34.7	24.3
Average number of characters per word	4.1	4.1	4.7	4.4
Average number of syllables per word	1.3	1.2	1.5	1.3

Note. The Fry Readability Graph score and Flesch Reading Ease score are based on three 100-word passages taken from the beginning, middle, and end of each document. The SMOG score is based on three continuous 10-sentence samples taken from the beginning, middle, and end of each document.

readability (lower difficulty) is assigned to written materials that have shorter sentences and fewer syllables. The lower average number of sentences and higher average number of syllables in the adult survey may explain why the Flesch Reading Ease score for the adult survey is lower than that for the child survey (a lower score indicates more difficult text).

The FRG scores can be verified by plotting the average number of sentences and average number of syllables on Figure 1. The FRG results show that for both the adult and child surveys, a seventh-grade reading level is required for comprehension.

Applying the FOG Index to the adult and child surveys resulted in similar but not identical results. With the FOG Index, readability levels of eighth grade for the adult survey and sixth grade for the child survey were obtained. These results are consistent with the higher Flesch score obtained for the child survey than adult survey.

Recall that the SMOG readability formula relies exclusively on counts of polysyllabic words found in three strings of 10 consecutive sentences selected randomly from the written material. Results from analyses using the SMOG are in agreement with results using the Fry graph indicating that a seventh-grade reading level is required for comprehension of both the adult and child surveys.

Table 2 also shows the results of the readability formulas applied to a children's story (Kayner, 1999) and an article from a national newspaper (*New York Times*, August 23, 1999, pp. A1, A23). The FRG, Flesch Reading Ease score, FOG Index, and SMOG consistently rated the newspaper article at a higher level than either survey. The results of these analyses place the children's story at a reading level near that of both surveys.

Table 2 also shows the results of counts of syntactic and semantic components of the surveys and other materials. The sentence complexity counts (words per sentence, syllables per sentence, and characters per sentence) indicate that the adult survey had greater sentence complexity than the child survey and the *Cricket* reader, and less sentence complexity than the newspaper article. The counts of semantic factors (one-syllable words, words with two or more syllables, number of characters per word, and number of syllables per word) are less easy to interpret. The newspaper had a greater average number of characters and syllables per word than either survey or the *Cricket* reader, indicating a greater use of longer words. The newspaper had a lower average number of one-syllable words and a greater average of words with two or more syllables, also indicating a greater use of longer words.

Spanish-Language Adult and Child CAHPS® Surveys

Table 3 shows the readability formula and word and sentence difficulty results for the adult and child Spanish-language surveys. The average number of sentences per 100-word sample was 6.8 for the adult survey and 4.4 for the child survey. The average number of syllables[4] per sentence per 100-word sample was 202.0 for the adult survey and 194.3 for the child survey. Although the adult survey has more sentences and more syllables than the child survey, the results of the FRG indicate a seventh-grade reading level for both.

The FRASE graph uses a similar method to the FRG to assess the readability of Spanish-language materials. The FRASE graph results indicate that both the adult and child surveys require an *intermediate* level of reading skill to be

[4]Unadjusted for the greater average number of syllables in Spanish-language materials than in English-language materials.

Table 3. Readability levels of Spanish-language CAHPS® 2.0 surveys, Spanish-language newspaper, and Spanish-language children's book

	CAHPS® 2.0 Adult Core	CAHPS® 2.0 Child Core	*La Opinion*	*Aventuras*
Fry Readability Graph	7th grade	7th grade	14th grade	1st grade
Average number of sentences per 100-word samples	6.8	4.4	2.8	16.7
Average number of syllables per 100-word samples	202.7	194.3	235.0	195.7
FRASE graph	Intermediate	Intermediate	Advanced	Beginning
Syntax indexes				
Average number of words per sentence	15.6	24.0	38.0	6.0
Average number of syllables per sentence	31.4	46.5	88.2	11.9
Average number of characters per sentence	74.0	110.5	191.6	26.0
Semantic indexes				
Average number of characters per word	4.8	4.6	5.0	4.3
Average number of syllables per word	2.0	1.9	2.4	2.0

Note. The Fry Readability Graph score is based on three 100-word passages taken from the beginning, middle, and end of each document. The FRASE assessment is based on five 100-word samples taken from the document.

fully comprehended. While the FRASE graph was intended to gauge the difficulty of materials used to teach Spanish as a second language, these results provide a useful indication of the readability level of the surveys. Furthermore, they provide a means of assessing the comparability of the child and adult survey readability levels.

Table 3 also shows the results of the readability formulas applied to an article from a Los Angeles Spanish-language newspaper entitled "Resultado mixto en reduccion de clase" (*La Opinion*, June 24, 1999, p. A1) and a beginning reader, *Aventuras* (Freeman & Freeman, 1997). Both the FRG and FRASE graphs rate the readability of the surveys lower than the newspaper and higher than the reader.

The syntactic counts (number of words, number of syllables, and number of characters) indicate that on average, the child survey sample had longer sentences than the adult survey. The semantic counts (number of characters and number of syllables per word) for both surveys were similar. The semantic counts of one- and two-syllable words were dropped from this analysis because the higher number of syllables in Spanish-language materials makes them unreliable indicators of vocabulary complexity.

Discussion

The results of this study suggest that the CAHPS® 2.0 adult and child core surveys require a seventh-grade reading level for adequate comprehension. The SMOG and Fry graphs both resulted in a seventh-grade-level readability assessment for the English language adult and child surveys. However, the FOG Index and the Flesch Reading Ease Score indicate that the adult survey may have a higher readability requirement than the child survey. This discrepancy may be due to a greater sensitivity of the FOG Index and Flesch formulas to differences in the number of sentences and/or number of syllables between the adult and child surveys than are possessed by either the SMOG or FRG. While the FOG Index suggests that the magnitude of the difference between the adult and child surveys may be as great as two grade levels, it is difficult to determine the significance of the difference between Flesch scores of 71.3 and 89.6, since these scores are not tied to specific grade levels.

This study also shows that the English and Spanish versions of the CAHPS® surveys have comparable readability levels. Based on the Fry graph, both the English and Spanish adult and child versions for the core CAHPS® surveys have seventh-grade readability levels. The similarity of the readability levels provides support for the success of the translation from English to Spanish.

Although the seventh-grade reading level may be appropriate for commercially insured populations, it may be too high for Medicaid populations. According to the National Adult literacy survey (Kirsch et al., 1993), as many as 75% of welfare recipients read at or below the eighth-grade level and 50% read at or below the fifth-grade level. This suggests that the reading level required by the CAHPS® core surveys for full comprehension exceeds the reading ability of more than 50% of welfare recipients. When one considers particular Medicaid beneficiary subgroups, the mismatch may be even greater.

A recent Public Policy Institute of California study reported that 42% of California Medicaid beneficiaries had less than a high school education (MaCurdy & O'Brien-Strain, 1997). Among recent immigrant Medicaid beneficiaries, 54% had less than a high school education; among Hispanic immigrant Medicaid beneficiaries who had arrived in the United States before 1985, 71% had less than a high school education. Since self-reported educational attainment tends to overstate literacy, the problem of low literacy and illiteracy among these groups is likely to be dramatic.

Poor comprehension of survey questions among those responding to patient surveys may also lead to unreliable results. For instance, adults with low literacy skills may not

comprehend the term "health insurance plan." Indeed, cognitive interviews suggested that Medicaid beneficiaries frequently rated their overall care when asked to rate their health plan (Brown, 1996; Brown et al., 1999). Cognitive interviews also found that Medicaid beneficiaries had trouble understanding the concept of a primary care provider or regular doctor and had trouble differentiating between a health plan and Medicaid (Brown, 1996; Brown et al., 1999).

Limitations of Readability Formulas

It is widely acknowledged that reading is an interactive process that occurs between the text and the reader. In fact, research shows that readers use experiences, knowledge, and information processing skills to comprehend text (Johnston, 1983).

Readability formulas, being strictly text-based, do not address the interactive nature of the reading process. Most reading formulas, including those used in this study, employ syntactic and semantic factors and do not directly address factors related to communicating meaning. For instance, readability formulas do not distinguish between written discourse and nonsensical combinations of words (Dreyer, 1984). Moreover, formulas can not assess other critical factors such as the reader's interest, experience, knowledge, or motivation, all of which may influence the reader's ability to comprehend the cognitive task asked by a survey (Duffy, 1985). Other factors related to readability and not assessed by a readability formula include typographical and temporal factors (e.g., time allotted to complete the reading task).

According to a recent paper on communicating with Medicaid beneficiaries, producing readable health materials requires thinking carefully about the audience to assess whether the intended respondents have the information with which to respond to the kinds of questions the survey asks (Hibbard, Slovic, & Jewett, 1997). It means organizing the material covered by the survey to make the survey easier to respond to, and eliminating extra material that can overflow a page and overwhelm the survey respondent. It also means formatting a survey so that the instructions are simple to follow, and using 12- to 14-point serif type, ample margins, and headers to aid in organization. Finally, the overall content and design of the survey must be friendly, appealing, and culturally appropriate to gain respondents' attention and increase their comprehension of important messages (Root & Stableford, 1999).

Many of the domains mentioned in the paragraph directly above were addressed during the development of the CAHPS® surveys. Cognitive interviews were used to identify items or terms that were difficult to comprehend, questions that were misinterpreted, and response options that that were inappropriate for the question or failed to capture the respondents' true experience (Harris-Kojetin et al., 1999). Literacy experts were consulted to improve readability of the survey (Brown et al., 1999). And careful translation procedures were followed to ensure the comparability of the English and Spanish versions of the surveys (Weidmer et al., 1999). These efforts provide additional evidence of the overall quality of the CAHPS® surveys.

This study is not intended to provide the definitive assessment of the readability of the CAHPS® surveys. Rather, it aims to provide an additional rough gauge of their readability. Incidentally, a readability assessment by two literacy experts placed the readability level of the CAHPS® surveys between the sixth and seventh grades (Julie Brown, personal communication, August 20, 1999).

Conclusions

Although the current readability level of the CAHPS® surveys may be appropriate for commercially insured populations, lower readability is desirable for those who are publicly insured. As many as 50% of welfare recipients may fail to respond to the CAHPS® surveys because of a mismatch between the readability level of the surveys and the reading level of the intended respondents.

This situation may be exacerbated for certain subgroups of Medicaid beneficiaries, such as immigrants and refugees from less-developed countries. According to research, non-English-speaking patients and patients with low literacy skills face the greatest threat of receiving poor quality of care (Baker et al., 1996; Morales, Cunningham, Brown, Lui, & Hay, 1999). Paradoxically, patients with low literacy skills also face the greatest barriers to responding to self-administered quality assessment tools such as the CAHPS® surveys.

Lowering the readability of the CAHPS® surveys, however, may be difficult. For reports about the CAHPS® surveys to help consumers make an informed choice about their health plan, the surveys need to collect information on a range of complex topics that require respondents to be familiar with concepts and vocabulary unique to health care. Shortening the survey and simplifying the vocabulary too much may cause the level of information gleaned from the CAHPS® surveys to fall, defeating the original purpose of CAHPS®.

Finding a balance between collecting important information and maintaining a reasonable level of survey readability will be an important consideration for researchers as future versions of the CAHPS® surveys are developed.

References

Baker, D., Parker, R., Williams, M., Pitkin, K., Parikh, N., Coates, W., & Imara, M. (1996). The health care experiences of patients with low literacy. *Archives of Family Medicine, 5,* 329–334.

Brown, J. (1996). Report on cognitive interviews with Medicaid mothers for the Consumer Assessment of Health Plans Study. DRU-1471-AHCPR. Santa Monica, CA: RAND.

Brown, J., Nederend, S., Hays, R., Short, P., & Farley, D. (1999). Special issues in assessing care of medicaid recipients. *Medical Care, 37*(3), MS79–MS88.

Cheng, L., & Yang, P. Q. (1996) The "Model Minority" deconstructed. In R. Waldinger & M. Bozorgmehr (Eds.), *Ethnic Los Angeles* (pp. 305–344). New York: Russell Sage Foundation.

Coke, E., & Rothkopf, E. (1970). Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology, 54,* 208–210.

Dale, E., & Chall, J. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin, 28,* 37–54.

Dreyer, G. (1984). Readability and responsibility. *Journal of Reading, 27,* 334–339.

Duffy, T. (1985). Readability formulas: What's the use? In T. Duffy & R. Walker (Eds.), *Designing usable texts* (pp. 113–143). Orlando, FL: Academic Press.

Flesch, R. (1948). A readability yardstick. *Journal of Applied Psychology, 32,* 221–233.

Freeman, D., & Freeman, Y. (1997). *Aventuras*. Boston: Houghton Mifflin.

Fry, E. (1969). The Readability graph validated at primary levels. *Reading Teacher, 22,* 534–538.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading, 21*(3), 242–252.

Fry, E. (1990). A readability formula for short passages. *Journal of Reading,* May, 594–597.

Gates, A., & MacGinite, W. (1965). *Gates-MacGinite reading tests.* New York: Teachers College Press.

Gilliam, B., Peña, S., & Moutain, L. (1980). The Fry graph applied to Spanish readability. *Reading Teacher,* January, 426–430.

Gunning, R. (1968). The Fog Index after 20 years. *Journal of Business Communication, 6,* 3–13.

Harris-Kojetin, L., Fowler, F., Brown, J., Schnaier, J., & Sweeny, S. (1999). The use of cognitive interviews to develop and evaluate CAHPS® 1.0 core survey items. *Medical Care, 37*(3), MS10–MS21.

Hibbard, J., Slovic, P., & Jewett J. (1997). Informing consumer decisions in health care: Implications from decision-making research. *Milbank Quarterly, 75*(3), 395–414.

Johnston, P. (1983). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Kayner, G. (1999). Sun flower. *Cricket, 26*(12), 4–7.

Kincaid, J., Fishburne, R., Rodgers, R., & Chissom, B. (1975). Derivation of new readability formulas for Navy enlisted personnel (Branch Report 8-75). Millington, TN: Chief of Naval Training.

Kirsch, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America.* Princeton, NJ: Educational Testing Service.

Klare, G. (1974). Assessing readability. *Reading Research Quarterly, 1,* 62–102.

Leadbetter, C., Hall, S., Swanson, J., & Forrest, K. (1990). Readability of commercial versus generic health instructions for condoms. *Health Care for Women International, 11,* 295–304.

Lewis, M., Merz, J., Hays, R., & Nicholas, R. (1995). Perceptions of intoxication and impairment at arrest among adults convicted of driving under the influence of alcohol. *Journal of Drug Issues, 25,* 141–160.

Lopez, D. (1996). Language and assimilation. In R. Waldinger & M. Bozorgmehr (Eds.), *Ethnic Los Angeles* (pp. 139–163). New York: Russell Sage Foundation.

MaCurdy, T., & O'Brien-Strain, M. (1997). *Who will be affected by welfare reform in California?* San Francisco: Public Policy Institute of California.

McCall, W., & Crabbs, L. (1961). *Standard test lessons in reading.* New York: Bureau of Publications, Teachers College, Columbia University.

McLaughlin, G. (1969). SMOG grading—A new readability formula. *Journal of Reading, 12,* 636–646.

Morales, L., Cunningham, W., Brown, J., Lui, H., & Hays, R. (1999). Are Latinos less satisfied with communication by providers? *Journal of General Internal Medicine, 14,* 409–417.

Morrow, G. (1980). How readable are subject consent forms? *JAMA*, *244*(1), 56–58.

Romano, R. (1979). *Readability in cancer communications: Methods, examples and resources for improving the readability of cancer messages and materials*. Bethesda, MD: U.S. Department of Health, Education and Welfare, Public Health Service, National Institutes of Health, National Cancer Institute.

Root, J., & Stableford, S. (1999). Easy to read consumer communication: A missing link in Medicaid managed care. *Journal of Health Politics, Policy and Law, 24*(1), 1–26.

Rush, R. (1985). Assessing readability: Formulas and alternatives. *Reading Teacher, 39,* 274–283.

Vari-Cartier, P. (1981). Development and validation of a new instrument to assess the readability of Spanish prose. *Modern Language Journal, 65*(Summer), 141–148.

Weech-Maldonado, R., Weidmer, B. O., Morales, L. S., & Hays, R. D. (in press). Cross-cultural adaptation of survey instruments: The CAHPS® experience. In D. O'Rourke (Ed.), *Health survey research methods: Seventh conference proceedings.*

Weidmer, B., Brown, J., & Garcia, L. (1999). Translating the CAHPS 1.0 survey instruments into Spanish. *Medical Care, 37*(3), MS89–MS96.

A Challenge to the Cross-Cultural Validity of the SF-36 Health Survey: Maori, Pacific, and New Zealand European Ethnic Groups

Kate M. Scott, Diana Sarfati, Martin I. Tobias, and Stephen J. Haslett

Introduction

Assessment of the structural model of a questionnaire via factor analysis is a fundamental aspect of determining the instrument's construct validity (Nunnally, 1978; van de Vijver & Leung, 1997). Comparison of the structural model of a questionnaire across different cultural groups sheds light on whether a health status measure is understood and interpreted in a similar manner by different populations. Examination of structure involves observation of the pattern of correlations between the scales of a questionnaire; principal component factor analysis derives "components" that are linear combinations of the scales, representing the dimensions (constructs) that underpin the questionnaire.

The SF-36 is one of the most widely used instruments internationally to measure health-related quality of life (McHorney, Ware, & Raczek, 1993; Ware, Kosinski, & Keller, 1994; Bullinger, 1995; Sullivan, Karlson, & Ware, 1995; Ware et al., 1998). The SF-36 consists of 36 items grouped into eight scales, ranging from 0–100, each measuring a different aspect of health, with higher scale scores representing better self-reported health. The scales are Physical Functioning (PF), Role Physical (RP) (the impact of physical health on performance of everyday role), Bodily Pain (BP), General Health (GH), Vitality (VT), Social Functioning (SF), Role Emotional (RE) (the impact of emotional health on role performance), and Mental Health (MH). The SF-36 was constructed to represent two major dimensions of health: physical health and mental health (Ware, Snow, Kosinski, & Gandek, 1993). This hypothesized two-dimensional structure was supported in a principal component factor analysis of the SF-36 among the U.S. general population (Ware et al., 1994). Additionally, the International Quality of Life Assessment (IQOLA) project (1994–1998), a 4-year project to translate and adapt the SF-36 for use in 15 countries, found that the two-dimensional structure was supported in 9 Western European countries, and that the interpretation of the two derived components as physical and mental health was straightforward and robust across countries and across age and gender subgroups within countries (Ware et al., 1998). Ware et al. also concluded that the IQOLA project factor analysis results confirmed the appropriateness of the SF-36 mental and physical health summary scores (formed using the factor score coefficients) and recommended their use in multinational comparisons.

However, an important qualifier about the IQOLA project results is that although the factor structure of the SF-36 was found to replicate across Western European and U.S. populations, the results of the same analysis in Japan were much more variable (Fukuhara, Ware, Kosinski, Wada, & Gandek, 1998). This should not seem surprising. The two-dimensional model of health, with physical and mental health seen as distinct and largely uncorrelated, rests on the assumption of mind-body dualism. Such an assumption may have widespread credence in the United States and Western Europe, but it cannot be assumed to dominate views of health in all populations.

The New Zealand population is made up of a number of ethnic subpopulations, including New Zealand Europeans (72%), Maori (15%), and Pacific people (6%). Maori are the indigenous people of New Zealand, and the Pacific community is made up of immigrants (and their New Zealand descendants) from the South Pacific islands. New Zealand Europeans are likely to have a similar construction of health as people living in other "Western," English-speaking countries around the world. There is generally an implicit understanding of mental and physical health as separable entities, consistent with the hypothesized two-dimensional (mental and physical) structure of the SF-36. However, Maori and the various Pacific groups have very different traditional concepts of health and disease. Although Maori and Pacific views of health are diverse, some common themes emerge. Health is not differentiated from well-being, and it reflects environmental, social, spiritual, psychological, and physical dimensions. The separation of mind from body is not recognized. Moreover, health is quintessentially a social phenomenon or state, a property of family or even of community, rather than of the individual (Kinloch, 1985; Durie, 1994).

Given these traditional views of health, the possibility is raised that the two-dimensional structure of the SF-36 found in Western European countries may not replicate among Maori and Pacific people. This possibility was tested in the following analysis using the same statistical methods that Ware et al. used in the IQOLA project. Although a test of comparative structural models would be more formally accomplished through confirmatory factor-analytic techniques such as structural equation modeling or maximum

Kate M. Scott, Diana Sarfati, and Martin I. Tobias are at the Ministry of Health, Wellington, New Zealand.

Stephen J. Haslett is at the Statistics Research and Consulting Centre, Massey University, Palmerston North, New Zealand.

likelihood analysis, the statistical assumptions of such techniques are not met by complex (nonrandom) survey data. However, the present analysis used the same published criteria used by Ware et al. (1998) as part of the IQOLA project to compare the principal component factor structure found in nine European countries against the hypothesized two-dimensional model.

Method

Survey Design

The population for the 1996–97 New Zealand Health Survey was defined as the usually resident, noninstitutionalized, civilian population residing in private households. The SF-36 was administered to individuals 15 years of age and older. The sampling frame used a clustered, stratified design based on geographic areas called primary sampling units (PSUs), each containing 50 to 100 dwellings. The total adult sample size was 7,862 (a response rate of 73.8%). Maori and Pacific people were oversampled to improve the reliability of the estimates. The sample size for each ethnic group was New Zealand European, 5,647; Maori, 1,321; and Pacific, 645. Data were collected over a 1-year period.

Statistical Analysis

The analysis was performed using the SAS® (SAS Institute Inc.) and SUDAAN® (Research Triangle Institute) statistical packages. The factor analysis was performed using the Proc Factor component of SAS specifying two factors and varimax rotation. The appropriateness of factor analysis in each group was assessed via the MSA (measure of sampling adequacy) test, which on a scale of 0–1 should be higher than .70, and via inspection of the anti-image correlation matrix, which should show low correlations across the matrix (Hair, Anderson, Tatham, & Black, 1995).

A survey weight (uniquely assigned to each respondent) was applied, which adjusted for varying probabilities of selection among members of the sample population, and post-stratified the age and sex distribution of the sample so that it matched the age and sex distribution of the New Zealand population. The principal components analysis was performed on the weighted data. Comparisons of the factor structure across ethnic groups were age and sex standardized, but further analyses within ethnic group were stratified by age.

The five criteria used by Ware et al. (1998) to evaluate support for the hypothesized two-dimensional physical and mental health model, and adopted in the present study, were as follows: (1) eigenvalues for the first two components greater than 1; (2) greater than 60% of the total variance in scale scores explained by the two components; (3) the PF scale correlating highest with the physical component, followed by the RP and BP scales, and all three scales correlating lowest with the mental component; (4) the MH scale correlating highest with the mental component, followed by the RE and SF scales, and all three scales correlating lowest with the physical component; and (5) the GH and VT scales correlating moderately with both physical and mental components, with the GH scale correlating higher with the physical component and the VT correlating higher with the mental component.

Factor Analysis and Complex Survey Data

As mentioned above, formal statistical comparison of factor patterns between groups is not currently possible through existing statistical software for complex survey data. In fact, the main basis for comparison in this study is how well the factor structure of each group met the criteria used by Ware et al. (1998) to evaluate the hypothesized two-factor model. However, consideration of the standard errors of the correlation matrix (from which the principal components were extracted), and the small degree of influence on these standard errors exerted by the complex survey design, leads to the conclusion that the differences between ethnic groups in the factor patterns presented below are reliable and statistically robust (data not shown).

Results

Descriptive Statistics and Data Quality

Scale internal consistency reliability (Cronbach's alpha) and the percentage of missing data were examined for each ethnic group (data not shown). Alpha reliabilities were generally high across scale and group (.79–.93), although lower on the VT and SF scales among Pacific people (.71 and .70, respectively). The level of missing data was highest in Pacific people but not particularly high in any group by international standards (McHorney et al., 1994; Sullivan et al., 1995). Item-total correlations were also examined for each ethnic group (data not shown). Although these reflected the general patterns that emerged in the scale correlation matrix (discussed below), there was no evidence of "aberrant" items behaving substantially differently in a given group. Item-total correlations in each group conformed to published standards for item internal consistency and discriminative validity (Ware et al., 1993).

Factor Structure

In each ethnic group the MSA test statistic was around .87, indicating high suitability for factor analysis, with generally low anti-image correlations for each group (although these were higher in the Pacific group compared with New Zealand Europeans). Eigenvalues were greater than 1 for the first two factors for both New Zealand Europeans and Maori; for Pacific people the second eigenvalue was .91. Therefore, the first criterion in support of the two-factor model was not met for the Pacific group. However, given the conservative nature of the eigenvalue criterion (Hair et al., 1995), and to enable a check of the other four criteria in the Pacific sample, it was

decided to continue with the rotation of two factors.

To facilitate comparison with the U.S. and IQOLA factor analyses, rotated factors have been labeled physical and mental in a manner consistent with the U.S. and IQOLA results, although these labels have been put in quotation marks for the Pacific sample to indicate the questionable nature of these labels for this group. The factor loadings (the product-moment correlations between scales and components) for each ethnic group are plotted in Figures 1–3.

In the New Zealand European population (Figure 1) the factor structure of the SF-36 was very similar to that found in the U.S. and Western European populations (Ware et al., 1994, 1998). All of the IQOLA project criteria supporting the two-factor model were met, with the minor exception that the relative loadings of the VT and SF scales were exchanged, with VT loading higher on the mental health component than SF. This may have been influenced by the wording change in the New Zealand/Australian version, which substituted the word "life" for the American word "pep."

In the Maori population (Figure 2) the MH and PF scales loaded onto the respective mental and physical components in a similar manner as for the New Zealand European population (second criterion), and the third criterion—that the PF scale would correlate highest with the physical health component, followed by RP and BP—was met. The GH scale also loaded in a similar manner to that found in Western European countries (fifth criterion). However, criterion 4, that MH would correlate highest with the mental health component followed by SF and RE, was met only in the former respect, as the SF and RE scales loaded less highly on the mental component than VT. Nonetheless, both SF and RE had stronger correlations with the mental health component than with the physical health component, consistent with the Western European pattern.

Pacific people (Figure 3), however, showed quite a different pattern, and none of criteria 3 to 5 was met. In the Pacific population it was the VT scale that loaded highest on the "mental health" component (followed by the MH scale) and the RP scale that loaded highest on the "physical health" component (followed by the PF scale). The BP scale, which in Western European populations, New Zealand Europeans, and Maori loaded reasonably well onto the physical health component, loaded higher on the "mental component" in the Pacific population. The same applied to the GH scale. The RE scale showed a similar reversal in correlating more strongly with the "physical component" than with the "mental component."

To assess the stability of the observed structural models across gender and age, the same principal component analysis was carried out with the sample stratified by gender (no difference was found across males and females) and by age (two groups: 15–44 years, and 45 and over). No substantial age-related differences in factor structure were found for either Pacific or New Zealand European groups. In the comparison of younger (<45 years) and older (≥45 years) Maori, however, substantial differences did emerge. Young Maori showed a fairly similar factor structure to New Zealand Europeans and to the Western European standard, but in older Maori there was no statistical support for the extraction of two factors (the eigenvalue for the second factor was .75).

Figure 1. Plot of SF-36 scale factor loadings on orthogonal physical and metal components: New Zealand European (*n* = 5,467)

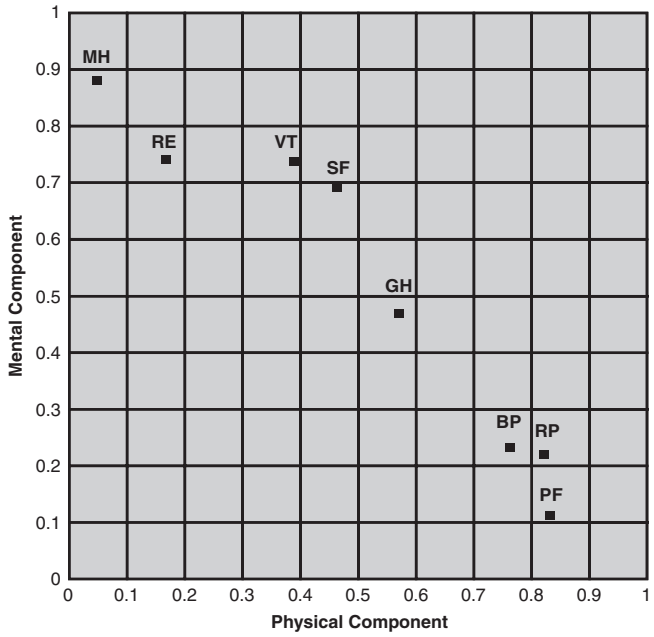


Figure 2. Plot of SF-36 scale factor loadings on orthogonal physical and mental components: Maori (*n* = 1,296)

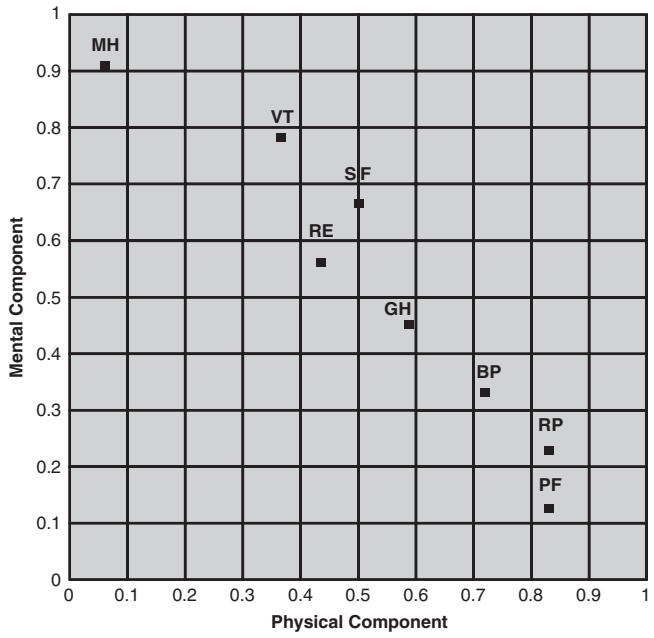
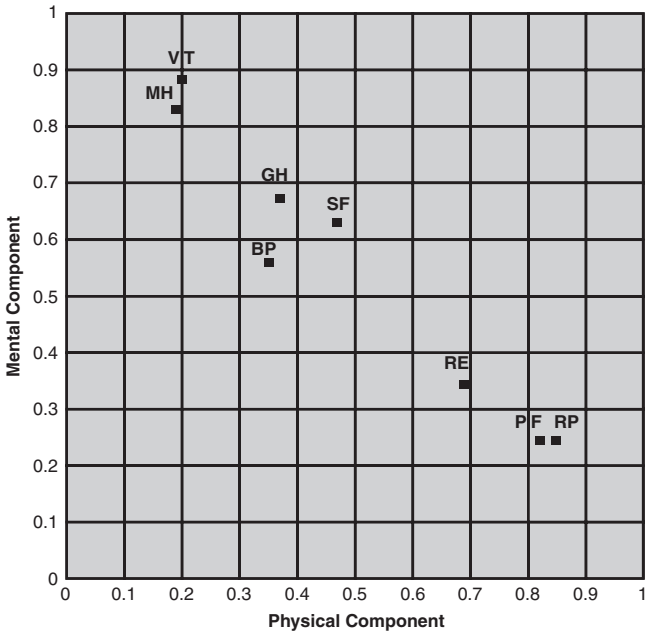


Figure 3. Plot of SF-36 scale factor loadings on orthogonal "physical" and "mental" components: Pacific (*n* = 618)



Scale Correlation Matrix

The scale correlation matrix was examined for each group (Table 1) to shed further light on the ethnic group differences in factor structure. In older Maori, the correlations tended to be fairly high across the matrix (in comparison with younger Maori and New Zealand Europeans), supporting a one-factor structure of the questionnaire in this group. For Pacific people, evidence of clusters of correlations do emerge—in particular, between RE, RP, and SF, and between MH and VT. The low intercorrelation between PF and MH supports a two-factor solution, but given the factor loadings and nature of the correlation clusters obtained for this group, the interpretation of the components as physical and mental health is highly questionable.

Discussion

As expected, for New Zealand Europeans the questionnaire structure was very similar to that found in the U.S. and Western European populations (Ware et al., 1994, 1998). This suggests that for this subpopulation the construct validity of the SF-36 is supported, in that the factor structure is as hypothesized by the developers of the instrument, and similar to that found in a number of other countries.

For the Maori sample as a whole, the factor structure was similar to the hypothesized two-dimensional model. However, the results in Maori were age-related. The two-factor model was supported in Maori under 45 years of age; this may reflect the fact that many younger Maori are urbanized, often with severed tribal ties and weakened cultural affinity, including language. In older Maori only one factor emerged, which argues for the explanation that, despite the construction of the questionnaire items to attempt to discriminate between physical and mental health, traditional Maori views of health that do not differentiate physical and mental health have dominated the interpretation of responses.

Among Pacific people, although there was some suggestion that the questionnaire structure was two-factor rather than one-factor (in the low correlation between the PF and MH scales), statistically the two-factor solution was borderline (the eigenvalue of the second factor being less than 1). Additionally, interpretation of the two factors is rendered highly problematic, with four of the eight scales loading on the opposite dimension of health from that predicted by the hypothesized physical and mental health dimensional model.

With regard to the use of SF-36 specifically, these results should represent a cautionary note in the evaluation of the instrument's cross-cultural validity. In terms of the construct validity of the measure, these results support the findings of Ware et al. (1998) that among Western European and neo-European cultures, the questionnaire is being interpreted consistently to measure partly independent dimensions of physical and mental health. For other populations, however, which may include some ethnic minorities in Western Europe and the United States, this study indicates that the SF-36 may be being interpreted differently, and those differences will complicate ethnic group comparisons of self-reported health status. In particular, the PCS and MCS summary scores derived from the principal component factor coefficients do not appear valid for use among Pacific people and a significant proportion of Maori. These limitations need to be borne in mind in the light of the recommendation by Ware et al. (1998) that these summary scores be used for multinational comparisons.

The more general conclusion that may be drawn from these results is that no matter how carefully an instrument is constructed to reflect a prevailing model of health, it is the respondents' model of health that will determine responses. Where the developer's and the respondents' models of health do not coincide, the performance of the instrument will be seriously affected and cross-cultural validity will be lost or impaired. Of course, it is hardly a new finding that respondents' ethnic background influences how they interpret survey questions (van de Vijver & Leung, 1997), but these issues are often explored at the item and question wording level. The present study emphasizes how higher-level constructions of health (that respondents may not have explicit access to) may also be influential.

In the common metric they provide across disease and population groups, generic measures of health-related quality of life would seem to have great potential in the monitoring of health inequalities and related policy development. However, these results question the ability of any health-related quality-of-life-instrument, regardless of how carefully it has been translated, to reflect anything other than the normative constructs of the society or culture within which it was created. In some (e.g., Western European) countries where there is

Table 1. Scale correlations by ethnic group

	PF	RP	BP	GH	VT	SF	RE	MH
NZ European								
Physical Functioning		0.61	0.50	0.51	0.40	0.46	0.28	0.20
Role Physical	0.61		0.58	0.46	0.45	0.55	0.37	0.25
Bodily Pain	0.50	0.57		0.46	0.46	0.48	0.30	0.28
General Health	0.51	0.46	0.46		0.59	0.48	0.34	0.43
Vitality	0.40	0.45	0.46	0.59		0.59	0.45	0.63
Social Functioning	0.46	0.55	0.48	0.48	0.59		0.57	0.55
Role Emotional	0.28	0.37	0.20	0.35	0.45	0.57		0.49
Mental Health	0.20	0.25	0.28	0.43	0.63	0.55	0.49	
Maori 15–44 Years								
Physical Functioning		0.49	0.38	0.33	0.26	0.36	0.23	0.20
Role Physical	0.49		0.53	0.41	0.39	0.55	0.42	0.27
Bodily Pain	0.37	0.53		0.40	0.39	0.49	0.32	0.28
General Health	0.34	0.41	0.40		0.47	0.42	0.29	0.40
Vitality	0.26	0.34	0.39	0.47		0.55	0.46	0.68
Social Functioning	0.36	0.55	0.49	0.42	0.55		0.57	0.60
Role Emotional	0.23	0.42	0.32	0.29	0.46	0.57		0.49
Mental Health	0.20	0.27	0.28	0.40	0.68	0.60	0.49	
Maori 45 Years and over								
Physical Functioning		0.65	0.58	0.62	0.61	0.62	0.54	0.41
Role Physical	0.65		0.62	0.53	0.60	0.62	0.63	0.39
Bodily Pain	0.58	0.62		0.62	0.64	0.61	0.47	0.53
General Health	0.62	0.53	0.62		0.70	0.58	0.44	0.53
Vitality	0.61	0.60	0.64	0.70		0.67	0.48	0.65
Social Functioning	0.63	0.62	0.61	0.57	0.67		0.63	0.52
Role Emotional	0.54	0.63	0.47	0.44	0.48	0.64		0.44
Mental Health	0.41	0.39	0.53	0.53	0.65	0.52	0.44	
Pacific								
Physical Functioning		0.62	0.34	0.46	0.39	0.46	0.45	0.34
Role Physical	0.62		0.39	0.45	0.40	0.49	0.56	0.35
Bodily Pain	0.34	0.37		0.38	0.47	0.45	0.38	0.38
General Health	0.46	0.45	0.38		0.60	0.49	0.37	0.50
Vitality	0.39	0.40	0.47	0.60		0.57	0.41	0.68
Social Functioning	0.46	0.49	0.45	0.49	0.57		0.58	0.56
Role Emotional	0.45	0.58	0.34	0.37	0.41	0.58		0.47
Mental Health	0.34	0.35	0.38	0.50	0.68	0.56	0.47	

substantial overlap in the origins of many ethnic groups, this may not present such a difficulty. In countries such as the United States, Australia, and New Zealand, however, where there is considerable ethnic diversity and thereby diversity in models of health, the resulting structural inequivalence of measures such as the SF-36 limits their usefulness in gauging ethnic differences in health status. This issue seems set to be an increasing challenge to the cross-cultural validity of health research as changing demographics mean that ethnic minorities are making up increasing proportions of national populations.

References

Bullinger, M. (1995). German translation and psychometric testing of the SF-36 health survey: Preliminary results from the IQOLA project. *Social Science and Medicine, 41,* 1359–1366.

Durie, M. (1994). *Waiora: Maori health development.* Oxford: Oxford University Press.

Fukuhara, S., Ware, J. E., Kosinski, M., Wada, S., & Gandek, B. (1998). Psychometric and clinical tests of validity of the Japanese SF-36 health survey. *Journal of Clinical Epidemiology, 51,* 1045–1053.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Kinloch, P. J. (1985). *Talking health but doing sickness: Studies in Samoan health.* Wellington: Victoria University Press.

McHorney, C. A., Ware, J. E., & Raczek, A. E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31,* 247–263.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Sullivan, M., Karlson, J., & Ware, J. E. (1995). The Swedish SF-36 health survey: I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Social Science and Medicine, 42,* 1349–1358.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey: Manual and interpretation guide.* Boston:Health Institute.

Ware, J. E., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A user's manual.* Boston: Health Institute.

Ware, J. E., Kosinski, M., Gandek, B., Aaronson, N. K., Apolone, G., Bech, P., Brazier, J., Bullinger, M., Kaasa, S., Leplege, A., Prieto, L., & Sullivan, M. (1998). The factor structure of the SF-36 health survey in 10 countries: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51,* 1159–1165.

Methods for Increasing Recruitment and Retention of Ethnic Minorities in Health Research Through Addressing Ethical Concerns

Vickie M. Mays

The National Institute of Health's (NIH) Revitalization Act of 1993 mandated the inclusion of women and ethnic minorities in NIH clinical research studies. By 1994, NIH had developed a set of guidelines that stated that women and members of minority groups and their subpopulations were to be included in all NIH-supported biomedical and behavioral research projects that involve human participants (Hayunga & Pinn, 1996). Despite the increase in the numbers of nonminorities involved in the collection of data in ethnic minority populations, little in the way of training has emerged to educate researchers about the communities they are entering (Cochran & Mays, 1998; Mays & Cochran, 2000).

This is unfortunate on several levels. A lack of knowledge about the historical and social context of the communities being studied robs researchers of valuable insights that should be integrated into their research endeavors. Also, being unfamiliar with the community increases the chance that research conducted with ethnic minorities may produce less benefit and have greater propensity to generate harm. Although in some instances there is little actual physical harm attributable to a particular research project, the relationship between the ethnic minority communities and the research community is harmed. This harm can occur due to perceptions arising out of historical relationships of mistrust and prior research misdeeds that derail current efforts to recruit ethnic minority respondents (Gamble, 1997; Hatch, Moss, Saran, Presley-Cantrell, & Mallory, 1993; Mays & Cochran, 1996).

This paper will focus on modules that can be incorporated into research training early in the process of conducting research in ethnic minority populations.

When and with Whom to Begin the Process

Although unrealistic, the best place to begin the process of integrating community concerns in a way that informs research activities is in the early stages of writing a grant application (Israel, Schulz, Parker, & Becker, 1998). Basically, this is because some of the suggested techniques for improving ethnic minority participation can raise project costs.

Vickie M. Mays is at the University of California, Los Angeles

This paper was supported by grants from the National Institute of Allergy and Infectious Diseases (RO1 Al38216, T15Al07566).

Module 1: Ethics and Responsibility in Working with Ethnic Minority Populations

Giving serious consideration to who the stakeholders are in the research process may influence who one will decide to hire or how many indigenous community members are needed in order to reach recruitment goals. Identifying the stakeholders will also help to identify the benefits and burdens of each in the research process itself.

To understand the previous relationships among the stakeholders, it may be important to understand the complexity, needs, burdens, and benefits of ethnic minorities as participants in the research process (Buchanan, 1996; Turner Advisory Committee on Human Radiation Experiments, 1996; Wermeling & Selwitz, 1993). Many of us conduct our research under the auspices of our university or survey organization. In applying for funds we benefit from the long history and multitude of activities by our colleagues. However, we sometimes fail to recognize that, in working with diverse ethnic populations—many of whom have not had full and equal access to our university's resources, ranging from admission for their children to ease in obtaining medical care—the same organizational name that facilitates our research can carry with it more burden than benefit.

It is important as we design studies to understand not just the past history in general of a particular ethnic group but rather understand in particular the history of our own organizations in relation to the ethnic communities. There are several areas in which having an understanding of both the institution's history with the community as well as the general, overall abuses that have occurred to the ethnic population can be useful in project design. Some of these follow.

Subject Participation

If one is conducting a study of contraceptive practices, for example, there is a long and checkered history, including the testing of early oral contraceptives in Puerto Rican women (Corea, 1985; Davis, 1990). Knowing this history should lead one first to question whether one will be able to enroll and maintain Puerto Rican women successfully in this study. Is the project integrated well enough into the community that there will be strong supporters who can and will work to explain why it is important for the research project to pursue the scientific goals, in the face of this history of abuse?

Knowing the history of abuse, are there nonetheless specific benefits to Puerto Rican women that this study will bring to that population?

Hiring Staff

Often in the hiring of "indigenous interviewers" one looks for someone who is from the particular ethnic group under study. What is often overlooked is that the community is diverse and it may be difficult for one person to serve as the representative both to the research group and on behalf of the community. Rarely in interviewing such persons for jobs with projects do we try to understand who this one person can and cannot represent. Often their class background, where they have chosen to live, the churches they attend, or the social groups to which they belong will facilitate their knowledge and access to a particular segment of the community. Is that segment of the community the one that the project is most interested in recruiting? Will this person be viewed as a credible source? For example, does a university-educated ethnic minority woman of middle-class background with an MPH have much in common with poor women from the west side of Chicago attending county health clinics or Puerto Rican women in Washington, D.C., who receive care in a family planning clinic in the impoverished corridors of the city? Or would such an interviewer's MPH, association with one of the up-scale Catholic churches, and lack of community contacts serve to *limit* whom she can reach?

Module 2: Interplay of Cultural, Racial/Ethnic, and Psychological Dynamics in Research Participation

This module focuses on volunteerism and its cultural/ethnic roots. Social science research has a long history in studies on prosocial behavior, yet little work has focused on how this varies by gender and ethnicity. Prosocial behavior typically arises from unselfish human nature or from the desire for such rewards as fame and recognition, self-fulfillment, empowerment, and the attainment of employable skills or monetary compensation. The extent to which these motivations vary by gender and ethnicity is important to the design of recruitment and retention strategies (Mays, Cochran, & Lin, 2000).

Looking at the history of volunteerism in ethnic minorities, we find that ethnic groups were historically more likely to volunteer in ethnic organizations that have benefited their ethnic groups (Gallegos & O'Neill, 1991). Appeals to ethnic group members that indicate benefits specific to their ethnic groups are more successful. There are a number of ways to accomplish this goal. In a random-digit-dialing survey we conducted of African Americans and Latino/as in South Central Los Angeles, we told participants that if they completed our survey, we would donate money to one of three organizations as an incentive for their consenting to participate. For African Americans, this included the Urban League, the United Negro College Fund, and a well-known local group. A similar list was generated for Latino/as. Even though the mechanisms of getting a small donation to each of them would have made the study impossible, offering to contribute to a minority organization that was known for its contributions to the community was a small but effective gesture of our respect for their time.

Module 3: Culture, Gender, Age, Individual Autonomy, and Community Responsibilities

There are circumstances in which the agreement to participate in research has broad implications for the family or community. The consent and participation in research by an individual may have implications for spouses, significant others, families, or the broader community (Kuczewski, 1996), despite actions in the informed-consent process that focus only on the individual. As an ethical researcher, it may be useful to ask where those close to the participant fit into the participant's decision-making, particularly for ethnic minority families where individualism is at times at odds with cultural values.

Often in recruitment, we ask the person to participate and expect a decision to be made immediately after our discussion. It's a "bird-in-hand" philosophy. Yet there are times when such an approach can harm communitywide recruitment. Allowing potential respondents the time to consult with others can indirectly improve recruitment by demonstrating the project's concern for the community. There are some segments of the ethnic minority community, particularly women, whose participation influences others to participate. It would be well worth our while as investigators to figure out those times when it would be better for participants that we wish to enroll in our study to be encouraged to consult with others first.

Module 4: The ICF (Informed Consent Form)

If researchers have engaged in many of the steps previously recommended, it will be reflected in the informed consent form. Researchers will ensure not only that their ICFs meet the legal requirements of their institution's Institutional Review Boards (IRBs) but that they can truly discuss risks and benefits with their study participants.

Critical to the ICF is not just the preparation but the delivery of this document. Interviewer training should emphasize that a well-written ICF is an opportunity for the study to communicate with the study participant what they can gain from participation. It is also important to find a way to ensure that all passages of the ICF are at a literacy level that will enhance understanding by the study participants.

Many of us often comment that there are study nurses or particular recruiters who are excellent with the study participants. Often the reason is that these particular persons make study participants feel cared for, understood, and listened to. The more that these feelings can permeate other parts of the study, such as in administering the ICF, the greater the chance that respondents will be retained in the study and that they will act as ambassadors for their community. The ICF is one

of those tools that researchers can effectively use to convey to study participants our understanding of what they are sacrificing and how their participation will make a contribution. Taking this process seriously will increase the likelihood that they will stay and encourage others to participate.

References

Buchanan, A. (1996). Judging the past: The case of the human radiation experiments. *Hastings Center Report, 6* (3), 25–30.

Cochran, S. D., & Mays, V. M. (1998). Use of a telephone interview survey to assess HIV risk among African American and Hispanic Los Angeles County residents. In *Proceedings of the 27th Public Health Conference on Records and Statistics and the National Committee on Vital and Health Statistics 47th Annual Symposium.* Washington, DC: USDHHS.

Corea, G. (1985). *The hidden malpractice: How American medicine mistreats women.* New York: Harper & Row.

Gallegos, H. E., & O'Neill, M. (Eds.). (1991). *Hispanics and the non-profit sector.* New York: Foundation Center.

Gamble, V. (1997). Under the shadow of Tuskegee: African Americans and health care. *American Journal of Public Health, 87,* 4–9.

Hatch, J., Moss, N., Saran, A., Presley-Cantrell, L., & Mallory, C. (1993). Community research: Partnership in Black communities. *American Journal of Preventive Medicine, 9,* 27–31.

Hayunga, E. G., & Pinn, V. W. (1996). NIH response to researcher's concerns. *Applied Clinical Trials, 5* (11), 59–64.

Israel, B. A., Schulz, A. J., Parker, E.A., & Becker, A. B. (1998). Review of community-based research: Assessing partnership approaches to improve public health. *Annual Review of Public Health, 19,* 173–202.

Kuczewski, M. G. (1996). Recovering the family: The process of consent in medical decision making. *Hastings Center Report, 26* (2), 30–37.

Mays, V. M., & Cochran, S. D. (1996). Is there a legacy of Tuskegee? AIDS misbeliefs among inner city African Americans and Hispanics. In *Proceedings of the Eleventh International Conference on AIDS, Vancouver, British Columbia, Canada.*

Mays, V. M., & Cochran, S. D. (2000). Methods for increasing the relevance of telephone and field survey research to community needs. In *Proceedings of the National Center for Health Statistics Conference on National Health Statistics.* Washington, DC: USDHHS.

Mays, V. M., Cochran, S. D., & Lin, C. C. (2000). *Volunteering behavior of women and ethnic minorities: Implications of a legacy of discrimination and benign neglect for their participation in health research and clinical trials.* Monograph in preparation.

Turner Advisory Committee on Human Radiation Experiments (1996). Research ethics and the medical profession: Report of the Advisory Committee on Human Radiation Experiments. *Journal of the American Medical Association, 276* (5).

Wermeling, D. P., & Selwitz, A. S. (1993). Current issues surrounding women and minorities in drug trials. *Annals of Pharmacotherapy, 27* (7–8), 904–911.

Issues in Turning Concerns about Culture and Survey Error into Scientific Questions

Robert M. Groves

In this discussion I will give attention to the problem of studying the effects of cultural variation on survey quality, with only occasional references to the papers in this session to illustrate the central points of my remarks. The discussion will begin with a review of traditional frameworks of error in survey statistics and how the field has linked methodological research to common practice.

Framework for Survey Quality

In the past two decades, the dominant paradigm for notions of the quality of a survey statistic has been "total survey error," which notes that the deviation between a full population value of some statistic and its sample-based estimator is a function of many components. While only sampling variance is typically measured, both variable and fixed components of error can arise from coverage, nonresponse, sampling, modes of data collection, interviewers, respondents, and processing. Outside of that framework is a quality issue central to the papers of this session—the gap between the construct and its measures.

One way to portray this is Figure 1, which represents two parallel sets of logical activities that inevitably take place in every health survey. On the left side of the figure is the process by which the measurement procedures of the survey are determined; on the right is the process by which the persons measured by the process are identified.

The measurement process begins with an abstraction, the construct to be measured. This might be something as elusive as "health status" or something as tangible as "number of doctor visits in the last two months." In essence this is an idea, often describable in words, but not necessarily easily subject to succinct definition. The construct evokes in the mind of the researchers a set of potential indicators. These indicators in health surveys are often verbal questions posed to the survey respondent by an interviewer. They might also be electronic or medical measurement devices. These measures, when implemented, produce individual responses, the values of the respondents on these measures.

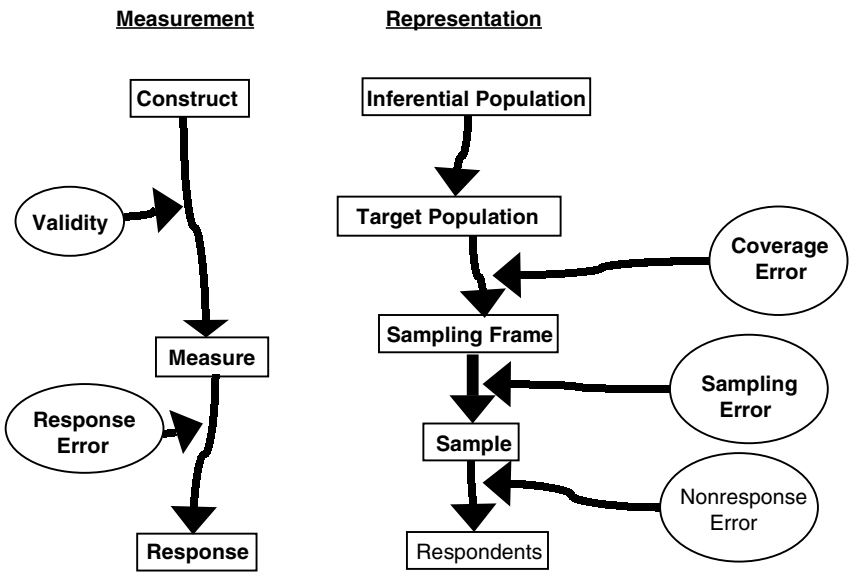
Paralleling the movement from abstract constructs to delivered answers for individual questions on the measurement side is a similar movement from the abstract to the concrete on the representation side. These activities determine what groups are to be described or studied by the measurement process. The activities begin with some notion of what population manifests the behavior or attributes of interest to the research. In many scientific studies, the real population of interest is nearly infinite in size (e.g., all humans who have lived or will ever live). In many descriptive health surveys, it is a population of fixed spatial and temporal dimensions (e.g., residents of the city of Los Angeles in 1997). Generally, however, this "inferential population" has attributes that make it difficult to fully describe. For example, how are persons who live in Los Angeles for only a short time during the year to be made eligible for the survey? Thus, a "target population," a more practical, concrete definition of the population, is described as the population that will be studied by the survey. This might restrict the population to those who can be unambiguously assigned to only one household for a time that includes the months of January and February 1997. The researcher then attempts to obtain the identities of all persons in this target population, through the development of a sampling frame. For this frame a sample design is specified and the subset of the sample to be sought for measurement is designated. Finally, the chosen sample is recruited for the survey and some, the respondents, are successfully measured.

These parallel movements from the abstract to the concrete are connected with key error components in the total survey error paradigm. On the population side, when we fail to measure some in the sample who have distinctive attributes on the survey variables, nonresponse error may arise. When the sample characteristics differ from the full sampling frame, sampling error results. When the sampling frame omits or includes types of population members differentially well, coverage error can result. On the measurement side, the total survey error paradigm focuses on the differences between the responses to a measure and what should be expected from the measure, labeling the result a "measurement" or "response" error.

Why is this review of error concepts relevant to this session? Generally lying outside the total survey error paradigm is the notion of what psychometricians might label "validity," a gap between the construct as created by the researcher and the measures actually used in the survey. In considering the impacts of cultural differences on the quality of survey statistics, one must, however, include the possibility that the new culture to which the survey is to be applied might not

The author is at the University of Michigan and Joint Program in Survey Methodology.

Figure 1. Parallel logical design flows from the abstract to the concrete: Measurement and representation.



incorporate the construct in the same way as the host culture. Alternatively, the construct may have no meaning in the new culture. We have glimpses of this possibility in the paper by Scott, Sarfati, Tobias, and Haslett, where the factor structure of the SF-36 might differ for the older Maori subpopulation in New Zealand.

The problem with focusing on the construct validity of a set of measures is that validity can be empirically measured only based on a specific structural and measurement model. That is, if my theory is correct (construct A causes construct B) *and* my indicators for construct A are highly correlated to those for construct B, then I conclude that the data support the judgment that the items have high construct validity. If my theory (upon which my validity estimates are based) is incorrect, all bets are off.

Hence, if (a) a survey protocol is applied to a new culture and (b) the measures perform differently (e.g., their covariance structure with other measures is different from that for the host culture), then it is possible that the causal structure (appropriate constructs) is different, but it is also possible that the response error properties are different. Distinguishing between the two possibilities is of key importance to the selection of repairs to the quality problem.

Evolution of Error Information Sources

A field's orientation to the quality of survey statistics seems to pass through three stages of evolution:

1. The "positivist" phase, in which analysts and consumers of the survey statistics treat the data as if they were without error. That is, the survey statistics are "truth."
2. A phase in which stand-alone methodological research is used to identify best practices for large sets of surveys, whether or not the essential survey conditions are uniform across the groups subjected to the survey.
3. The phase in which the error properties of survey statistics are available along with the statistics themselves.

In general, health surveys have passed beyond the first phase, and analysts are curious about the nature of the error properties of the estimates. However, in general, most surveys are not at the stage in which error properties are well understood for each of the statistics computed from the surveys. The exception to this is the widespread use of probability sampling techniques, yielding estimates of sampling variance in conjunction with the survey estimates themselves.

How is this relevant to the issue of cultural differences in reactions to survey protocols? Given that best practices are defined by methodological research divorced from the survey itself, the research must question whether the essential survey conditions of the methodological research resemble those of the circumstances for the new culture. For many of the ethnic and language groups described in this session, this may not be true. Most of the important methodological research defining best practices in health surveys was conducted on largely majority populations. Are their results relevant to the new cultures to be subjected to the survey?

Definitional Burdens of Studying the Effect of Culture on Survey Quality

Culture itself is an abstract construct, subject to diverse definitions and alternative indicators. In this session alone,

indicators included race, ethnicity, indigenous status, language groups, age, education, marital status, and sex. To some, "culture" is a set of shared norms and group identity. While each of these indicators might be presented as a candidate, none of them measures the perceived attachment of the person to the group in question. They are thus imperfect indicators.

To illustrate, the construction of a new set of measurement protocols for Hispanic members of a sample might be less effective if applied to all Hispanics, particularly those who identity little and have no language skills with the group.

In short, cultural attachment, as measured by language, shared norms, perceived attachment to the group, and so forth, is a dimension, not a discrete classification.

Research Design Implications of the Error Framework

There are important methodological research implications of the issues raised here. Of most importance is the fact that alternative error sources identified in Figure 1 may be the cause of implementation problems of a survey design in a new culture. The second is that cultural attachments need to be measured in the survey design itself in order to calibrate the utility of culture-specific methods in a survey sample.

First, let's illustrate the problem of locating the source of a survey error specific to a new culture's implementation of a survey design.

Assume (as in the paper by Weech-Maldonado, Weidmer, Morales, and Hays) that a translation–back translation review strategy is taken to adapt a questionnaire to a new language group. Assume that the back translation, in the eyes of the researcher, appears to connote a different meaning for some questions compared with the original text. What is the source of the problem in this case? There are several alternatives:

1. The problem is only "apparent," not real, arising from variation among translators. It would disappear with other translators.
2. The problem is real, arising from the fact that the culture represented by the language group does not incorporate the construct into its perspective on the phenomenon of interest (as Scott suspects is the case with the spiritual dimension of health among Asians).
3. There are language inequivalencies. The idea that back translation will identify weaknesses of the translation is wrong, because there are multiple phrases in each language to apply to the words used, each connoting a slightly different tone of meaning.

Given each of these alternatives, there are radically different fixes. Of greatest importance, because of its foundational role, is whether the construct actually does not apply in the new culture. If this is true, then the entire logical flow on the measurement side of Figure 1 must be redone, focusing exclusively on the new culture.

Another example comes much later, after the researcher has committed to a measurement strategy for the new culture and the old culture. Statistical analysis of the data reveals that covariance matrices from respondents in the two cultures are nonequivalent. The pattern of relationships among variables (usually those involved in some complex pattern of causal relations reflected in a structural model) appears different in one culture versus another. What is the problem in this case?

Again, there are several alternatives:

1. The constructs to be reflected by the diverse variables are nonequivalent in the two cultures. That is, a fundamental logical error was made at the first step of Figure 1 on the measurement side.
2. The adaptation of the host culture's measurement scheme (e.g., translation, mode switches) to the new culture has weaknesses. The indicators chosen are less valid for the new culture than for the old.
3. The response performance of the new culture differs from that of the old. The comprehension, memory retrieval, judgment process, or articulation of the responses differs between the two cultures (e.g., one culture does not commonly discuss within its group the issues being measured; the other does).
4. Any of the other survey design features impacts the measurement process differently (e.g., reactions to the mode of data collection produce different nonresponse characteristics).

Just these two examples are sufficient to illustrate the inherent complexity of the process of adapting a survey to a new culture. Ideally, the adaptation process would have design features that could sort out which of the alternative interpretations above is correct for the given problem.

Unfortunately, most procedures of adapting a survey design to a new culture focus on only one of the potential causes above. The researcher thus exerts a judgment process as to what is the proper cause. This judgment is affected by the researcher's own cultural lenses (as illustrated by Mays) but also by the discipline in which he or she was trained. Thus, anthropologists often tend to discover cultural differences as the cause of survey inequivalencies. Statisticians might, in contrast, seek explanations involving differential nonsampling errors in the two groups.

Other Issues Resembling the Issue of Culture and Survey Design

There are two debates in survey methodology that resemble that of culture and survey design. Lessons might be learned from the resolution of those debates in practice. The first is the controversy about whether structured questionnaires are superior to less structured interviewing procedures (e.g., Suchman and Jordan, 1990; Fowler, 1993). Examination of structured interviews sometimes reveals an apparent

breakdown of "shared meaning" between the interviewer and the respondent. Respondents appear unclear of the meaning of terms in questions, about what their role should be in the response process, or what conversational rules apply to the interview. The result can be the delivery of incorrect answers to questions.

While this debate remains unresolved at this writing, the extreme positions taken appear to be less desirable than some middle ground. One extreme is that each person represents his or her own language and semantic memory group. The measurement process must be customized to each person, potentially involving different words in questions, different probing behavior for the interviewer, and different quesition order. This position, while conceivable, is not practical for surveys of thousands of cases. Indeed, such an extreme threatens the very definition of a survey, which offers standardized information on a sample with known representational properties. It severely threatens the ability to cumulate data from diverse encounters into an interpretable whole describing the target population. There is no assurance that one is measuring the same constructs, even when they might apply equally well to all persons in the sample.

Another debate involves the ability to cumulate data from different modes of data collection. When are telephone and face-to-face interviewing productive of "equivalent" measures? Is it possible that some persons cannot provide comparable data in one mode versus another? Is it necessary to assign different modes of data collection to different parts of the population in order to obtain equally informative responses? The same extreme postures on the debate—each person must have a different mode versus all must have one mode—yield indefensible positions, in my opinion.

These two examples are similar to the discussion of the role of culture and survey design because they illustrate that there is no logical limit to sensitivity to cultural subgroups. Although the desirability of translating an instrument into the language commonly used by a new culture seems obvious, it is not obvious how many different dialects should be represented. Language itself can be operationalized in many different ways; terms mean different things in different regions of a country, despite the presence of a shared language. The extreme would be individualized verbal presentation of each question. That would threaten the ability to "scale up" survey measurements to large operations involving thousands of respondents.

An Ideal Type of Solution

While rejecting the extreme of customizing the survey design to each respondent, one can imagine some sort of "ideal type" of survey design process for two groups, once they have been identified. That is, conditional on the classification of a group as members of a new culture, how might we try to be sensitive to the alternative causes of problems illustrated above?

The preferred process would probably involve parallel construction of the design separately in the two cultures. Only then will the researcher be alerted to possible divergent construct frameworks between the two groups. After a series of qualitative investigations (focus groups, think-aloud interviews, paraphrasing steps), the two developmental groups would be joined together. That is, the research would be exposed to the possibility that the two cultures might have different conceptual frameworks for the phenomenon of research interest. If that were the case, different measurement strategies might apply.

Such a developmental process would be quite expensive, multiplying the design phase by the number of cultures identified in the first judgment. It would remain weak because the initial judgment of what groups constitute a homogeneous culture is based only on cumulative wisdom of the researcher.

Practical Steps

There is little hope for parallel development with large numbers of cultural subgroups for each survey conducted. There may be more hope for the inclusion of major subgroups in the questionnaire development phase, as collaborators in the construction of instrumentation (as is common in HIV studies these days).

Further, it seems scientifically necessary, if one suspects that cultural identification is an influence in the survey process, to measure directly for each sample unit the extent of their identification with each cultural group of importance to the design. That is, if culture affects survey quality, it should be measured as one would include any covariate of a phenomenon in its measurement. With measures in hand, the researcher would be better able to separate out the effects of culture on survey statistics from those of other factors.

Closing Remarks

The issues discussed in these papers are important but complex. Increasingly, they must be the focus of attention in health surveys because of the demonstrated variation in real health status and access to care across subcultures. Unfortunately, however, they are unlikely to yield to simple solutions. Hence, they require ongoing investigation aimed at sorting out when one is observing construct breakdown, alternative causal structures, differential response error, and confounding weaknesses of other survey design features.

References

Fowler, F. J., Jr. (1993). *Survey research methods.* Newbury Park, CA: Sage.

Suchman, L., and Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85,* 232–241.

Racial and Ethnic Populations: Cross-Cultural Considerations in Health Survey Research

Robert L. Santos

The participants in this session are commended for presenting five papers that illustrate the complexities of gathering accurate, meaningful information from multicultural populations. Indeed, the United States is home to a significant number of ethnic and racial subpopulations, offering a rich diversity in values, beliefs, language, and norms. The content of the presentations offered some insight, and the conclusions were often provocative. There are several statistical and methodological concerns with some of the papers, which will be shared later. But this should not detract from the welcome attention to this important area of methodological research in the health survey research (HSR) arena.

The papers generally focus on the cultural appropriateness of research designs and survey measures. These issues are important to survey research when the population of inference features diversity in racial/ethnic composition or when constructs or measures tailored to one subpopulation are applied to another. But certainly, they are not the only relevant methodological issues in cross-cultural survey research. In fact, each phase of the survey research process can benefit from a multicultural perspective—from conception of the research agenda and research questions, through the design of the survey, the sample design and selection, the collection and processing of data, to the data analysis and interpretation of results. More on this appears later.

This paper develops two principal themes, interweaving several topics and issues presented during the session: (1) the impetus for cross-cultural considerations in health survey research and (2) some multicultural perspectives in health survey research.

Theme 1: The Impetus for Cross-Cultural Considerations in Health Survey Research

It is crucial to discuss cross-cultural methodological research within the context of its relevance to health survey research. Why are cross-cultural considerations relevant to health survey research? What is the value added? What motivates the integration of multicultural perspectives into health survey research?

To develop effective health policy, one needs *accurate and reliable information*. Much of this is provided by sample survey data. The survey research industry has increasingly drawn from various substantive disciplines to measure and address a number of threats to survey data integrity (e.g., cognitive testing, focus groups, item response theory). It is tempting to immerse oneself in the intellectual puzzles that cross-cultural survey measurement issues present to the social scientist, health researcher, statistician, and survey methodologist. This is an area of research that is relatively fertile and is ideal for multidisciplinary approaches to uncover nuggets of insight whose results can be applied well beyond the HSR arena. But within the context of health survey research, we should keep in mind that such methodological investigations should be principally driven by the ultimate goals of improving access to care, health care utilization, quality of care, health care cost efficiency, health outcomes, and quality of life. Scientists and methodologists should be guided by these goals in their selection of cross-cultural methodological puzzles.

Vital statistics for the United States show clearly that (relative to the majority white population) racial and ethnic minorities generally have less access to care, have lower levels of health care utilization, engage in higher rates of risky health behaviors, and report poorer health status. For instance, both Hispanics and African Americans have higher incidences of AIDS and, together with Asians and Pacific Islanders, have the highest incidences of tuberculosis and percentages of births to adolescents (Plepys & Klein, 1995). Minorities also tend to be among the poorest and least educated in the nation. For instance, over 40% of Hispanic children and almost half of all African American children in the United States live in poverty. Data from the 1990 Latino National Political Survey show that over 60% of foreign-born adult Mexicans have 0–8 years of schooling (de la Garza, DeSipio, Garcia, Garcia, & Falcon, 1992). And the paper by Morales, Weidmer, and Hays cited data from the 1993 National Adult Literacy Survey showing that half of welfare recipients read at the level of fifth grade or below. People in poverty also suffer higher rates of poor health status and risky health behaviors relative to the majority white population. The health status of racial and ethnic minorities (and the poor in general) is a *significant problem.*

Health disparities between minorities and whites are sufficiently serious that federal policies have been developed to target these underserved, at-risk populations. Research communities see the results of these efforts through grant stipulations requiring that proposed research include women and

The author is at the Urban Institute.

minorities, funding set-asides for minority-focused health services research and research supplements, and training grants to increase the representation of minority scientists. In part, these initiatives are intended to increase a multicultural perspective in the conceptualization and conduct of health survey research. In terms of large research programs, minority oversamples in national health surveys are now a principal survey objective, especially if the survey is supported by federal funds. And, of course, national policy (via Healthy People 2000) has set a high priority for reducing the health disparities between minority/disadvantaged populations and the majority white population.

Thus, the *impetus for cross-cultural research* in HSR methodology lies in (1) the ultimate goal of health survey research to improve the human condition for *all* people; (2) pronounced disparities in risky health behaviors, health status, and access to/utilization of health care between racial and ethnic minorities and the majority white population; and (3) an explicit national policy to reduce these observed disparities in health. It is within this context that the research papers in this session were reviewed. We now incorporate this perspective into a discussion of the role of multicultural perspectives in health survey research.

Theme 2: Some Multicultural Perspectives in Health Survey Research

Despite known health deficiencies across minority populations, despite low levels of education and literacy among the (mostly minority) disadvantaged, despite explicit national policy and resources targeted at reducing disparities in health among racial and ethnic minorities versus whites, and in the face of mounting evidence of measurement error due to the use of culturally inappropriate measures across populations from different cultures, it appears that survey instruments continue to be designed primarily for comprehension by and relevance to the majority population. Furthermore, health survey research continues to be conducted from the perspective of a majority culture. We assert that this state of nature is self-perpetuating and generally reinforcing. (This is consistent with the principal theme of the Mays paper.)

As an illustration, cross-cultural differences in perceptions of physical and mental health were found in the SF-36 instrument by Scott et al. Although cultural differences are recognized, researchers may nevertheless decide to use this scale to achieve comparability with the U.S. population. In such a scenario, the U.S.-based "majority population" research perspective (evidenced by the U.S.-conceived SF-36) tacitly exerts an influence over a culturally different population (i.e., New Zealanders) in the conceptualization and measurement of well-being. The research goal of cross-cultural comparisons takes priority over the need to measure more accurately the culturally diverse, at-risk subpopulation. It is ironic that the goal of cross-cultural comparison between nations is precisely the mechanism that effects culturally inappropriate measurement within a nation (if, in fact, the SF-36 is left intact).

This example reveals the mechanism of influence by the majority population. Researchers from other institutes readily incorporate these multiculturally inappropriate survey questions into their own studies for the sake of standardization, comparability with other studies, and, of course, cost efficiency. But the more use and exposure these survey measures receive across different surveys, the more widely they are accepted in the research community. Before long, a survey scale with recognized limitations in cross-cultural measurement has found wide use in the research community. A standard scale is born. As many in the survey operations industry know, this scenario is realized all too often.

There are other avenues by which culturally inappropriate measures may be dispersed throughout the health survey research community. Consider the large-scale survey contract that includes minority oversampling but fails to explicitly identify in its statement of work an assessment of the cultural appropriateness of the survey instrument. (The incidence of such work statements appears to be decreasing.) In these situations, English versions of the questionnaire are almost always developed, tested, and virtually finalized before work begins on a Spanish or second-language translation. This eliminates or considerably dampens any effort to identify and integrate new constructs, such as perceptions of well-being, that might arise from, say, a focus group of immigrant Latinos. Unless explicit survey goals exist, pressures to maintain cost-competitiveness (at the bidding stage), budget lines (once the contract is awarded), and scheduling (to preserve timeliness of data release) tend to limit the level of resources and the amount of time devoted to culturally adapting an instrument to at-risk populations—those who would benefit most from accurate measurement.

The preceding two examples illustrate some of the processes that result in a recalcitrance to assess, adapt, and if necessary replace survey measures so that they are culturally appropriate to the diverse population of inference. Others could be provided, but space restrictions do not permit.

Perhaps there are more profound ways in which the majority population influences the form, content, and interpretation of health survey research. Consider the formation of research questions and research goals in a very important subarea of health survey research—health services research. Both are developed during the conception of the research project. Research questions that involve patients or human populations are almost always motivated by a conceptual framework of health behavior or health beliefs. For care providers and organizations, conceptual frameworks include organizational theory, resource dependence theory, and many others; for mixed populations (patient, provider, organization), models such as Aday's (1993) framework for vulnerable populations are invoked.

It is necessary to adapt these conceptual frameworks to reflect constructs relevant to racial or ethnic minorities (e.g., family, collectivism, language). They are then used to motivate the research questions and research hypotheses, which in turn drive the research design. For instance, Perez-Stable,

Marin, and Posner (1998) showed that a smoking cessation program for Mexican American adults was effective when it emphasized family health rather than individual health (i.e., used the more culturally appropriate construct of "collectivism" and "family values" rather than "individualism"). This is consistent with a conceptual model that emphasizes the influence of social and familial factors on health behavior, and a research question that hypothesizes the effectiveness of an intervention based on culturally relevant health constructs. Now suppose that a smoking cessation program designed for the "general population" and emphasizing individual health (rather than family health consequences of smoking) was administered to the same Mexican American adults. It is plausible to conclude that the program would not have been as effective. The interpretation of findings might easily be misconstrued (e.g., Mexican American adults are not concerned with their health), when the real problem was a fundamental misunderstanding of the underlying health behavior model. This illustrates the contention that culturally appropriate conceptual frameworks are vital to the development of informed, cogent research questions (and interventions, for that matter). Without them, research questions are misguided and misspecified, leading to inadequate research designs (e.g., missed confounders) and invalid research conclusions.

Turning to the issue of research goal development, the culture of the researchers inevitably influences the formulation of research goals and the utility of research findings. As a hypothetical illustration, consider the CAHPS® surveys. The goals of the CAHPS® include the preparation and dissemination of reports to consumers who are trying to select a health insurance plan. CAHPS® data collection involves gathering consumer data (which requires at least an eighth-grade reading level), analyzing and tabulating the data, and preparing reports for distribution to consumers who are shopping for plans. Clearly, a more highly educated "majority culture" is accustomed to and comfortable with using reports to inform their decision making. But consider the notion that an underserved, lower-educated, culturally different subpopulation will use such reports in the same fashion (or at all, for that matter). That exemplifies a majority population's imposition of a research goal (perhaps *research objective* is a more appropriate term) onto a culturally different minority population. It is unlikely (although possible) that previous research shows that racial and ethnic minorities and other at-risk populations (a) have sufficient reading and quantitative literacy skills to understand such reports or (b) culturally adopt such an analytic framework when selecting a health insurance plan. It is more realistic to expect that such decision-making models for racial and ethnic minorities rely more heavily on recommendations from family, friends, and community leaders (e.g., parish priests) than from careful review of a report (although, admittedly, such an assertion similarly reflects an assessment based on the background and culture of the author).

From a multicultural perspective, two important research questions arise: (1) How do underserved, at-risk populations make decisions about selecting a health insurer (when such a choice is available)? (2) How can CAHPS® information be packaged and disseminated so that it is relevant and useful to the decision-making processes of the at-risk, culturally diverse populations of the United States? Stated succinctly, if the only consumer products are quantitative reports, then CAHPS® appears to be designed for use primarily by the majority population. To take a multicultural perspective on CAHPS® research goals, perhaps research should be conducted on the decision making of consumers from different cultural backgrounds (specifically with regard to selecting health plans). Moreover, it may be worthwhile to investigate alternative methods of dissemination of CAHPS® information to increase relevance and utility to culturally diverse populations.

It should be emphasized that the illustrations in this discussion (specifically for CAHPS®) were developed without full knowledge of the products and dissemination strategies that are planned. So perhaps alternative, culturally tailored strategies are already being developed to provide CAHPS® information to the underserved, multicultural populations who have the opportunity to acquire health insurance. A positive indication in this direction was the use of a focus group of Latino women in the paper by Morales et al. to uncover constructs that were missing in the initial version of the CAHPS® instrument.

This discussion is not intended to be an indictment of health survey research in general or of the survey research community. As a six-year member of the Health Services Research Study Section of AHCPR, I have observed clear evidence in the extramural research arena of growing cultural sensitivity to the issues raised here. (But, of course, there is *much* that still needs to be accomplished.) And there is clear evidence that the federal statistical agencies are taking a leadership role in this arena (e.g., the Directive 15 research to standardize race and ethnic categories). Large-scale survey programs such as the National Household Survey of Drug Abuse and the National Survey of Family Growth have undertaken extensive efforts to improve the quality of data for multicultural, at-risk populations. But such efforts are best viewed within the context of an ongoing program of research rather than a one-time "hit of the reset button." So these efforts must be nurtured and integrated into our research designs, in the same spirit as the concept of "continuous quality improvement."

Thus far, this discussion has asserted that at-risk, culturally diverse subpopulations are those likely to be measured least accurately by survey instruments. But the problems facing health survey research extend beyond measurement error. Problems can also develop when multicultural perspectives are not incorporated into the development of research questions and research goals. The papers in this session support this assertion. Indeed, we assert that cross-cultural perspectives can benefit virtually every stage of the research process. Here are a few broad recommendations that span the research process and, through the infusion of a multicultural perspective, may be expected to benefit the quality and effectiveness of health survey research (this list is by no means complete):

Research Policy, Research Team, Research Question, and Goal Development

Research Policy

Set as a primary research goal the development of culturally appropriate health measures for underserved, vulnerable, diverse populations (e.g., reduce the reading level of self-administered questionnaires, investigate the feasibility of concurrently measuring more than one health construct, concurrently develop non-English-language instruments). Note that this reduces the priority of measuring trends over time.

Large federal HSR programs should exercise leadership by embedding methodological research in the work scope of contracts to address cross-cultural measurement issues, promote extramural research in relevant areas of cross-cultural survey methodology, and continue traineeships for minorities in HSR.

Research Team

Add a multicultural perspective to the research team (before the research questions are developed). Add cultural diversity to the research team at all levels. If needed, utilize available federal funding resources; three can be found at these Web sites:

- http://www4.nas.edu/osep/osephome.nsf
- http://www.hcfa.gov/ord/prioriti.htm
- http://grants.nih.gov/grants/guide/pa-files/PA-99-104.html

Research Question

Utilize a conceptual framework that is sensitive to multicultural populations (adapt the model if necessary). Use the framework to motivate research questions.

Research Goals

Use a multicultural perspective to develop research goals of HSR and to ensure that results can be used effectively in a multicultural environment (i.e., are meaningful and useful to a diverse population).

Research Design

Align the Survey Design to the Population

Include a multicultural perspective when matching the research design to the populations under study. Design parameters include sampling frame, screening and respondent selection methods, data collection (interviewing) mode, instrument characteristics (length, language, constructs, etc.), community involvement, and data collection protocols (including informed consent/IRB (Institutional Review Board) considerations, period of data collection, composition of interviewer staff, training).

Planning

Allow sufficient planning time and resources to develop culturally appropriate instruments, and test data collection protocols across diverse segments of the target population.

Survey Implementation

Instrument Development

Use qualitative research methods to validate the constructs proposed for the survey instrument and improve the measurement properties of the survey items. Conduct cognitive testing and pretests in all languages used in the study. Use an instrument translation process similar to that described in the Morales et al. paper, and allow for changes of questions in all languages to accommodate cultural appropriateness.

Data Collection Protocol Development

Use qualitative methods and pilot testing to develop culturally sensitive protocols for facilitating subject participation, informed consent, and (if the research involves an intervention) a more culturally appropriate intervention.

Interviewer Recruitment and Training

Evaluate the appropriateness of specifying race/ethnicity, gender, and age characteristics of field staff; training staff should be culturally diverse. Allow sufficient training in the languages that interviewers will be using (e.g., Spanish mock interviews).

Data Collection

Conduct quality control assessments in all languages used for data collection. Effective data collection strategies for some racial or ethnic populations (e.g., Native Americans residing in reservations) may require a longer data collection period with considerably different protocols. Allow for such "stratification" of procedures (being careful not to alter the essential survey conditions sufficiently to threaten data quality in other ways).

Post-Survey Data Processing

Coding staff and editing routines should be sensitive to cultural population differences, as well as differences within cultural groups. For instance, Mexicans and Puerto Ricans use identical Spanish terms to denote different levels of educational attainment: completion of high school and completion of

undergraduate college (the literal translation of the terms is "four years").

Analysis and Impact of Findings

Analysis

Conduct analysis with a team whose members collectively have knowledge of the cultures representing the population of inference. This is especially important for the interpretation of results.

Impact of Findings

Oftentimes in health survey research, considerable attention is given to policy implications and recommendations stemming from the research results. Development or enhancement of these recommendations can be gained through the conduct of post-survey qualitative research (e.g., focus groups of research participants, in-depth interviews of community leaders or health care professionals). Post-survey qualitative research can considerably enrich the policy recommendations section of a final report and strengthen the impact of the findings. Similar methods conducted with project staff, interviewers, and respondents can be highly useful for identifying future improvements in survey design and methodology.

Concluding Remarks on Themes

Health survey research is guided in large part by the ultimate goal of improving health care and health status of all members of the population. The papers presented in this session collectively acknowledge the scientific relevance of differences in the cultural concepts of health and health behavior across diverse subpopulations. They generally call for the adaptation of research so that it is relevant and applicable to a culturally diverse population. The goals of this discussion paper are to promote the cultural adaptation of research and to broaden its focus to include the entire research process, from the development of research agendas and priorities to the development of policy recommendations based on research findings. The key to effecting cultural adaptation is to create an environment in the research community that values and integrates multicultural perspectives in health survey research.

Specific Comments on the Papers

It would be inappropriate to conclude this discussion without briefly providing comments on each of the papers presented. The paper by Owens, Johnson, and O'Rourke on item nonresponse investigated differential patterns of item nonresponse across racial and ethnic groups for a variety of surveys. A number of factors in the analysis were identified as statistically significant in differentiating item nonresponse. However, this may be due to the high level of statistical power associated with large sample sizes rather than to substantively meaningful differences. Moreover, it may be appropriate to adjust reported p levels to account for (1) simultaneous statistical inference (using a Bonferroni type of adjustment, say) and (2) incorporation of the complex sample design into the statistical model. Regardless, most raw differences in item nonresponse rates are rather small, so the substantive relevance is questionable (this was appropriately pointed out in the paper).

The Mays paper calls for incorporating cultural and ethical perspectives into the design and implementation of health research. The essence of this paper is that health research should be conducted responsibly, respectfully, and ethically; with a focus on the culture, people, and community being studied; and with an eye toward specific culturally relevant benefits to the community being studied. The time to begin thinking about and planning for this is when the research application is being developed. This is an excellent message and one highly supported by this discussant. However, being heavily involved with survey operations over the years, I do not agree with some of the specific recommendations if they are intended as standards for all health survey projects. For instance, it is well recognized among survey operations staff that the more opportunities one provides a subject to decline participation, the more likely it is that the subject will not participate. Good field interviewers know this and tailor their approach to the door to increase as much as possible the likelihood of the subject's cooperation at the time of a visit (rather than on a subsequent visit). Informed consent is important, and this cannot be compromised in favor of securing a survey respondent. But other methods, such as advance letters and notification of media and community organizations, can be used prior to the first visit. Another concern involves the level of community participation. This too should be *tailored to the research being conducted;* otherwise this could contaminate an intervention or in other respects alter the attitudes, knowledge, or behavior (i.e., the dependent variables) of the community and thereby bias the results of the study.

The paper by Scott et al. focused on cross-cultural validity of the SF-36 instrument. This was an interesting example of how a construct developed for a "majority" population is not necessarily applicable to racial and ethnic subpopulations. The results raise concern about the cultural appropriateness of the SF-36 instrument for subpopulations of the United States such as Asians and Pacific Islanders, Native Americans, Hispanics, and some subgroups of African Americans (e.g., recent immigrants from Africa or the Caribbean Islands).

The paper by Weech-Maldonado et al. discussed the methodology used to culturally adapt the CAHPS® instrument to Spanish-speaking Latinos. While the methods seem generally reasonable, it is worth noting that no objective, quantitative measures were provided in the paper to assess the instrument at the end of the instrument development process. For instance, an assertion is made that the findings of the focus groups suggest that the instrument is culturally and substantively

appropriate, but no data are furnished. The same applies to the cognitive and field testing of the instrument. Essentially, the effectiveness of these procedures is left as a matter of faith.

The Morales et al. paper focuses on readability of the CAHPS® instrument. A number of algorithms were applied to the CAHPS® instrument to predict its corresponding reading level. One concern is that the reading algorithms appear to be designed for application to relatively homogeneous text, rather than to highly structured survey instruments. This tacit assumption of homogeneity is likely the reason for using simplistic, small sampling specifications. But survey instruments are highly structured and carefully sequenced to facilitate administration (e.g., easy questions appear first, and more taxing or more sensitive questions appear later). Thus, the implicit homogeneity assumption will not hold. A second concern is the elimination of response categories from the assessment. At best this means that the readability assessment applies to a *subset* of the instrument. At worst the entire methodology is called into question. Ultimately, the utility of readability formulae may be as a reality check on an instrument prior to the conduct of cognitive testing (with carefully selected subjects). If cognitive testing of a sample of individuals with an eighth-grade reading level shows a uniform level of comprehension across subjects, then readability scores add little if any value.

References

Aday, L. A. (1993). *At risk in America: The health and health care needs of vulnerable populations in the United States.* San Francisco: Jossey-Bass.

de la Garza, R., DeSipio, L., Garcia, F. C., Garcia, J., & Falcon, A. (1992). *Latino voices: Mexican American, Puerto Rican, & Cuban perspectives on American politics.* Boulder, CO: Westview Press.

Perez-Stable, E., Marin, G., & Posner, S. (1998). Ethnic comparisons of attitudes and belief about cigarette smoking. *Journal of General Internal Medicine, 13,* 167–174.

Plepys, C., & Klein, R. (1995). *Health status indicators: Differentials by race and Hispanic origin.* Healthy People 2000 Statistical Notes, No. 10. September 1995, U.S. Department of Health and Human Services.

Discussion Notes, Session 2

Terry DeMaio and Diane Makuc, Rapporteurs

The floor discussion following the presentations and discussants' remarks was lively. There were several strains to the discussion.

Two survey research issues were identified as being important for minority populations. One is applying the best existing methods in designing instruments for collecting data from minority populations (e.g., through cognitive testing with subgroup members); the other is improving what we are capable of doing by developing new methods to measure the health and health needs of minorities.

One way to improve existing survey methods is to involve members of the minority community at early stages of the research. The input of such persons, both in early survey development stages and in the field data collection stage, can improve the appropriateness of the questionnaire terminology for minority subgroups and the motivational aspects of reporting to increase levels of response.

Researchers need to be sensitive to the fact that the participant is the most important person in a survey. This is critical to eliciting cooperation from minority populations as well as others. Unfortunately, the majority of persons conducting research have not been trained to think in this way. It was noted, for example, that frequently those conducting research fail to consider the costs to the subjects of arrangements designed to make things easier for the investigator.

Definition and Measurement of Constructs

There may be cultural differences in understanding the constructs used in health surveys, as well as cultural differences that affect the measurement of these constructs. These are separate issues relevant in both clinical and survey research. There is a need to define constructs that are understandable to all respondents. Cognitive interviews can be useful in addressing some measurement issues if we listen to respondents. Another point raised was that we should start with the simplest solutions before attempting more complex ones, and that we cannot make assertions that there are cultural differences in the definition and measurement of constructs without evidence.

Racial and Ethnic Categories Can Be Misleading

An opinion was expressed that ethnic differences may be minor with respect to construct definitions and that there are probably larger differences by socioeconomic status. It was also suggested that in the United States there is a tendency to categorize populations into a few racial and ethnic groups and to ignore the heterogeneity of persons within those groups. In contrast, in European countries there is a much greater focus on socioeconomic status. It was pointed out that differences between persons of different educational levels are greater than those between different ethnic groups. In addition, there is substantial variation in the characteristics of different Hispanic subgroups such as Mexican Americans, Puerto Ricans, and Cubans. The overall category Hispanic masks important differences between subgroups.

Policy issues

As we are tasked with collecting more and more data about underserved populations, we are asking people for information the value of which may not be apparent to them, and the rewards of participating also may not be immediately apparent. Respondents may ask: Is this research ever going to translate into policy that will address my needs? If research does not improve health, then maybe resources should go into providing services rather than conducting surveys.

Others in the group pointed out that health research has been useful to minority populations. Health surveys have been used to focus attention on areas where improvements are needed and have resulted in programs to address these needs (for example, in the area of childhood asthma). It was pointed out that the value of health surveys in improving health was eloquently discussed by the Director of the Centers for Disease Control and Prevention at the 1999 NCHS Conference on Health Statistics (Koplan, 1999).

The discussion about the importance of having vulnerable populations speak for themselves as their needs are assessed fit well with the Saturday noon session on health policy and the use of data. As Cathy Schoen of the Commonwealth Fund noted in her comments, we need to become more adept at making research data public.

Connections between Researchers and Minority Communities

There was some concern expressed that there is a problem with continuing to ask for participation in studies if we cannot demonstrate that the studies are of benefit to those from whom data are requested. Lack of communication between

researchers and minority communities has contributed to misperceptions about the purpose and usefulness of research and distrust of researchers, affecting the ability of researchers to recruit minority respondents to surveys. It is important that researchers work harder to communicate the results of work addressing minority communities back to those communities in ways demonstrating that participating is in their interests. Helping respondents to see the relevance may help alleviate contact and response problems. More research is needed on how to do this effectively.

Issues addressed by Vickie Mays were noted to be especially useful in this regard. It was noted that agencies can and maybe should incorporate some of her observations as standards for conducting research with vulnerable populations as part of research grants and contracts. Such standards would increase the cultural sensitivity among those who reach out to respondents.

Reference

Koplan, J. P. (1999). Keynote address. National Conference on Health Statistics, National Center for Health Statistics, Washington DC, August 2.

Comparability of Data across Different Modes of Data Collection

A number of pervasive and thorny issues in survey research have affected the quality of survey data in the past and continue to do so, and a major goal of these conferences has been to critically review our progress with respect to these important, persistent issues. A classic example is the issue of survey nonresponse. Ostensibly, the issue of mode effects appears to fit this paradigm, but in some very significant respects it does not. Instead, while the issue of mode effects has been with us for some time and has been in evidence at most of these conferences, fundamental, rapid, and continuous changes in the ways in which data are collected (even from conference to conference!) give it a far different flavor, although the need for review and update is no less urgent. The major culprit, of course, has been the increasing—and rapidly changing—role of technology in the data collection process, notably the increasing use of computer-assisted survey information collection (CASIC) methods.

As Norman Bradburn notes in his discussion (p. 155), "In the beginning there were two modes: face-to-face interviewing and self-administered questionnaires. Then God created the telephone and a new day dawned." The use of computer-assisted interviewing (CAI), either by telephone (CATI) or in person (CAPI), is now standard in virtually every major health survey; and audio computer-assisted self-interviewing (A-CASI), considered somewhat exotic at the time of the previous conference, has been successfully and routinely implemented in several major surveys. Throw in other forms of CASI and the looming use of the Web, and the pace of evolution is clear. Moreover, the pace of change is such that we have virtually no time to "digest" each successive change in technology/mode. The practical effect of all this has been that we do not fully understand the nature and impact of mode effects even for some of our "late 20th century" innovations; and our ability to anticipate and predict the potential effects of different modes used within the same study, or to confidently compare results of studies using different modes of data collection, has been significantly compromised.

The featured papers in this session all involve data and comparisons that bear on this basic theme, although they each emerge from different motivations and objectives. The first (Green and Krosnick) and third (Midanik, Rogers, and Greenfield) papers address a fairly classic issue that has been with us for nearly two decades, and is still with us today (although it is rarely discussed in "polite company")—the conversion of an existing survey from in-person to telephone interviewing. Three other papers address other, less common mode combinations and comparisons, including (1) the influence of telephone versus self-administration on the psychometric properties of quality-of-life measures (Rockwood, Kane, and Lowry), (2) the feasibility of combining telephone surveys with biological specimen collection (Osmond, Catania, Pollack, Canchola, Jaffe, MacKellar, and Valleroy), and (3) the relative impact of two alternative mixed-mode data collection sequences (priority mail/telephone, or the reverse) on physician response rates (Moore, Gaudino, deHart, Cheadle, and Martin). The final paper provides a general, very preliminary review of the advantages, disadvantages, and some of the potential challenges associated with conducting Web-based research, a mode that will likely be used quite extensively in various areas of health survey research by the time of the next conference.

Comparing Telephone and Face-to-Face Interviewing in Terms of Data Quality: The 1982 National Election Studies Method Comparison Project

Melanie C. Green and Jon A. Krosnick

During the last three decades, American survey research has shifted from being dominated by face-to-face interviewing in respondents' homes (based on samples generated by block listing of residences) to telephone interviewing of samples generated by random digit dialing (RDD). Telephone interviewing has many practical advantages, including reduced cost, the possibility of a quicker turnaround time, and the possibility of greater standardization of administration through closer supervision of interviewers. Initially, telephone interviewing had another unique advantage as well: the possibility of computer-driven questionnaire presentation. With the advent of computer-assisted personal interviewing (CAPI), telephone interviewing's edge in this regard is gone, but it continues to maintain its other unique advantages.

Telephone interviewing also has obvious disadvantages. For example, showcards, which are often used to present response choices in face-to-face interviews, are more difficult to employ in telephone surveys, requiring advance contact and mailing of cards to respondents. Telemarketing has also made it more difficult to obtain response rates in telephone surveys as high as those obtained in face-to-face surveys. Furthermore, as of 1998, about 6% of the U.S. population did not have a working telephone in their household, prohibiting these individuals from participating in a telephone survey. Thus, it is not obvious that data quality in RDD telephone surveys will exceed that obtained from block-listed face-to-face surveys.

Over the years, a number of studies have been conducted to compare the quality of data obtained by these two modes. However, these studies have for the most part been atheoretical, looking for potential differences between modes with little conceptual guidance about what differences might be expected and why. Furthermore, the designs of these studies have often involved methodological confounds or limitations that restrict their internal validity and generalizability.

In this paper, we report the results of a new set of analyses exploring differences in data quality across modes. We begin by offering a series of theory-grounded hypotheses about possible mode differences, and we review what little evidence exists regarding their validity. We then report findings from an analysis of data from the 1982 National Election Study Method Comparison Project (MCP), an experiment designed to compare block-listed face-to-face interviewing with RDD telephone interviewing. Our focus is on three aspects of data quality: sample representativeness (gauged in terms of demographics), the amount of effort respondents devote to providing accurate answers (i.e., satisficing versus optimizing), and the extent to which people misportray themselves in socially desirable ways, rather than giving honest answers.

Hypotheses and Literature Review

Sample Quality

There are several reasons why sample representativeness may differ across modes. First, as we mentioned, telephone ownership is not universal, and people without telephones are automatically excluded from an RDD sample. These people may be disproportionately low in income (lack of money often prevents ownership of a working telephone), low in education (education is correlated with income), non-white (race is correlated with income), or young (young people usually have less disposable income and are more transient than older people).

In addition, the determinants of refusal to be interviewed may differ across modes. In particular, people who are socially disenfranchised and feel at greater social risk of manipulation and persecution may be reluctant to participate in telephone surveys, because it is more difficult to be sure exactly who is calling and what consequences might follow from the answers a respondent gives. But when an interviewer who seems to be friendly and trustworthy—and has documentation of his or her identification—appears on a person's doorstep, the importance and legitimacy of the enterprise may be more apparent, making such people less reluctant to participate. This may exacerbate the underrepresentation of individuals with lower social status in telephone surveys relative to face-to-face surveys.

Several studies have compared block-listing face-to-face surveys with RDD telephone surveys in terms of sample representativeness, three employing national samples (Groves & Kahn, 1979; Mulry-Liggan, 1983; Thornberry, 1987) and two employing local samples (Klecka & Tuchfarber, 1978; Weeks

The authors are at Ohio State University.

This research was commissioned by the National Election Study Board of Overseers. The authors wish to express their thanks to James Lepkowski and Robert Belli for their advice and to Kathy Cirksena for her support of this project. We are also grateful to Aldena Rogers and Chris Mesmer for their assistance with study design and data collection.

This research was supported by a grant from the National Science Foundation (SBR-9707741).

et al., 1983).[1] As expected, response rates were higher in the face-to-face surveys than in the telephone surveys. Whereas gender was unrelated to sampling method, RDD samples consistently overrepresented individuals aged 25 to 44 and underrepresented individuals age 65 or above, compared with block-listed samples. RDD samples consistently included a greater proportion of whites and a smaller proportion of nonwhites than block-listed samples, although this difference was not always statistically significant. Four studies found that RDD samples contained a greater proportion of high income respondents and a correspondingly smaller proportion of individuals with low incomes (Groves & Kahn, 1979; Klecka & Tuchfarber, 1978; Thornberry, 1987; Weeks et al., 1983). All five studies found that RDD samples contained more individuals with a great deal of formal education and fewer with little education. All of this suggests that RDD samples underrepresent segments of the population with lower social status. However, these studies only compared the two sampling methods with each other; none used benchmarks (e.g., census data) to assess which sampling method represented the population more accurately. The current study will provide such a comparison.

Satisficing

A second potential set of mode effects involves satisficing. Krosnick's (1991) theory of survey satisficing is based upon the assumption that optimal question answering involves a great deal of cognitive work (Tourangeau, 1984). Yet there are a variety of reasons why people may not expend all this effort. People can shortcut cognitive processing in one of two ways—via either weak satisficing or strong satisficing. Weak satisficing involves executing all the cognitive steps involved in optimizing, but less completely and with more bias. Strong satisficing involves seeking to offer responses that seem reasonable to an interviewer without having to do any memory search or integration of information at all.

The likelihood that a respondent will satisfice is thought to be a function of three classes of factors: respondent ability, respondent motivation, and task difficulty. People who have relatively limited abilities to carry out the cognitive processes required for optimizing and those who are minimally motivated to do so are the most likely to shortcut them. And people are most likely to shortcut when the cognitive effort demanded by a question is substantial.

Interview mode may influence the likelihood of satisficing by affecting respondent motivation and task difficulty. During a face-to-face interview, the interviewer's engagement in and enthusiasm for the process of exchange is likely to be conveyed through visual, nonverbal behavior and is likely to be infectious. Respondents whose motivation flags or who question the value of a survey can observe the interviewer's obvious seriousness and commitment to the enterprise, which may motivate them to generate thoughtful answers. Respondents interviewed by telephone cannot observe such nonverbal cues and so may be less motivated.

Telephone interviews are typically conducted at a quick pace, much quicker than face-to-face conversations normally go. The fast pace makes interpreting questions more difficult and may press respondents to generate answers more quickly and superficially compared with a slower pace of presentation. Therefore, telephone interviewing may increase the likelihood of respondent satisficing and may therefore decrease data quality.

Some previous research offers evidence that tests this hypothesis. Consistent with the satisficing hypotheses, some past studies found more acquiescence in telephone interviews than in face-to-face interviews (e.g., Groves & Kahn, 1979; Jordan, Marcus, & Reeder, 1980). Furthermore, various studies found that respondents said "don't know" significantly more often in telephone interviews than in face-to-face interviews (e.g., Aneshensel, Frerichs, Clark, & Yokopenic, 1982; Aquilino, 1992; Groves & Kahn, 1979; Herzog, Rogers, & Kulka, 1983; Jordan, Marcus, & Reeder, 1980; Locander & Burton, 1976), though one found no significant mode difference (Rogers, 1976). And a meta-analysis by de Leeuw (1992) confirmed a general trend toward fewer "don't know" responses in face-to-face interviews relative to telephone interviews.

Social Desirability

Another consideration relevant to mode differences is social desirability response bias–the notion that respondents sometimes intentionally lie to interviewers (Paulhus, 1984). There is reason to believe that social desirability response bias can vary depending upon data collection mode. Many past studies suggest that people are more likely to be honest when there is a greater distance (both physical and psychological) between themselves and their interviewers. Distance seems to be minimized when a respondent is being interviewed face-to-face in his or her own home. The more remote telephone interviewer has a lesser ability to convey favorable or unfavorable reactions to the respondent, and may therefore be seen as meriting less of the respondent's concern. Consequently, more social desirability bias might occur in face-to-face interviews than over the phone.

Surprisingly, however, the few studies done to date on mode differences do not offer support for this hypothesis. Some studies have found no reliable differences between face-to-face and telephone interviews in reporting of socially desirable attitudes (Aquilino, 1998; Colombotos, 1965; Rogers, 1976; Wiseman, 1972). Other work has found that reliable differences run opposite to the social distance hypothesis. For example, Aquilino (1994) found more reporting of socially undesirable behaviors in face-to-face interviews than in telephone interviews; Johnson, Haugland, and Clayton (1989) found similar results in a college student sample. And Groves (1979) found that respondents expressed more discomfort

[1]Gfroerer and Hughes (1991) compared RDD and block-listed samples. However, because different methods were used to oversample minorities in the two modes, this study does not provide an accurate test of demographic differences between modes. Similarly, Freeman, Kiecolt, Nicholls, and Shanks (1982) compared the two sampling methods but report demographics only for the head of the household rather than the respondent.

about discussing sensitive topics (e.g., racial attitudes, political opinions, and voting) over the telephone than face-to-face. This may occur because the telephone does not permit respondents and interviewers to develop as comfortable a rapport. Consequently, respondents may not feel they can trust their interviewers to protect their confidentiality as much as they might in face-to-face interviews, so they are more reluctant to reveal embarrassing facts. But the limited array of evidence on this point again calls for further testing.

Data

To test the hypotheses that sampling and data collection mode might affect sample representativeness, satisficing, and social desirability response bias, we analyzed data from the 1982 National Election Studies Method Comparison Project, a study designed to explore mode differences. Specifically, this study compared RDD telephone interviews to block listing sampled face-to-face interviews.

Data Collection

The 1982 MCP involved 998 complete or partial telephone interviews and 1,418 face-to-face interviews, all conducted during the three months following the 1982 congressional elections. All of the face-to-face interviews were conducted by the University of Michigan's Survey Research Center (SRC). The telephone interviews were randomly split between the Michigan SRC and the University of California at Berkeley's Program in Computer-Assisted Survey Methods. Essentially identical questionnaires were used for all interviews, although showcards used in the face-to-face interviews were replaced by spoken explanations in the telephone interviews. The survey was similar in length to other National Election Studies (which require approximately one hour to complete) and asked about a range of political beliefs, attitudes, and behaviors.

Measures

The 1982 MCP data set included measures that we could use to gauge three forms of strong satisficing (selection of no-opinion response options, nondifferentiation, and mental coin flipping) and social desirability response bias (Krosnick, 1991). To gauge the tendency to select a no-opinion response option, we employed the seven questions explicitly offering respondents such options. For each respondent, we calculated the percentage of these questions he or she was asked that were answered "don't know"/"haven't thought much," which was recoded to range from 0 to 1.

To gauge the tendency to nondifferentiate (i.e., rate a series of objects identically on a single rating scale), we focused on two batteries. The first was a set of seven 101-point feeling thermometers, and the second battery involved nine ratings of Ronald Reagan's personality traits. For the feeling thermometers, we recoded the 0–100 scale into 10 segments (0–10, 11–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80, 81–90, 91–100). We then counted up the maximum number of identical ratings made by each respondent for each battery. These two numbers were then each rescaled to range from 0 to 1, and the two batteries' scores were averaged to yield a single assessment of nondifferentiation for each respondent.[2]

"Mental coin flipping" was assessed by examining the strength of association between presidential candidate preference (as gauged by the difference between attitudes toward Jimmy Carter and Ronald Reagan measured on feeling thermometers) and various predictors (all variables being coded to range from 0 to 1). Mental coin flipping should attenuate such associations.

Social Desirability Response Bias

Only five questions in the 1982 MCP seemed likely to have widely shared social desirability connotations, involving interest in politics, voting in previous elections, and support for government aid to blacks (the latter among Caucasians only). Admitting animosity toward African Americans is presumably not respectable among Caucasians. Additionally, interest and participation in politics are presumed to be civic virtues in this culture.

House Effects

We approached the assessment of mode differences in two ways. To gain the maximal statistical power by using the full array of cases, we compared the face-to-face interviews to the full set of telephone interviews. However, this comparison confounds mode with house, because Michigan conducted all the face-to-face interviews but half the telephone interviews were done by Berkeley. Therefore, although we did not expect significant house differences, if the standard interviewing practices at either institution differentially encouraged or discouraged satisficing, this comparison would be misleading about the effects of mode per se.

To deal with this problem, we conducted two additional sets of complementary analyses. First, we did statistical tests comparing the extent of satisficing in the Michigan and Berkeley telephone samples, to explicitly test for house effects. We also conducted less powerful tests comparing only the Michigan telephone respondents to the face-to-face respondents.

Results

Sample Comparability

We first examined whether the samples of respondents interviewed face-to-face and by telephone differed in their

[2]Although it may seem reasonable to use the variance of the responses as a measure of differentiation, we do not use this measure because we are interested in whether or not the respondent is giving different answers versus the same answers to a set of questions, not how extreme the differences in answers might be.

Table 1. Unstandardized regression coefficients predicting interview mode with demographic variables

Predictor	Unstandardized Regression Coefficients: All Predictors Entered Simultaneously	Unstandardized Regression Coefficients: Each Predictor Entered Individually
Age	–.08*	–.06**
Education	.03	.07**
Gender	.04	.02
Race	–.10**	–.07**
Income	.16**	.12**
R^2	.02	
N	2097	

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

demographic characteristics. To do so, we initially regressed a dummy variable coded 0 for people interviewed face-to-face and 1 for people interviewed by telephone individually on one of the following demographic variables at a time: age (coded in years), education (coded in years), gender (coded 0 for males and 1 for females), race (coded 0 for Caucasians and 1 for others), and income (coded in a set of discrete categories representing ranges of dollars). As column 2 of Table 1 shows, the telephone respondents were significantly younger ($b = –.06$, $p < .01$), significantly better educated ($b = .07$, $p < .05$), significantly more likely to be Caucasian ($b = –.07$, $p < .01$), and of significantly higher income ($b = .12$, $p < .01$) than the face-to-face respondents. And in a regression using all the demographics to predict mode, the age, race, and income differences remained significant (see column 1 of Table 1).[3]

To assess which mode provided a more representative sample, we compared both samples to data from the 1980 U.S. Census. As shown in Table 2, the face-to-face sample matches the Census figures more closely than the telephone sample. For example, the face-to-face sample differs from the Census by 2.2% in terms of gender, whereas the telephone sample's discrepancy is 4.5%. Likewise, the face-to-face sample misses the percentage of Caucasians by .7%, whereas the telephone sample does so by 4.7%. And the average discrepancy between the face-to-face sample and the Census in terms of the percentages of people in the seven income categories is 3.2% on average, compared to 5.9% on average for the telephone sample. As expected, the telephone sample underrepresented low-income, less-educated, and non-white respondents more so than the face-to-face sample did.

The differences in samples created by block listing and RDD suggest that individuals in telephone samples should be more motivated and able to provide high-quality data. In order to provide a fair test of our hypotheses about data quality, we controlled for these variables in the analyses to follow, as well as other demographics that previous research suggests are related to the use of satisficing response strategies.

[3]Throughout this study, significance tests of directional predictions are one-tailed, and tests of nondirectional predictions are two-tailed.

Table 2. Demographic characteristics of face-to-face and telephone samples and the nation (according to the 1980 U.S. Census)

Demographic	1982 MCP: Face-to-Face	1982 MCP: Telephone	1980 U.S. Census
Gender			
Male	44.7%	42.4%	48.6%
Female	55.3	57.6	51.4
Race			
White	88.5	92.5	83.1
Non-white	11.5	7.5	16.9
Age			
18–24	11.4	14.0	16.0
25–29	12.1	13.4	13.3
30–39	23.2	24.5	19.8
40–49	13.3	13.6	14.2
50–59	13.4	13.6	14.5
60–64	8.0	4.6	6.3
65 and over	18.6	16.3	15.9
Education			
Grade 8 or less	11.2	7.8	15.7
Grade 9–11	11.0	9.6	16.1
High school diploma	34.6	35.1	36.4
Some college	23.2	24.5	17.4
College graduate	20.0	23.0	14.4
Income			
Less than $5,000	10.7	4.4	13.2
5,000–9,999	13.7	12.1	15.9
10,000–14,999	14.8	11.0	15.3
15,000–19,999	10.9	12.0	14.1
20,000–24,999	13.7	13.8	12.4
25,000–34,999	17.3	22.4	15.7
35,000–49,999	11.2	14.2	8.6
50,000 and over	7.7	10.1	4.6

No-Opinion Responses

The first three columns of Table 3 display the mean proportions of no-opinion responses for the face-to-face respondents, the Michigan telephone respondents, and the Berkeley telephone respondents. The first row of the table reports results for the full sample and shows higher levels of no-opinion responding in the telephone samples (Michigan mean = 26%, Berkeley mean = 22%) than in the face-to-face sample (mean = 18%), consistent with the satisficing hypothesis. The difference is more pronounced when only the Michigan data are considered than when the Berkeley data are added in. There were in fact significantly fewer no-opinion responses in the Berkeley telephone data than in the Michigan telephone data ($b = .04$, $p < .05$, $N = 851$).

We tested the significance of the mode effect in two ways. In both cases, we conducted an OLS regression predicting the

Table 3. Analyses comparing rates of satisficing across modes in the 1982 NES

	Adjusted Means			Regression Coefficients							
		Telephone		Mode							
Sample	Face-to-Face	Mich.	Berk.	Mich.	Mich. & Berk.	Age	Educ.	Gender	Income	Race	R^2
No-opinion											
Full sample	.18	.26	.22	.09**	.07**	.03*	–.30**	.01+	–.07**	.05**	.18
				(1682)	(2095)						
Low education	.34	.54	.40	.21**	.15**	.03	–.46**	.07**	–.09+	.005	.09
				(350)	(410)						
High education	.14	.20	.18	.06**	.05**	.00	–.23**	.01	–.05**	.05**	.10
				(1332)	(1685)						
Nondifferentiation											
Full sample	.34	.37	.38	.02*	.03**	–.04**	–.02	–.01+	–.02+	–.01	.02
				(1684)	(2097)						
Low education	.34	.38	.40	.02	.05**	.01	.07	–.02	.06+	–.01	.03
				(351)	(411)						
High education	.35	.37	.39	.02*	.03**	–.05**	.00	–.01	–.03*	–.01	.02
				(1333)	(1686)						

N's appear in parentheses underneath coefficients. The effects of demographics are from an equation including the Michigan and Berkeley data.
+ $p < .10$
* $p < .05$
** $p < .01$

proportion of no-opinion responses with a dummy variable coded 0 for face-to-face respondents and 1 for telephone respondents, and controlling for a series of other demographic variables. This regression was done once only with the Michigan respondents and again with the Berkeley respondents folded in. The tests of mode effects are shown in the fourth and fifth columns of Table 3.

Consistent with the satisficing hypothesis, the mode effect was significant in both analyses, though a bit weaker when the Berkeley data were included ($b = .09$, $p < .01$ for the Michigan data only; $b = .07$, $p < .01$ including the Berkeley data). Furthermore, the effects of the demographic variables were largely consistent with prior research (Krosnick & Fabrigar, forthcoming). No-opinion responses were more common among respondents with less education ($b = –.30$, $p < .01$), those with lower incomes ($b = –.07$, $p < .01$), those who were older ($b = .03$, $p < .05$), those not Caucasian ($b = .05$, $p < .01$), and those who were female ($b = .01$, $p < .10$). These findings generally validate our analytic approach.

The satisficing hypothesis predicts that respondents' dispositions may interact with situational forces in determining the degree to which any given person will satisfice when answering any given question (Krosnick, 1991; Krosnick, Narayan, & Smith, 1996). That is, satisficing may be most likely when a person is disposed to do so *and* when circumstances encourage it. This logic suggests that the mode effect we observed might be strongest among respondents who were most disposed to satisfice. A great deal of research suggests that an especially powerful disposition in this regard is cognitive skills, which are very strongly correlated with years of formal education (Ceci, 1991) and can therefore be effectively measured in that way. We tested this interaction here.

Rows 2 and 3 of Table 3 display our findings when we tested this prediction by splitting the sample into respondents who had not graduated from high school and respondents with more education (for the rationale for this split, see Narayan & Krosnick, 1996). As expected, the mode effect was especially pronounced among the least-educated respondents. Looking only at the Michigan data, the average proportion of no-opinion responses increased from 34% in the face-to-face interviews to 54% on the telephone ($b = .21$, $p < .01$). The difference is smaller but nonetheless significant when the Berkeley data are folded in ($b = .15$, $p < .01$). The mode effect is much smaller in the highly educated subsample, though it is statistically significant there as well (Michigan data only: $b = .06$, $p < .01$; Michigan and Berkeley data: $b = .05$, $p < .01$).

Nondifferentiation

The second panel of results in Table 3 pertains to nondifferentiation. Here again, we see evidence consistent with the satisficing hypotheses. First, there was more nondifferentiation in the telephone samples (Michigan mean = .37, Berkeley mean = .38) than in the face-to-face sample (mean = .34). The latter rate is significantly lower than the telephone rate, whether we exclude the Berkeley data ($b = .02$, $p < .05$) or include it ($b = .03$, $p < .01$). The rate of nondifferentiation in the Michigan telephone sample was not significantly different from that in the Berkeley telephone sample ($b = .02$, n.s., $N =$

851), suggesting that all forms of satisficing were not uniformly less common in the latter.

Very little is known about the demographic correlates of nondifferentiation, other than the fact that it tends to be more common among less-educated respondents (Krosnick & Alwin, 1988; Krosnick, Narayan, & Smith, 1996; Rogers & Herzog, 1984). This trend was apparent here but was not statistically significant ($b = -.02$, $p > .10$); in fact, even the simple bivariate relation of education to nondifferentiation in the full sample was not significant ($b = -.01$, n.s., $N = 2403$). However, nondifferentiation was significantly or marginally significantly more common among respondents with lower incomes ($b = -.02$, $p < .10$), those who were younger ($b = -.04$, $p < .01$), and those who were male ($b = -.01$, $p < .10$).

When only the Michigan data were considered, the mode effect was no stronger in the least educated group ($b = .02$, n.s.) than in the more-educated group ($b = .02$, $p < .05$); (see the second and third rows of the bottom panel of Table 3). But when the Berkeley data were included, the mode effect was nearly twice as large in the least-educated group ($b = .05$, $p < .01$) as in the more educated group ($b = .03$, $p < .05$).

Mental Coin Flipping

To assess mental coin flipping, we gauged the strength of associations between variables via OLS regression. In all of them, the dependent variable was the difference between feeling thermometer ratings of Ronald Reagan and Jimmy Carter. Each regression included only a single predictor (in order to avoid problems that might be caused by multicollinearity and redundancy). The list of predictors included many well-documented strong correlates of candidate preference: party identification, job performance evaluations, perceptions of traits and emotions evoked, and assessments of personal finances and the national economy.

In the full sample, the average association was .44 for face-to-face respondents, compared with .41 for the Michigan telephone respondents and .43 for the Michigan and Berkeley telephone respondents combined. The difference between the face-to-face and telephone respondents was significant when only the Michigan telephone respondents are used ($z = 2.96$, $p = .002$) and when both the Michigan and Berkeley telephone respondents are used ($z = 1.90$, $p = .03$). Among the least-educated respondents, the mode effect was larger (as expected): the average association is .39 for the face-to-face respondents, .33 for the Michigan telephone respondents, and .36 for the Michigan and Berkeley telephone respondents ($z = 2.13$, $p = .02$, and $z = 0.90$, $p = .18$, respectively). Among more-educated respondents, the face-to-face mean association was .45, compared with .44 for both the Michigan only and the Michigan and Berkeley telephone respondents ($z = 1.36$, $p = .09$, and $z = 1.27$, $p = .10$, respectively).

Social Desirability

Two of the five tests of social desirability response bias differences by mode yielded significant results. Respondents interviewed by telephone reported higher interest in politics than respondents interviewed face-to-face ($b = .05$, $p = .01$). And Caucasians interviewed by telephone reported more support for government aid to blacks ($b = .06$, $p < .01$). Thus, telephone respondents were apparently more reluctant to report some socially undesirable interests and attitudes than were face-to-face respondents.[4]

Discussion

These analyses suggest that interview mode can affect both the sample representativeness and response patterns observed in surveys. In particular, individuals who were socially disadvantaged were undersampled in the telephone survey, relative to both the face-to-face survey and the population. Furthermore, data obtained from telephone interviews were more distorted by satisficing and by a need to appear socially desirable than were data obtained from face-to-face interviews. Individuals interviewed over the telephone showed more nondifferentiation and gave no-opinion responses more often, and these people showed an increased tendency toward socially desirable responding. These patterns are consistent with the notion that the rapport developed in face-to-face interviews inspires respondents to work harder at providing high-quality data, even when doing so means admitting something that may not be socially admirable. Given the concordance of these findings with ones from past studies, there seems to be a justifiable basis for confidence in their generalizability and validity. Furthermore, these findings suggest that there is validity to satisficing theory's claims and utility to its perspective for understanding survey phenomena.

The book is far from closed on the issue of interview mode and data quality, and the question remains an important one for survey researchers. Although telephone interviewing may be appealing to researchers because of the financial benefits, there may be significant costs associated with this method. Particularly when disenfranchising socially vulnerable members of a population is important (which it may always be), RDD telephone methods may not be worth the cost savings when budgets can permit block listing and face-to-face interviewing instead.

References

Aneshensel, C. S., Frerichs, R. R., Clark, V. A., & Yokopenic, P. A. (1982). Telephone versus in-person surveys of community health status. *American Journal of Public Health, 72,* 1017–1021.

Aquilino, W. S. (1992). Telephone versus face-to-face interviewing for household drug use surveys. *International Journal of the Addictions, 27,* 71–91.

[4]Parallel results were generated with only the Michigan respondents, except that the significant effect for interest in politics apparent in Table 5 falls to nonsignificance, likely due to the smaller sample size.

Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58,* 210–240.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27,* 703–722.

Colombotos, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly, 29,* 457–458.

De Leeuw, E. D. (1992). *Data quality in mail, telephone, and face to face surveys.* Amsterdam: TT-Publikaties.

Freeman, H. E., Kiecolt, K. J., Nicholls, W. L., & Shanks, J. M. (1982). Telephone sampling bias in surveying disability. *Public Opinion Quarterly, 46,* 392–407.

Gfroerer, J. C., & Hughes, A. L. (1991). The feasibility of collecting drug abuse data by telephone. *Public Health Reports, 106,* 384–393.

Groves, R. M. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly, 43,* 190–205.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Academic Press.

Herzog, A. R., Rogers, W. L., & Kulka, R. A. (1983). Interviewing older adults: A comparison of telephone and face-to-face modalities. *Public Opinion Quarterly, 47,* 405–418.

Jordan, L. A., Marcus, A. C., & Reeder, L. G. (1980). Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly, 44,* 210–222.

Klecka, W. R., & Tuchfarber, A. J. (1978). Random digit dialing: A comparison to personal surveys. *Public Opinion Quarterly, 42,* 105–114.

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

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52,* 526–538.

Krosnick, J. A., & Fabrigar, L. R. (forthcoming). *Designing good questionnaires: Insights from psychology.* New York: Oxford University Press.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation, 70,* 29–44.

Locander, W., & Burton, J. P. (1976). The effect of question forms on gathering income data by telephone. *Journal of Marketing Research, 13,* 189–192.

Mulry-Liggan, M. H. (1983). A comparison of a random digit dialing survey and the Current Population Survey. In *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp. 214–219). Washington, DC: American Statistical Association.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly, 60,* 58–88.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609.

Rogers, T. F. (1976). Interviews by telephone and in person: Quality of responses and field performance. *Public Opinion Quarterly, 40,* 51–65.

Rogers, W. L., & Herzog, A. R. (1984). Response style characteristics and their relationship to age and item covariances. Unpublished manuscript, Institute for Social Research, Ann Arbor, MI.

Thornberry, O. T. (1987). *An experimental comparison of telephone and personal health interview surveys.* Hyattsville, MD: U.S. Department of Health and Human Services, Public Health Service, National Center for Health Statistics.

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

Weeks, M. F., Kulka, R. A., Lessler, J. T., & Whitmore, R. W. (1983). Personal versus telephone surveys for collecting household health data at the local level. *American Journal of Public Health, 73,* 1389–1394.

Mode of Administration Considerations in the Development of Condition Specific Quality of Life Scales

Todd H. Rockwood, Robert L. Kane, and Ann Lowry

Introduction

The past few decades have seen a dramatic increase in the development and utilization of standardized condition-specific health-related quality-of-life (CSQoL) scales in health outcomes research (Kane, 1997). While general health status and health-related quality-of-life (HRQoL) scales such as the SF-36 (Ware, Snow, Kosinski, & Gandek, 1993) and Sickness Impact Profile (Bergner, Bobbit, Pollard, Martin, & Gilson, 1981) continue to be used, many researchers are turning to CSQoL instruments to assess the outcomes of treatment for particular populations and to provide a sensitive and responsive assessment of particular behaviors and/or state-trait characteristics associated with a specific health/medical condition that general health status questionnaires might not capture (Kane, 1997; Streiner & Norman, 1989). This paper discusses the importance of taking nonrandom measurement error into account in either the development or the use of HRQoL instruments in general and CSQoL instruments in particular.

The presentation of CSQoL scales in the literature is usually accompanied by traditional psychometric[1] evaluations for reliability and validity, but material on potential nonrandom measurement error factors, such as context and mode of administration, is rarely addressed. This paper will present findings from a study in which mode effects for a CSQoL instrument, the Fecal Incontinence Quality of Life (FIQL) scale (Rockwood et al., in press), were evaluated.

Information from HRQoL/CSQoL scales can potentially affect the treatment a person receives (or doesn't receive) as well as the allocation of resources; consequently, the measurement and psychometric requirements placed on these scales is necessarily high. The core question for this paper is: Are the psychometric properties of a scale developed using data collected through self-administered procedures (mail mode) retained in a similar population in which the instrument was administered using the telephone mode? As a follow-up question, we ask: If the psychometric properties are not retained, is there evidence that mode of administration effects contribute nonrandom measurement error to the measurement properties of the instrument?

Fecal Incontinence Quality of Life Scale

The purpose of the FIQL is to assess HRQoL issues related specifically to fecal incontinence. In the development of items for the FIQL, input from patients and physicians was sought using focus groups as well as cognitive interviews. After the initial item pool had been reduced to 41 items, the instrument was field-tested and then administered in a population with fecal incontinence (FI) to evaluate the psychometric properties of the FIQL. The final version of the FIQL is composed of 29 items that form four scales: Lifestyle (10 items), Coping/Behavior (9 items), Depression/Self-Perception (7 items), and Embarrassment (3 items).

Theoretical Framework

Psychometrics

The core purpose of psychometric evaluations is to establish the reliability and validity of a measure. In general, reliability assessment is intended to assess measurement error (Cronbach, 1990; Nunnally & Berstein, 1994). Historically, measurement error has been defined as: $\text{Score}_{\text{Observed}} = \text{Score}_{\text{True}} + \text{error}$; in this model error can be random or nonrandom. In reliability evaluation the goal is to evaluate the amount and impact of random measurement error (Cronbach, 1990: Nunnally & Bernstein, 1994). Alternatively, validity has focused on "how well it measures what it purports to measure" (Nunnally & Berstein, 1994, p. 83) or more generally, if the inferences drawn from the measure are valid. (Wainer, Braun, & Educational Testing Service, 1998). Assumed within the assessment of validity is the evaluation of nonrandom measurement error.

Nonrandom Measurement Error

Knowledge concerning nonrandom measurement error has increased significantly over the past three decades (Biemer,

Todd H. Rockwood is Assistant Professor, Division of Health Services Research and Policy, University of Minnesota.

Robert L. Kane is Director, Clinical Outcomes Research Center and Professor, Division of Health Services Research and Policy, University of Minnesota.

Anne Lowry is at the Department of Surgery, School of Medicine, University of Minnesota.

[1]The domain of psychometric evaluation for the purpose of this paper is constrained to refer only to standardized survey instruments.

Groves, Lyberg, Mathiowetz, & Sudman, 1991; Groves, 1989; Schuman & Presser, 1981; Schwarz & Sudman, 1992, 1996; Tanur & Social Science Research Council (U.S.) Committee on Cognition, 1992). With these advancements this knowledge can now be utilized and should be taken into account in developing and evaluating HRQoL and CSQoL scales. Research on context effects, such as question order (Schwarz, Groves, & Schuman, 1998; Smith, 1992), response categories (Krosnick & Alwin, 1987; Rockwood, Sangster, & Dillman, 1997), and recent research that has begun to model the cognitive processes involved in response formation (Sudman, Bradburn, & Schwarz, 1996; Tourangeau & Rasinski, 1998) has demonstrated that many factors can contribute to the emergence of subtle and at times not-so-subtle nonrandom measurement error. In this paper the focus will be limited to effects associated with mode of survey administration.

The relationship between mode of administration and certain types of nonrandom measurement error, such as social desirability and primacy/recency (extremeness) has been recognized for some time, but detailed studies of the ways in which mode of administration can influence context effects have not been explored historically (Dillman, Sangster, Tarnai, & Rockwood, 1996; Schwarz, Strack, Hippler, & Bishop, 1991). With the increased use of different modes of administration and the emergence of mixed-mode surveys (Dillman & Tarnai, 1988), increased attention has been paid to the impact of mode of administration on nonrandom measurement error (Bishop, Hippler, Schwarz & Strack, 1988; de Leeuw & Collins, 1997; Dillman & Tarnai, 1988; Schwarz et al., 1991).

While theory of mode effects is still developing, fundamental models have emerged. For example, Schwarz and colleagues have proposed a theoretical model arguing that self-administered surveys (mail mode) should be resistant to most types of context effects while acknowledging that mode is a primary contributor to interaction effects, such as social desirability (Schwarz et al., 1991). Others have begun to develop more conceptual models, focused on the specification of theoretically based causal pathways through which mode of administration can affect response (Dillman et al., 1996; Rockwood et al., 1997) while reserving judgment as to whether or not mode of administration has a significant role in relation to context effects.

HRQoL instruments in general, and CSQoL in particular, often ask about sensitive issues such as depression, sexual activity, drug use, and other threatening behaviors as well as about specific behaviors or events for which recall could be problematic. It is expected that effects associated with mode of administration could be present with such instruments due to either interaction effects, such as social desirability (Aquilino, 1994; Locander, Sudman, & Bradburn, 1976) or control over the pace of the interview (Dillman et al., 1996; Rockwood et al., 1997; Schwarz et al., 1991).

Study Design

Two separate studies conducted in five colon and rectal surgery clinics provide the data for this study. Four of the five clinics are located in the Midwest (Omaha, NE, Minneapolis, MN, Cleveland, OH, St. Louis, MO); the remaining clinic was located in the South (Ft. Lauderdale, FL). Any patient seeing a colon and rectal surgeon for the evaluation and/or treatment of FI was eligible for inclusion in the study. Patients were enrolled consecutively in each of the studies. The same survey was administered in each study.

The first study was a self-administered survey (drop-off/mail-back) conducted by the Clinical Outcomes Research Center at the University of Minnesota. A total of 193 patients were enrolled in this study, of which 118 returned completed questionnaires, a response rate of 61%, (Rockwood et al., in press). The purpose of this study was to collect data for the initial psychometric evaluation of the FIQL.

Immediately following the end of this study, another small study was conducted drawing upon patients from the same five clinics. In this sample the survey was done using the telephone mode. These surveys were conducted by the Survey Center, which is part of the Division of Health Services Research and Policy at the University of Minnesota. The Survey Center specializes in telephone and face-to-face interviewing for health-related surveys. As with the mail survey, consecutive patients seeing a colon and rectal surgeon for the evaluation and/or treatment of FI were selected. A total of 61 patients were enrolled in the study; of these, 47 completed the telephone survey (response rate 77%).

It is important to discuss the design limitations associated with this study. A split-mode experiment in which respondents were randomly assigned to mode of administration would have provided a stronger design (Campbell & Overman, 1988; Cook & Campbell, 1979; Kaplan, 1964). Even though the design is not as strong as it could be, given that the studies were conducted sequentially and that there were no events within each clinic, we believe that the comparison of data from these studies will, at a minimum, be suggestive.

Findings

Psychometrics

Given the data available, two primary psychometric aspects of the FIQL can be evaluated: construct validity and internal reliability. The most important of these is construct validity. As with any psychometric evaluation, construct validity, while not a sufficient basis to establish the validity of a scale, is a necessary one (Cronbach, 1990; Nunnally & Berstein, 1994).

Table 1 presents the factor loading scores for each of the items in the four FIQL scales for the telephone mode. The factor loading scores of items in the telephone mode do not conform to the established scale structures, which were developed based on data collected using the mail mode. If the structure of each of the scales as developed in the mail mode is taken as the "actual" structure, then the telephone mode does not demonstrate construct validity (since the scales do not demonstrate acceptable factor loadings, the

Table 1. Factor loadings FIQL sub-scale for telephone mode of administration

		Lifestyle	Coping	Depression	Embarrasssment
Lifestyle					
Q4d	I avoid staying overnight away from home	0.91	0.22	0.13	–0.04
Q5q	I avoid going out to eat	0.85	0.30	0.15	0.12
Q4c	I avoid visiting friends	0.87	0.31	0.11	0.28
Q4j	I avoid traveling	0.64	0.08	0.67	–0.03
Q5p	I avoid traveling by plane or train	0.62	0.16	0.21	0.33
Q4e	It is difficult for me to get out and do things like going to a movie or to church	0.82	0.25	0.47	0.06
Q4f	I cut down on how much I eat before I go out	0.46	0.17	0.71	0.26
Q4i	It is important to plan my schedule (daily activities) around my bowel pattern	0.49	0.32	0.72	0.30
Q4a	I am afraid to go out	0.65	0.51	0.41	0.08
Q5c	I cannot do many of the things I want to do	0.60	0.63	0.11	0.39
Coping					
Q4g	Whenever I am away from home, I try to stay near a restroom as much as possible	0.34	0.65	0.46	0.33
Q4r	I can't hold my bowel movement long enough to get to the bathroom	0.18	0.71	0.34	0.09
Q4t	I try to prevent bowel accidents by staying very near a bathroom	0.08	0.54	0.76	0.22
Q5e	I worry about bowel accidents	0.30	0.82	0.39	0.17
Q4k	I worry about not being able to get to the toilet in time	0.37	0.59	0.68	0.10
Q5r	Whenever I go someplace new, I specifically locate where the bathrooms are	0.32	0.61	0.15	0.22
Q4l	I feel I have no control over my bowels	0.31	0.79	0.30	0.26
Q5n	The possibility of bowel accidents is always on my mind	0.07	0.75	0.42	0.21
Q5l	I have sex less often than I would like to	0.26	0.88	0.08	0.18
Depression					
Q6	During the past month, I have felt so sad, discouraged, hopeless, or had so many problems that I wondered if anything was worthwhile	0.12	0.30	0.30	0.65
Q5f	I feel depressed	0.40	0.52	0.17	0.61
Q5j	I feel like I am not a healthy person	0.01	0.40	0.20	0.80
Q5k	I enjoy life less	0.40	0.82	0.22	0.16
Q5m	I feel different from other people	0.08	0.74	0.56	–0.09
Q5o	I am afraid to have sex	0.17	0.81	0.22	0.37
Ql	In general, would you say your health is	–0.15	0.00	0.03	–0.93
Embarrassment					
Q5b	I leak stool without even knowing it	0.21	0.68	0.05	0.53
Q5g	I worry about others smelling stool on me	0.42	0.81	0.05	0.15
Q4s	I feel ashamed	0.22	0.64	0.64	0.15

internal validity analysis Cronbach's alpha will not be presented).

The factor structure of any given scale will differ across administrations and populations, and the psychometric properties of any instrument are to some extent an artifact of the population it was developed in. For example, we looked at the factor loadings of the SF-36 in the last five studies we used it in; not once did the SF-36 meet the traditional criteria for factor loadings (+.60 on scale, >.30 on other scales), and in every instance some of the items jumped scales. While it is plausible that there are major psychological or physiological differences between the mail and telephone study groups, it is not likely; the primary difference between the two populations is mode of survey administration.

To exactly reproduce the factor structure of any given scale in all studies (even in similar populations) is not a reasonable expectation. But, is it safe to assume that the core structure of the scales should remain constant, unless nonrandom measurement error has a significant impact on the measurement properties of the items. The differences shown in Table 1 suggest that nonrandom measurement error is influencing the psychometric properties of the FIQL.

Nonrandom Measurement Error

While the individual items in the survey are not complex, many of them are of a highly personal nature. While many "fecal"-related words and phrases are a part of popular culture, normal social discourse about "fecal incontinence" as a personal problem is not something that one is likely to hear. Given this, it is expected that effects such as social desirability could influence responses in the telephone mode. Only 4 of the 21 items in the instrument demonstrated significant differences between the mail and telephone modes (see Table 2, Bonferroni-adjusted *t*-test of means). Many items that would usually be considered prime suspects for social desirability effects, such as "I have sex less often than I would like to," did not demonstrate significant differences. For the four items with significant differences (Q1, Q4s, Q5g, and Q5n), respondents in the mail mode tended to provide more "negative" responses, a distribution consistent with a social desirability expectation. Although Q1 does present something of a quandary as to why mode differences were found, the other three items are not as perplexing. What is encouraging is that, given the highly personal nature of many of the items in the instrument, more mode effects were not found.

Table 3 shows that two of the four scales found in the FIQL demonstrate significant differences between the mail and telephone modes of administration. In the mail mode the mean for the depression scale is 2.8 (std. dev. 81), and in the telephone mode it is 3.2 (std. dev. 89, $p < .01$). One of the seven items in this scale also demonstrated significant differences (Q1, health in general). The other scale showing significant differences is the embarrassment scale (mail mean/std. dev.: 2.0/.84, telephone mean/std. dev.: 2.8/1.1, $p < .01$). Two of the three items in this scale demonstrated significant differences: Q4s (leak stool without knowing it) and Q5g (worry about others smelling stool). The distribution of responses for all scales demonstrate a pattern that is consistent with a social desirability expectation, in which respondents in the telephone mode are more likely to report a higher quality of life.

The final aspects of measurement to be considered are floor and ceiling effects. In considering nonrandom measurement error relative to floor and ceiling effects, attention is immediately drawn to satisficing, primacy/recency, and extremeness effects (Dillman & Tarnai, 1991, 1992; Krosnick & Alwin, 1987; Rockwood et al., 1997; Schuman & Presser, 1981; Schwarz, Hippler, & Bishop, 1991a, 1991b). Given the structure of the questions and response categories used in this survey, floor effects would be associated with primacy and ceiling effects with recency. Table 4 presents the percentage of respondents demonstrating floor and ceiling effects for each of the scales in each mode of administration. Consistent with theory, the mail mode demonstrates marginally more primacy (floor) effects. A large number of recency (ceiling) effects were found in the telephone mode compared with the

Table 2. Individual questions demonstrating significant differences between the mail and telephone modes of administration (significance test is a Bonferroni-adjusted *t*-test of means)

	Mail		Telephone		
Question	***N***	**Mean/Std. Dev.**	***N***	**Mean/Std. Dev.**	**Signif.**
Q1. In general, would you say your health is	117	3.3/.93	47	2.6/1.2	.001
Excellent	3%		19%		
Very good	14%		34%		
Good	41%		28%		
Fair	32%		9%		
Poor	10%		11%		
Q4s. I leak stool without even knowing it	118	2.2/.98	46	3.0/1.2	.001
Most of the time	29%		15%		
Some of the time	40%		20%		
A little of the time	19%		13%		
None of the time	13%		52%		
Q5g. I worry about others smelling stool on me	94	2.0/1.1	47	2.8/1.3	.001
Strongly agree	49%		28%		
Somewhat agree	24%		17%		
Somewhat disagree	9%		6%		
Strongly disagree	18%		49%		
Q5n. The possibility of bowel accidents is always on my mind	117	1.8/1.0	46	2.4/1.3	.05
Strongly agree	55%		37%		
Somewhat agree	26%		20%		
Somewhat disagree	9%		11%		
Strongly disagree	11%		33%		

Table 3. Mean scale score comparison between telephone and mail modes of administration (significance test is a Bonferroni-adjusted *t*-test of means; scale range is 1–4)

	Mail			Telephone			
Scale	***N***	**Mean**	**Std. Dev.**	***N***	**Mean**	**Std. Dev.**	**Signif.**
Lifestyle	110	2.8	.94	46	3.0	.92	ns
Coping	118	2.1	.87	46	2.5	1.1	ns
Depression	112	2.8	.81	46	3.2	.89	.01
Embarrassment	116	2.0	.84	47	2.8	1.1	.01

Table 4. Percent of items demonstrating floor and ceiling effects (minimum and maximum scores) by mode of administration

	Floor		Ceiling	
Scale	**Mail**	**Telephone**	**Mail**	**Telephone**
Lifestyle	2%	2%	12%	28%
Coping	7	4	5	22
Depression	1	0	1	0
Embarrassment	16	13	7	2
	Fischer's exact (two-tail) $p = .94$		Fischer's exact (two-tail) $p = .24$	

mail mode for all scales except depression. While the distribution is not significantly different based on Fischer's exact test, the differences are cause for concern relative to measurement and the use of the data for analytic purposes.

Even though this research did not use a pure experimental design, the findings are indicative that nonrandom measurement error is a real concern relative to the measurement properties of some of the items in the FIQL and two of the four scales in the instrument. The ceiling effects found in the telephone mode are also indicative of nonrandom measurement error influencing the measurement process through the introduction of recency effects. Overall, the findings point to a conclusion that nonrandom measurement error is influencing measurement. While the direct question of the influence of this nonrandom error on the scales' psychometric properties cannot be addressed given this study design, it is plausible to expect that the psychometric properties of the FIQL, especially validity, might be influenced by this nonrandom error.

Conclusions

The message from this paper differs between the "user" of HRQoL and CSQoL instruments and the "developers" of such scales. For people who utilize HRQoL or CSQoL instruments in their work, this paper offers the caution that mode of administration and possibly other sources of nonrandom measurement error (Groves, 1989) can affect the measurement properties of the items in a scale as well as the scales themselves and the inferences based upon said measures. For the user, the consideration of what is "valid" and "reliable" needs to extend beyond a consideration of alphas and factor loading scores and must incorporate the wealth of knowledge that has come from the study of measurement error in survey research (Biemer et al., 1991; Groves, 1989; Schwarz & Sudman, 1992; Tanur & Social Science Research Council (U.S.) Committee on Cognition, 1992).

For those who work in the area of developing HRQoL and CSQoL instruments, the message is that while an instrument might demonstrate acceptable validity based upon psychometric techniques, this does not mean that nonrandom measurement error has been eliminated. For developers, the consideration of factors other than those traditionally assessed by psychometrics must become a concern, especially in relation to measurement error in survey research (Groves, 1989).

The knowledge that has been accumulated on measurement error in survey research (Biemer et al., 1991; Groves, 1989; Schwarz & Presser, 1981; Schwarz & Sudman, 1992; Tanur & Social Science Research Council (U.S.) Committee on Cognition, 1992) is essential to both the use and development of HRQoL or CSQoL scales. Even the "bible" of psychometrics, Nunally and Bernstein's *Psychometric Theory,* does not have a single reference to classic works on measurement error in survey research, such as *Questions and Answers in Survey Research* (Schuman & Presser, 1981) or *Survey Errors and Survey Costs* (Groves, 1989). Within psychometrics, there seems to be an underlying assumption that the techniques used to evaluate validity will ferret out sources of nonrandom measurement error. The findings from this study point to the need to formally introduce the consideration of nonrandom measurement error into the evaluation of psychometric properties of HRQoL scales.

References

Aquilino, W. S. (1994). Interview mode effects in drug surveys. *Public Opinion Quarterly, 58* (2), 210–240.

Bergner, M., Bobbitt, R., Pollard, W., Martin, D., & Gilson, B. (1981). Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care, 19*, 787–805.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (1991). *Measurement errors in surveys*. New York: Wiley.

Bishop, G. F., Hippler, H.-J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. I. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 321–340). New York: Wiley.

Campbell, D. T., & Overman, E. S. (1988). *Methodology and epistemology for social science: Selected papers*. Chicago: University of Chicago Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally College Publishing Co.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

de Leeuw, E., & Collins, M. (1997). Data collection methods and survey quality: An overview. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 199–220). New York: Wiley.

Dillman, D. A., Sangster, R. L., Tarnai, J., & Rockwood, T. H. (1996). Understanding differences in people's answers to telephone and mail surveys. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research* (vol. 70, pp. 110). San Francisco: Jossey-Bass.

Dillman, D. A., & Tarnai, J. (1988). Administrative issues in mixed mode surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. I. Nicholls, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 509–528). New York: Wiley.

Dillman, D. A., & Tarnai, J. (1991). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 73–93). New York: Wiley.

——— . (1992). Questionnaire context as a source of response differences in mail and telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 115–130). New York: Springer-Verlag. .

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Kane, R. L. (1997). *Understanding health care outcomes research*. Gaithersburg, MD: Aspen Publishers.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. San Francisco: Chandler Publishing Co.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51* (Summer), 201–219.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method threat and response distortion. *Journal of the American Statistical Association, 71* (354), 269–275.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rockwood, T. H., Church, J. M., Fleshman, J. W., Kane, R. L., Mavrantonis, C., Thorson, A. G., Wexner, S. D., Bliss, D., & Lowry, A. C. (in press). FIQL: A quality of life instrument for patients with fecal incontinence. *Diseases of the Colon & Rectum*.

Rockwood, T. H., Sangster, R. L., & Dillman, D. A. (1997). The effect of response categories on questionnaire answers: Context and mode effects. *Sociological Methods and Research, 26* (1), 118–140.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Academic Press.

Schwarz, N., Groves, R. M., & Schuman, H. (1998). Survey methods. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of social psychology* (4th ed., vol. 1, pp. 143–179). Boston/ New York: McGraw-Hill; distributed exclusively by Oxford University Press.

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991a). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly 55* (Spring), 3–23.

———. (1991b). The impact of administration mode on response effects in survey research. *Applied Cognitive Psychology, 5*, 193–212.

Schwarz, N., & Sudman, S. (1992). *Context effects in social and psychological research*. New York: Springer-Verlag.

Schwarz, N., & Sudman, S. (1996). *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (1st ed.). San Francisco: Jossey-Bass.

Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 163–186). New York: Springer-Verlag.

Streiner, D. L., & Norman, G. R. (1989). *Health measurement scales: A practical guide to their development and use*. Oxford (England) New York: Oxford University Press.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology* (1st ed.). San Francisco: Jossey-Bass.

Tanur, J. M., & Social Science Research Council (U.S.) Committee on Cognition. (1992). *Questions about questions: Inquiries into the cognitive bases of surveys*. New York: Russell Sage Foundation.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103* (May), 299–314.

Wainer, H., Braun, H. I., & Educational Testing Service. (1988). *Test validity.* Hillsdale, NJ: L. Erlbaum Associates.

Ware, J., Snow, K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey: Manual and interpretation guide*. Boston: Health Institute, New England Medical Center.

Mode Differences in Reports of Alcohol Consumption and Alcohol-Related Harm

Lorraine T. Midanik, John D. Rogers, and Thomas K. Greenfield

Introduction

Researchers who conduct national surveys must balance their needs for reliable and valid data from a large sample against increasing costs for collecting data using face-to-face interviews (Gfroerer & Hughes, 1991, 1992). While large surveys have for many years been conducted using face-to-face and telephone questionnaires, it is not clear how much of an effect mode of administration has on the quality of survey data and, in particular, on estimates of alcohol consumption and problems.

For alcohol use, comparisons of telephone and face-to-face interviews have yielded some findings of lower reports obtained by telephone (Aquilino, 1992, 1994), but in other cases minimal or no differences were found (Aquilino & Wright, 1996; de Leeuw & van der Zouwen, 1988; McAuliffe, Geller, LaBrie, Paletz, & Fournier, 1998). In contrast, Hochstim (1962) and Sykes and Collins (1988) found higher rates of alcohol use using the telephone than through in-person interviewing. Most of these studies have relied on dichotomous drinking measures and have not focused on level or pattern of drinking. Even less is known about how mode may affect reports of alcohol-related harm. Sampling differences as well as social desirability and interviewer effects may all contribute to differential reporting of alcohol-related harm in alcohol surveys.

From a research standpoint, differences by mode are important and should lead researchers to pursue further studies to reconcile these conflicting results. Added to these research concerns is the reality of cost involved in conducting large-scale national population surveys. The current U.S. National Alcohol Survey (NAS), which is part of a large-scale survey series undertaken at approximately 5-year intervals, is converting from a face-to-face to a telephone methodology this year. Thus, to continue work on trends using this cross-sectional series by comparing the 1999 NAS with earlier national surveys (Greenfield, Midanik, & Rogers, in press; Midanik & Greenfield, in press), it is necessary to understand potential biases in mode of administration and attempt to adjust for these biases in comparing prevalence rates of alcohol use and alcohol-related problems in planned analysis.

In two *between-subjects* analyses, assessment of prevalence rates of alcohol consumption measures and alcohol-related harm from two national surveys in 1990 with different modes of administration (face-to-face versus telephone interviews) yielded mixed findings. Few differences were found in alcohol consumption prevalence estimates (last 12 months) by interview mode. However, the findings indicate that lower-income individuals were less represented in the telephone sample, which suggests that weights should be applied to adjust for this discrepancy (Greenfield et al., in press-a) .

Higher rates of alcohol-related *health* harm and *work* harm in the last 12 months were found in the telephone survey as compared with the face-to-face study after controlling for demographic characteristics (gender, age, ethnicity, income, and education) and number of heavy drinking days in the past year (Midanik, Greenfield, & Rogers, 1996). There are several possible explanations for these findings. First, with telephone interviewing, anonymity may be increased and social desirability factors may be decreased. It has been argued that telephone interviews may provide more accurate responses to sensitive items due to anonymity (Schwarz, Strack, Hippler, & Bishop, 1991). However, it is not clear that measures of health and work harm due to alcohol use are any more sensitive than other measures of potential alcohol-related harm, such as friendships/social life, home life or marriage, or financial position, which showed no mode differences. A second explanation could be that cognitive processes may differ by mode (Sudman, Bradburn, & Schwarz, 1996). Thus, the demands placed on respondents to retrieve information from their memory relatively rapidly and comprehension problems that may occur might differ for respondents on the telephone as opposed to those in face-to-face interviews. A recent study using protocol analysis to obtain drinking data both by telephone and face-to-face interviews from 30 heavier drinkers explored these issues further. The findings indicate no differences by mode (Midanik, Hines, Greenfield, & Rogers, in press). An alternative explanation for the mode differences in reports of alcohol-related harm is the placement of items in each questionnaire. In the face-to-face survey, the harm items were

Lorrraine T. Midanik is a professor at the at the University of California at Berkeley and an affiliate senior scientist at the Alcohol Research Group, Berkeley, California.

John D. Rogers and Thomas K. Greenfield are at the Alcohol Research Group, Berkeley, California.

This research was supported by a National Alcohol Research Center Grant (AA-05595-16) from the U.S. National Institute on Alcohol Abuse and Alcoholism to the Alcohol Research Group, Public Health Institute, Berkeley, California.

placed after a long list of alcohol-related problem items, which may have enhanced respondent fatigue. Further, respondents may have "learned" that if they answered "yes" to an "ever" harm question, they would automatically be asked an additional question pertaining to the last 12 months. Thus, they could avoid additional questions by answering "no" initially to the harm item (Millwood & Mackay, 1978). In the telephone survey, the harm items were asked immediately after the alcohol consumption items; thus, there was less opportunity for respondent fatigue.

While the *between-subjects design* assesses one dimension of telephone versus face-to-face interviewing modes, in which mode is confounded by sampling design differences (random-digit dialing reaches only those with residential telephones), a *within-subjects* design addresses whether mode differences influence responses with respondents serving as their own controls. Here the underlying sampling design is held constant. The purpose of this paper is to compare prevalence rates of alcohol consumption and alcohol-related harm measures for respondents who participated in both face-to-face and subsequent telephone interviews as part of our 1995 National Alcohol Survey (NAS) Telephone Follow-up study.

Methods

Study Population

Data for this study were obtained from a subsample of the National Alcohol Survey (NAS) conducted in 1995 (4,925 face-to-face interviews total) who were given a telephone follow-up (1,047 respondents of 1,348 contacted, a 78% response rate). A case was considered eligible for the telephone follow-up if the original interview was completed in English between September 1, 1995, and April 30, 1996, and a confirmed home telephone number was provided by the respondent at the time of the original interview. Table 1 compares demographic, alcohol use, and alcohol-related harm variables for the original 1995 NAS with the telephone sample. There were no significant differences between those included or not included in the follow-up sample in terms of age, alcohol use (mean daily volume and mean days of heavier, 5+, drinking), or reports of any harm in the last 12 months. However, the respondents not included in the telephone follow-up were more likely to be male, to be African American, and to have lower educational levels. Respondents in the telephone follow-up sample were more likely to have income levels below the median.

Originally, the telephone follow-up was designed to be a six-week follow-up study involving substantial numbers of Caucasian, African American, and Latino respondents. However, because data collection began later than planned, two steps were taken to meet the objective of 1,000 telephone follow-up cases. First, a Spanish-language version of the follow-up was developed to avoid selectivity and make available more Latino respondents. Second, the time interval between surveys was extended to ensure an adequate sample size. The mean interval between surveys was 17.3 weeks (S.D. = 8.2, range = 4.3 to 40.7 weeks) with 75% of the cases less than 5 months after the initial face-to-face interview. While this is a longer time interval than originally planned, the drinking items used in this study are based on last-12-month time frames, which should be relatively stable over time. Also, the majority of the cases have overlapping time frames of approximately 7–8 months, making the reporting period fairly similar.

Table 1. Characteristics of 1995 National Alcohol Survey respondents included and not included in the telephone follow-up sample

	Follow-up Sample	Not in Follow-up
N	1047	3878
Gender*: % male	44.0	48.9
Age		
18–29	25.0	23.5
30–49	44.0	46.9
50+	31.0	29.6
Ethnicity**		
White	33.0	34.1
African American	30.7	37.3
Latino	34.0	25.4
Other	2.3	3.2
Income**: % below median	67.5	59.5
Education**: % HS or less	34.1	44.6
Reported: % any harm	6.8	6.1
Mean (SD) daily volume†	.94 (1.8)	.93 (1.9)
Mean (SD) days 5+†	24.0 (66.3)	22.9 (65.3)

*$p < .01$
**$p < .001$
†Drinkers only at Time 1 (face-to-face) interview: $n = 622$ in follow-up sample; $n = 2{,}195$ in non-follow-up sample. Means based on raw data; *t*-tests based on logged data.

Measures

Alcohol consumption was measured in two ways for the last 12 months: (1) average daily volume of alcohol consumed and (2) heavy drinking (number of days in which 5 or more drinks were consumed per day). Both measures were developed using a graduated frequencies (GF) approach (Hilton, 1989). The GF series begins with a maximum-quantity question (largest quantity of alcohol consumed in a single day during the last 12 months) and then asks respondents at each corresponding level of alcohol use (e.g., 12 or more drinks, 8–11 drinks, 5–7 drinks, 3–4 drinks, 1–2 drinks), the frequency for which they drank alcohol at that level. The algorithms used for constructing drinking volume (expressed as average number of drinks per day) and (heavy drinking number of days in the prior 12 months during which an individual consumed 5 or more drinks) were identical for the NAS and the telephone follow-up survey (see Rogers & Greenfield,

1999). The frequency bands were first converted to days (in the last 12 months), using the midpoints of the amount ranges in drinks, except for the 12+ drinks category, which was coded conservatively as 13. Logarithmic (base 10) transformations were applied to the volume and frequency of 5+ drinks to reduce skewness and better approximate normality in each distribution (Greenfield, et al., in press).

Alcohol-related harm was measured identically in both surveys by six separate items and one composite (any harm) index (Midanik, et al., 1996). The harm items were worded as follows: ". . . was there ever a time when you felt *your* drinking had a *harmful* effect on: (1) your friendships and social life? (2) your outlook on life? (3) your health? (4) your homelife or marriage? (5) your work and employment opportunities? (6) your financial position?" If respondents indicated that an alcohol-related harm occurred during their lifetime and they were current drinkers, they were then asked if it had occurred during the last 12 months. Only the last-12-months harm items are used in this analysis.

Analysis

Differences between consumption estimates and prevalence rates of alcohol-related harm were assessed using McNemar chi-square tests and paired *t*-tests. Logarithmic (base 10) transformations were applied to mean daily volume and number of days 5+ to reduce skewness. Logistic regression models were also developed to determine if demographic variables or length of time between interviews was significantly related to inconsistent reporting.

Table 2. Odds ratios (95% confidence limits) of reporting an inconsistent drinking status

	Any Inconsistent Drinking Status	Drinker to Nondrinker Status	Nondrinker to Drinker Status
N	189	34	155
Male	.90	.86	.92
	(.64, 1.27)	(.42, 1.77)	(.63, 1.34)
Age	1.01	.99	1.01
	(.99, 1.01)	(.96, 1.01)	(.99, 1.02)
Ethnicity†			
African American	1.41	.67	1.70*
	(.92, 1.02)	(.28, 1.60)	(1.06, 2.75)
Latino	1.64*	.92	1.88*
	(1.03, 2.61)	(.37, 2.28)	(1.13, 3.15)
Other	1.42	.80	1.65
	(.50, 3.98)	(.10, 6.58)	(.53, 5.14)
Lower income	1.55**	3.43**	1.23
	(1.05, 2.29)	(1.34, 8.77)	(.81, 1.88)
Lower education	1.23	.88	1.33
	(.85, 1.76)	(.42, 1.83)	(.89, 1.99)
Weeks between interviews	.99	1.01	.98
	(.97, 1.01)	(.96, 1.05)	(.96, 1.01)

*$p < 0.05$
**$p < 0.01$
†Caucasian as the reference group.

Results

Alcohol Use

Of the 1047 respondents who participated in both surveys, 25.9% (n = 271) reported no drinking during the past year at both interviews, 56.1% (n = 587) remained current drinkers at both times, and 18.1% (n = 189) shifted their drinking status between interviews. Of those who changed their drinking status, 82% (n = 155) were originally categorized as nondrinkers, but reported current drinking at follow-up, while the remainder (n = 34) reported alcohol use in the last year during their initial interview and then no drinking at the telephone follow-up. For 96% of respondents who changed their drinking status, daily volume (either at initial interview or follow-up) was less than one drink per day.

To assess the effect of length of time between interviews and demographic variables on inconsistency of self-reported drinking status, three logistic regression models were developed using the following three dependent variables: (a) any (bidirectional) inconsistent reporting of alcohol use versus consistent reporting; (b) respondents who changed their drinking status from drinker to nondrinker versus all other respondents, and (c) respondents who changed their drinking status from nondrinker to drinker versus all other respondents (Table 2). Inconsistent drinking status between interviews was significantly associated with being Latino (as compared with Caucasian) and having a lower income. Lower income was also significantly related to changing one's status from a current drinker during the face-to-face interview to a nondrinker at the telephone interview. Being African American or Latino was significantly associated with changing one's drinking status from nondrinker to drinker. For all three models, length of time between interviews was not significant.

Based on data from respondents who were current drinkers at both periods of data collection (n = 587), there were no significant differences by mode for the number of drinks per day reported during the last year (raw mean daily volume from face-to-face interview = .94 drinks [sd =1.75] versus 1.03 drinks [sd = 2.07] from telephone interview, t = 1.91, ns). A significantly higher mean number of heavier (5+) drinking days was found for the initial, face-to-face interview (raw mean 5+ days from face-to-face interview = 24.10 days [sd = 66.32] versus 27.75 days [sd = 77.93] from the telephone interview, t = 3.59, p < .001). However, when the comparisons are calculated using only those respondents who reported having five or more drinks at least once during the last year (n = 177), there is no difference between interviews (raw mean 5+ days from face-to-face interview = 66.53 days [sd = 98.61] versus 81.54 days [sd = 117.49] from the telephone interview, t = 1.89, ns).

Table 3. Percentage of current drinkers (N = 587) at both interviews who reported any 5+ days, monthly 5+ drinking, and weekly 5+ drinking

	Face-to-Face					
	Any 5+ Days		Monthly 5+		Weekly 5+	
Telephone	No	Yes	No	Yes	No	Yes
Any 5+ days						
No	49.6	15.5				
Yes	6.3	28.6				
χ^2	22.0*					
Monthly 5+						
No			72.2	9.0		
Yes			6.5	12.3		
χ^2			2.2			
Weekly 5+						
No					84.2	4.6
Yes					5.3	5.9
χ^2					.2	

*$p < .0001$

Another way to assess differences in the number of 5+ days reported in the interviews is to compare categorized frequencies of 5+ drinking (any 5+ days, 5+ on a monthly basis, 5+ on a weekly basis) for current drinkers at both interviews (Table 3). A significantly larger proportion of respondents reported at least one day of 5+ drinks in the face-to-face interview compared with the telephone follow-up (15.5% versus 6.3%); however, there were no differences in reports of monthly 5+ days or weekly 5+ days. Logistic regression models were run to determine if any demographic variables or the length of time between interviews was associated with inconsistent reporting of any 5+ days (data not shown). In all three models (any inconsistent reporting of 5+ drinking days, reported 5+ days initially and not at follow-up; 5+ drinking days not reported initially but reported at follow-up), length of time between interviews was not associated with any of the dependent variables. Only one demographic variable, lower education (high school graduate or less), was significantly related to any inconsistent reporting of a 5+ day.

Alcohol-Related Harm. There were no significant differences between the two surveys for the rates of reporting alcohol-related harm during the last 12 months. For the face-to-face interview versus the telephone survey, the percentage of current drinkers reporting alcohol-related harm is as follows: social harm (4.4 versus 4.8), outlook on life (4.3 versus 6.2), home life or marriage (5.0 versus 3.9), financial (4.1 versus 3.2), work (2.4 versus 1.9). health (5.0 versus 5.3), and any harm (10.7 versus 10.6).

Discussion

In summary, this within-subjects design yielded data that indicate no differences between the face-to-face and telephone interviews for mean daily volume or for reports of alcohol-related harm for current drinkers at both interviews. The mean number of days in which respondents drank five or more drinks was significantly higher in the initial face-to-face interview, but was not significant when only those who reported at least one 5+ day were included in the analysis. Reports of a weekly or monthly pattern of heavier episodic drinking (5+) did not differ by survey, but the prevalence of reporting any 5+ drinking day in the prior 12 months was significantly higher for the initial, face-to-face interview.

One potentially disturbing finding in this study is the shift of drinker/nondrinker status that occurred between the face-to-face and telephone interviews. A large proportion of those respondents who were classified as nondrinkers at the initial interview reported that they consumed less than once in the last year. This does not necessarily imply that the respondent did not drink during the last 12 months, because the time period is not explicitly specified in the original frequency question. That is, if one's frequency of drinking is less than yearly, one might, in principle, consume alcohol in any given year, or even the present year. Nonetheless, there could be variation over time. Thus, for our next national survey we will be including an item that specifically asks infrequent drinkers when they had their last drink. This additional item may improve our categorization of drinking status (current drinker versus nondrinker) early on in the survey, which, in turn, determines who will be asked more questions about current alcohol use and alcohol-related problems.

Interestingly, unlike previous research findings that African American and Latino respondents report lower frequencies of drunkenness (Aquilino, 1994) and that African Americans report lower estimates of alcohol use (Aquilino and LoSciuto, 1990) on telephone interviews compared with in-person interviews, in the present study African Americans and Latino respondents were more likely to report being a current drinker on the telephone survey compared with the face-to-face interview.

Overall, as we make the transition to a telephone methodology for the next National Alcohol Survey, the lack of differences by mode in both the between-subjects analyses (Greenfield et al., in press; Midanik et al., 1996) and within-subjects analysis presented here provides additional support for continuing trend studies with adjustments to account for sampling and other biases.

References

Aquilino, W. (1992). Telephone versus face-to-face interviewing for household drug use surveys. *International Journal of Addiction, 27,* 71–91.

Aquilino, W. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58,* 210–240.

Aquilino, W., & LoSciuto, L. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly, 54,* 362–395.

Aquilino, W., & Wright, D. (1996). Substance use estimates from RDD and area probability samples: Impact of differential screening methods and unit nonresponse. *Public Opinion Quarterly, 60,* 563–573.

de Leeuw, E., & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative metaanalysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–299). New York: Wiley.

Gfroerer J., & Hughes, A. (1991). The feasibility of collecting drug abuse data by telephone. *Public Health Reports, 106,* 384–393.

Gfroerer, J., & Hughes, A. (1992). Collecting data on illicit drug use data by phone. In C. Turner, J. Lessler, & J. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 277–295). DHHS Publication No. (ADM) 92–1929 U.S. Dept. of HHS, Public Health Service, ADAMHA.

Greenfield, T., Midanik, L., & Rogers, J. (in press-a). Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*.

Greenfield, T., Midanik, L., & Rogers, J. (in press-b). A ten-year national trend study of alcohol consumption: Is the period of declining drinking over? *American Journal of Public Health*.

Hilton, M. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction, 84,* 1085–1092.

Hochstim, J. R. (1962). Comparison of three information-gathering strategies in a population study of sociomedical variables. *Proceedings of the Social Statistics Section, American Statistical Association*, 154–159.

McAuliffe, W., Geller, S., LaBrie, R., Paletz, S., & Fournier, E. (1998). Are telephone surveys suitable for studying substance abuse epidemiology? Cost, administration, coverage and response rate issues. *Journal of Drug Issues, 28,* 455–482.

Midanik, L., & Greenfield, T. (in press). Trends in social consequences and dependence symptoms in the United States: The National Alcohol Surveys, 1984–1995. *American Journal of Public Health*.

Midanik, L., Greenfield, T., & Rogers J. (1996). Reports of alcohol-related harm by heavier drinkers: A comparison of telephone versus face-to-face interviews. Paper presented at the 124th Annual Meeting of the American Public Health Association, New York, November.

Midanik, L., Hines, A., Greenfield, T., & Rogers, J. (in press). Face-to-face versus telephone interviews: Using cognitive methods to assess alcohol survey questions. *Contemporary Drug Problems*.

Millwood, J., & Mackay, A. (1978). Measurement of alcohol consumption in the Australian population. *Community Health Studies, II,* 123–133.

Rogers, J., & Greenfield, T. (1999). Beer drinking accounts for most of the hazardous alcohol consumption reported in the U.S. *Journal of Studies on Alcohol, 60,* 732–739.

Schwarz, N., Strack, F., Hippler, H., & Bishop, G. (1991). The impact of administration mode on responses effects in survey measurement. *Applied Cognitive Psychology, 5,* 193–212.

Sudman, S, Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers. The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Sykes, W., & Collins, M. (1988). Effects of mode of interview: Experiments in the UK. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 301–320), New York: Wiley.

Obtaining HIV Test Results with a Home Collection Test Kit in a Community Telephone Sample

Dennis H. Osmond, Joseph Catania, Lance Pollack, Jesse Canchola, Deborah Jaffe, Duncan MacKellar, Linda Valleroy

Introduction

Monitoring the AIDS epidemic by obtaining estimates of the present burden of disease (seroprevalence) and the rate of new infections (seroincidence) is an important public health goal that has been difficult to achieve because population-based samples of persons at high risk for infection are costly and difficult to obtain. Results from venue-based samples of these populations are difficult to interpret. HIV seroprevalence estimates that rely on population-based samples of the entire U.S. population, such as the third National Health and Nutrition Examination Survey (NHANES III), underrepresent populations at high risk. The NHANES III estimate of HIV seroprevalence in the United States, obtained from antibody tests of the sample, was 461,000 (McQuillan, Khare, Karon, Schable, & Vlahov, 1997). The researchers acknowledged substantial underrepresentation of men who have sex with men (MSM) and injecting drug users (IDU) and estimated that an additional 190,000 infected persons should be added to the estimate, a 41% upward adjustment that had to rely on other data sources.

In May 1996, the Food and Drug Administration (FDA) licensed HIV home collection kits, and many states moved quickly to remove legal barriers to their use. Prior to their licensure, the use of such kits moved from general opposition by most AIDS prevention experts several years ago to acceptance and active encouragement in recent years (Bayer, Stryker, & Smith, 1995). The FDA-approved home collection kit is based on a blood sample obtained by fingerstick. The Orasure HIV sample collection device (Epitope, Portland, Oregon) is FDA-licensed for collecting a specimen of oral fluids in medical settings for HIV testing. Although it is not licensed for a home collection kit, it has been used in the home collection format in research settings, such as with the NIH-sponsored HIV vaccine trial cohorts.

Effective methods of obtaining population-based samples with biological specimens would be a useful addition to methods of monitoring the AIDS epidemic as well as other infectious diseases. The Urban Men's Health Study (UMHS) provided a population-based telephone sample of MSM for testing the acceptability of HIV home collection kits as a method of obtaining HIV test results (Binson et al., 1996). We conducted a study that offered an HIV test using an oral home collection kit to a subsample of MSM previously interviewed in the UMHS. We are not aware of a previous study using HIV home testing with telephone sampling. The specific aims of the study were:

1. To determine what proportion of MSM contacted by telephone will use an oral-based HIV home collection kit to be tested for HIV
2. To investigate factors associated with willingness to participate in HIV testing using home collection
3. To assess technical problems of using an oral specimen device for a home collection kit with a phone sample

Methods

Sampling MSM: The Urban Men's Health Study

The UMHS conducted a telephone survey of MSM from four urban centers that represent "gay meccas" in the United States: San Francisco, Los Angeles, New York, and Chicago. In these four cities, the proportion of households with telephones is approximately 95% (Anderson, Nelson, & Wilson, 1998; U.S. Bureau of the Census, 1997). The qualitative and quantitative aspects of the current sample design are described in a series of papers (Binson et al., 1996; Blair, 1999; Mills et al., 1998). Preliminary work identified moderate- to high-density MSM geographic areas (zip codes) within each city through mapping of MSM-relevant health, commercial, and

Dennis H. Osmond is at the University of California, San Francisco, Department of Epidemiology and Biostatistics, and the AIDS Research Institute and Center for AIDS Prevention Studies, University of California, San Francisco.

Joseph Catania, Lance Pollack, and Jesse Canchola are at the AIDS Research Institute and Center for AIDS Prevention Studies, University of California, San Francisco, and the University of California, San Francisco, Department of Medicine.

Deborah Jaffe is at the University of California, San Francisco, Department of Medicine.

Duncan MacKellar and Linda Valleroy are at the Centers for Disease Control and Prevention.

Primary support for the Urban Men's Health Study was provided by NIMH Grant No. MH54320, and supplemental support for this study was provided by the Centers for Disease Control and Prevention's Division(s) of HIV/AIDS Prevention through cooperative agreement TS235-12/12 from the Association of Teachers of Preventive Medicine. SmithKline Beecham (Raritan, New Jersey) provided Orasure collection devices at no cost for the first 300 participants.

census data sources. We used disproportionate and adaptive sampling techniques (Blair, 1999; Kalton, 1993; Hansen, Hurwitz, & Madow, 1953; Sudman, 1976) to construct a random-digit-dial (RDD) sample for the designated target areas in each city. The prevalence of MSM households across zip codes in the obtained sample ranged from 1.6% to 4.0% in the lower-density areas and from 4.1% to 33.6% in the higher-density areas. In all four cities, the sampled areas account for an estimated majority of all MSM households. Men age 18 or older who reported having sex with another man since age 14 or who defined themselves as gay or bisexual qualified for the study. Only one respondent per household was interviewed. Data collection occurred from November 1996 through February 1998.

UMHS interviewed a diverse sample of 2881 MSM. Approximately 21% of respondents were completely or semi-closeted (i.e., had either never told anyone or told less than half of family or friends that they had/have same-gender sex); approximately 7% did not fall into standard definitions of gay or bisexual orientation; 30% had a high school education or less; 16% were late middle-aged or elderly; and 21% were ethnic minority men. Prior reports from the UMHS found that men outside the densest neighborhoods provide greater representation of important social groups to the HIV epidemic, including lower-socioeconomic-status men, ethnic minorities, and closeted men (Mills et al., 1998). Thus, the UMHS provides a glimpse into a broad community of MSM residing within large urban centers of the United States.

HIV Testing with a Home Collection Kit

A subsample of 615 men were called back from April 1997 through June 1998 and offered an HIV test using an oral home collection kit based on the Orasure oral collection device. Only subjects who had consented to being recontacted after their UMHS interview were called (2402/2881, or 83.4%). The intent was to secure sufficiently large numbers of HIV-positive, HIV-negative, and never-tested respondents so that comparisons of acceptance rates could be made across serostatus groups. Recontacts were attempted with all men never previously tested for HIV and with all available African-American, Hispanic, Asian, and Native American men until the quota of 600 interviews was reached in order to permit race/ethnicity comparisons of acceptance rates. UMHS respondents who were interviewed in Spanish ($n = 17$) or who had been tested but refused to disclose their HIV serostatus ($n = 16$) were not recontacted. The remaining sample was distributed between men reporting previous HIV-negative and -positive tests approximating the proportions of self-reported HIV status in the UMHS sample. The sample distribution of completed interviews by city was 240 in New York, 190 in San Francisco, 111 in Los Angeles, and 74 in Chicago.

During telephone callbacks, short interviews were conducted to assess history of HIV testing, familiarity with HIV home collection testing, and willingness to be HIV tested using an oral home collection kit. Subjects were asked to consent to using the home collection kit with an ID number that linked their HIV test result with the information they had previously given in the UMHS interview. Subjects declining a linked HIV collection kit were offered the option of participating without linkage to the UMHS interview. Subjects choosing an unlinked test made up an alphanumeric ID when they received the kit, retaining the made-up ID to allow them to call for their test result. They were asked to give their age group, race/ethnicity, education, and city of residence on a card included in the kit. A $25 incentive was offered for participation.

Kits were made up by project staff using an Orasure collection device, a ziplock bag and gauze to contain and protect it, instructions for use, a two-part perforated card with the ID number on both parts (one part to be mailed to the lab with the specimen and the other part to be retained in order to call for the result), and a mailing box with a first-class permit addressed to the laboratory. Kits were sent by Federal Express or by U.S. mail if a post office box was given as an address. Participants received a follow-up phone call to confirm that they had received the kit and understood how to use it. If the kit was not mailed back within two weeks, additional follow-up calls were made to encourage completion of participation, and a replacement kit was offered if the first had been misplaced or misused.

HIV testing of oral specimens was performed in the California Department of Health Services Viral and Rickettsial Disease Laboratory (VRDL) using the FDA-approved Vironostika HIV-1 MicroElisa system (Organon Teknika). Enzyme immunoassay (EIA) repeatedly reactive samples were confirmed by Western Blot. If the test result was inconclusive or if the sample received at the VRDL was insufficient for HIV testing, the lab informed the study investigators, and the subject was recontacted and mailed a replacement kit. Specificity is reported in the Orasure package insert as 97.7% in a high-risk population and 99.6% in a low-risk population. Both sensitivity and specificity of the Organon HIV-1 assay on oral fluids collected with the Orasure device were reported to be 99.9% in a study of 3570 known positive and negative subjects (Gallo, George, Fitchen, Goldstein, & Hindahl, 1997).

HIV test results and counseling were available at a toll-free telephone number from trained counselors at the San Francisco AIDS Health Project and, later in the study, from trained HIV counselors at the Center for AIDS Prevention Studies (University of California, San Francisco). A protocol for phone counseling was developed by the AIDS Health Project for informing participants of their results by telephone, counseling them on the meaning of the results and on safe sex practices, and referring them to services if appropriate.

The Committee on Human Research at the University of California, San Francisco approved the study protocol.

Measurements and Statistical Analysis

For subjects permitting linkage of HIV test results with their UMHS interview, data were combined to permit analysis of associations with self-reported sexual behavior and demographic variables. A dichotomous variable for high-risk sex was created based on whether or not unprotected anal

intercourse was reported during the 6 months prior to the UMHS interview. Age was treated as an ordinal variable by dividing subjects into four groups: 18–29, 30–39, 40–49, and 50+ years. Associations in tabular data were tested with the chi-square statistic. Logistic regression was used to assess multivariate relationships with indicator variables entered for nominal variables. Stepwise and "best variables (score statistic) " regression models were examined and checked for evidence of interactions.

Results

Overall, 79.7% (490/615) of men recontacted consented to a home collection HIV test and 84.1% (412/490) mailed a specimen to the lab (Table 1). Thus, 67% of the pilot study sample provided HIV test results. Technically, the home collection kits performed quite well. Only two subjects had indeterminate test results. One had tested HIV-positive in the past, and his retest on a blood-based home collection kit was positive. The second had not been tested previously; his retest was negative on a blood-based kit. Ten subjects returned devices with insufficient sample to be tested; seven returned a second kit and were tested successfully. There were few other problems using the kit. Only five subjects reported a problem: Two had difficulty breaking off the plastic handle of the device, one said the liquid spilled out of the device container, one wasn't sure when the device was moist, and one felt that the device dried out his gums. Some who didn't report a problem took the offered opportunity to comment. The most common negative comment was dislike for the taste of the device (n = 11). Many participants commented on how easy it was to use.

Results from HIV testing followed self-reported HIV status very closely. All those reporting a prior HIV-positive test (n = 77) tested positive. Only 1.5% (4/266) reporting a prior HIV-negative test were positive, and 2.9% (2/69) with no prior HIV test were positive.

Prior self-reported HIV test status had the strongest association with both initial consent and participation. Among those who initially consented, 94% of men with a prior positive test mailed back the device, compared with 83% of those whose last test was negative and 79% of those with no prior test result. These latter two percentages were not significantly different. Participation varied from 83% in those with a prior positive test to 54% in those without a prior HIV test (Table 1). Only 3.5% (17/490) chose a test kit option unlinked to their UMHS interview, and only 17.7% (3/17) of these mailed a specimen to the lab.

Other aspects of prior HIV testing history were not significantly associated with participation, nor was prior knowledge of HIV home collection kits (Table 1). Ninety-two percent of the sample had heard of HIV home collection tests. Among the demographic variables, only city of residence approached statistical significance (p = 0.07, Table 2). Individuals reporting recent unprotected receptive anal intercourse were only slightly more likely to participate than those not reporting high-risk behavior, and this difference was not statistically significant.

Table 1. Proportion of subjects participating in HIV testing with an oral home collection test kit by measures of prior HIV testing history

	N	Consented to Kit and Mailed Back Specimen *N* (%)	
Total sample	615	412 (67.0)	
Prior HIV test status			p = 0.001*
Prior positive test	90	75 (83.3)	
Prior negative test	393	266 (67.7)	
No prior HIV test**	132	71 (53.8)	
Year of last HIV test			p = 0.19
'97–'98	185	128 (69.2)	
'95–'96	150	107 (71.3)	
'94 or earlier	154	91 (59.1)	
Number of prior tests			p = 0.37
One	111	74 (66.7)	
Two to four	210	145 (69.1)	
Five or more	166	123 (74.1)	
Prior use of home collection HIV test kit†			p = 0.12
Yes	12	6 (50.0)	
No	392	336 (70.7)	
Heard of home collection HIV test kit			p = 0.48
Yes	568	386 (68.0)	
No/Don't know	36	21 (58.3)	
Seen advertising for home collection HIV test kit			p = 0.53
Yes	463	320 (69.1)	
No/Don't know	104	66 (63.5)	

*P-value from chi-square test.
†Asked only of those with prior HIV test.
** Includes six persons with prior test who did not receive test result.

On several measures differences in the percentage consenting to be mailed a kit were offset by differences in the opposite direction of the percentage that mailed in a specimen (data not shown). For example, compared with whites and African-Americans, more Hispanics consented but fewer mailed in a specimen. Similarly, subjects from Chicago had a higher consent but a lower specimen return rate. Men over age 30 were the least likely to consent to a kit but the most likely to send in a sample if they did consent.

Prior HIV test status was only weakly confounded with city of residence, age, and race/ethnicity. No evidence of interactions was found. Consequently, a multivariate logistic regression model retained prior HIV test status as the strongest association with participation, and the association with city was still marginally significant, as New York City residents were somewhat less likely to participate (OR = 0.7, 95% CI = 0.4–1.1, p = 0.08, Table 3).

A separate model run only on those who consented to receive a kit showed that men older than 30 years were more likely to mail back a test specimen (OR = 1.4, 95% CI = 1.1–2.0, p = 0.03, full model not shown).

Table 2. Participation by demographic and behavioral characteristics

	N	Consented to Kit and Mailed Back Specimen *N* (%)	
City of residence			$p = 0.07$*
San Francisco	190	140 (73.7)	
New York City	240	151 (62.9)	
Chicago	74	52 (70.3)	
Los Angeles	111	69 (62.2)	
Race/ethnicity			$p = 0.44$
White	426	286 (67.1)	
African-American	47	31 (66.0)	
Hispanic	76	54 (71.1)	
Other	64	40 (62.5)	
Age			$p = 0.41$
18–29 years	349	231 (66.2)	
30–39 years	149	103 (69.1)	
40–49 years	60	44 (73.3)	
50+ years	57	34 (59.6)	
Sexual risk behavior			$p = 0.34$
Unprotected anal intercourse (≤ 6 mo)	164	115 (70.1)	
No unprotected anal intercourse (≤ 6 mo)	433	286 (66.1)	

*P-value from chi-square test.

Table 3. Multivariate logistic regression* model predicting consent to HIV testing and mailing back the oral device from a home collection test kit (*N* = 615)

Variable	Odds Ratio for Returning Specimen	95% CI	*p*-Value
Prior HIV test status			
Prior HIV+ test	4.1	2.1–8.0	0.0001
Prior HIV– test	1.9	1.3–2.9	0.002
No prior test result	Ref.	—	—
City of residence			
New York City	0.7	0.4–1.1	0.08
Los Angeles	0.7	0.4–1.1	0.11
Chicago	1.1	0.6–2.0	0.78
San Francisco	Ref.	—	—

Table 3. Multivariate logistic regression* model predicting consent to HIV testing and mailing back the oral device from a home collection test kit (*N* = 615)

Variable	Odds Ratio for Returning Specimen	95% CI	*p*-Value

*Logistic models selecting the best four variables based on score statistic, backward selection, and forward selection ($p < 0.15$) gave a model including prior HIV+ test, prior HIV– test, NYC, and LA. Chicago was included in this final model to complete representation of all four sites.

All subjects were asked if they had concerns about testing with an HIV home collection kit before the conditions of the study were explained. Forty-two percent of the sample said they had at least one concern (Table 4). Two of those concerns, the cost of a kit and performing a fingerstick, are relevant to the current commercial blood-based kit but not to the kit we subsequently asked them to accept. The most common concerns were uncertainty about the accuracy of the test (23%) and the lack of in-person counseling (14%). Subjects with a prior HIV-positive test result were more likely to be concerned about the lack of in-person counseling (69.8% of those with a concern) than subjects whose last HIV test was negative (35.3%) or subjects with no prior test (41.5%). They were less likely to be concerned about the validity of the test (33.4% compared with 61.7% and 52.8%, respectively, in the other two groups). Those men who had no previous HIV test were slightly more concerned about a fingerstick blood draw (11.2%) compared with HIV-positive men (2.6%) and men whose last test was negative (7.2%).

Subjects refusing participation were asked to give their reasons for declining (Table 4B). The reason most frequently given was concern about confidentiality (36%), followed by not wanting an HIV test (22%) and uncertainty about the accuracy of the test (21%).

Discussion

We found that a high proportion of MSM who had been identified by a random-digit-dial telephone survey would subsequently participate in HIV testing with a home collection kit. The combination of telephone interview and home collection HIV testing appears to be a feasible way to obtain population-based HIV seroprevalence data. Probability sampling of households by telephone is less costly than probability sampling by in-person contact, and mailing test kits is a minimal additional cost. Any form of probability sampling is

likely to be more costly than venue-based sampling, but probability samples remain the gold standard against which other types of sampling should be assessed.

Less than 5% of subjects gave reasons for refusing participation that were based on home collection methodology. The largest proportion of subjects refusing either were concerned about confidentiality, a concern not restricted to home collection testing, or just did not want an HIV test. Only 3.1% (19/615) refused to participate because they were worried about the accuracy of a home collection test, and only 1.6% (10/615) did not want to be informed of an HIV test result by phone. In addition, home collection kits had only recently been licensed when this study was done. With a longer time on the market and greater familiarity, their acceptability is likely to increase.

The strong association with prior HIV test status points to a limitation of all HIV seroprevalence surveys that are not based on blinded samples originally drawn for another purpose. Persons not previously tested are the most likely to decline participation. We obtained a 54% participation rate in this group. Although not optimal, we believe these results show it is possible to get useful data for this hard-to-study group. A small percentage may have participated because they were concerned about blood drawing, and we offered an oral-specimen test. The complete agreement between a self-reported prior positive HIV test result and the home collection test result indicates that self-report from HIV positives who have been tested is valid.

Table 4. Concerns about HIV home collection testing

Concern	Number	Percent
A. Concerns Among All Subjects Interviewed (*n* = 568)*		
Any concern	241	42
Unsure test accurate	130	23
Concerned about lack of in-person counseling	81	14
Worried about confidentiality	26	5
Doesn't want to stick finger for blood	18	3
Doesn't want to be informed by phone	13	2
Thinks kits are too expensive	10	2
Other concerns	99	17
B. Concerns Among Subjects Declining to Participate (*n* = 125)**		
Didn't want any HIV test now	45	36
Worried about confidentiality	28	22
Unsure test accurate	26	21
HIV+, doesn't need test or see point	15	12
Will test only with doctor	11	9
Doesn't want to be informed by phone	9	7
Didn't want kit to be mailed	2	2
Too busy, can't be bothered	5	4
No risk, not sexually active, no point	5	4
Don't know/declined to answer	14	11
Other concerns†	8	6

*Three responses allowed; concerns asked prior to consent question, not specifying type of kit (oral or blood specimen) or cost.
** Three responses allowed; asked of those refusing after the type of kit and financial incentive had been specified and the consent question asked.
† Other reasons included: moving soon, loss of control, didn't see the point, and not believing the hypothesis that HIV causes AIDS.

The 67% participation rate is comparable to cooperation rates with providing a serum sample for in-person probability household surveys. For example, 72% of participating adults in NHANES and 71% of adults in a household survey of multiethnic neighborhoods in San Francisco provided serum samples (Siegal et al., 1994; Hahn, Magder, Aral, Johnson, & Larsen, 1989). The 78% of identified eligible persons who participated in the parent UMHS survey was also comparable to participation rates in probability household surveys: 64% of eligible persons participated in the multiethnic neighborhood survey in San Francisco (Siegal et al., 1994), and 82% participated in NHANES III (McQuillan et al., 1997).

We cannot know for certain what participation rate we would have obtained if the HIV test kit had been part of the original UMHS study protocol and included in the initial interview. Seventeen percent of interviewed UMHS participants did not give permission to be recontacted for future interviews and were therefore not included in our HIV testing substudy. However, it cannot be assumed they would not have participated in HIV testing. The original UMHS telephone interview was exceptionally long (mean time = 75 minutes), requiring more than one session in many instances to com-

plete. The substudy was also not as well funded as the UMHS to pursue completions. The 84% return rate of the home collection kit among those who consented was somewhat lower than we expected, but a limited number of follow-up calls were made, and no additional incentive was offered to procrastinators. On the other hand, the length of the interview could have had an intervention effect, increasing willingness to be HIV tested. A less burdensome initial interview coupled with the offer of a home test kit and more resources to follow up on those not returning specimens would have been a somewhat different study.

The methodology was well accepted by MSM of color and by a range of different age and socioeconomic groups, and with only small geographic variation among the four cities. We observed only modest differences in participation by age, race/ethnicity, and city of residence. The lower participation among older MSM may reflect a lower self-perceived risk in this group. A telephone survey of MSM in Seattle reported similar rates of prior HIV testing (82% in their study compared with 88% in UMHS) and reported that older men were less likely to have been tested, and frequently (57%) gave perceived low risk as a justification (Campsmith et al., 1997). All of the six HIV-positive men who had not previously tested positive in our sample were below age 40.

Linkage of the test result to questionnaire information was not a significant issue for subjects. Only 3.5% wanted a test unlinked to their UMHS interview. Since the majority of those agreeing to an unlinked test did not send the kit back, no meaningful increase in participation was obtained by offering this option. The lack of interest in the unlinked test option may be a result of the quasi-anonymous nature of the UMHS interview. Respondents were identified only by first name, so no fully identifying information was obtained, although the contact files did of necessity include a telephone number. Kits were mailed in a manner specified by the participant, including mailing to post office boxes, and no additional information obtained for mailing was recorded.

HIV prevalence estimates obtained from combining RDD phone sampling with HIV home collection testing is limited by the coverage of the initial phone sample. The UMHS methodology required screening for current sexual identity or sex with a male since age 14. Some individuals undoubtedly declined to answer the screening questions and were not included in the survey. Nondisclosure of sexual behavior is a feature of any attempt to study MSM, whether in person or by phone, with a probability sample or a convenience sample. It may appear less of a problem in convenience samples taken, for example, from gay bars because persons identified in such venues have in effect already disclosed their sexual orientation. Consequently, what appears to be greater disclosure may well be a function of the origin of the sample. National samples have found no difference between in-person and telephone interviews in the proportion of respondents who disclose same-gender sexual behavior (Binson et al., 1996). The UMHS represented only targeted zip codes, but UMHS coverage of MSM within those zip codes is estimated to be from 88.2% to 98.2% in the four cities and overall to represent the majority of the MSM population in all four cities. The UMHS inclusive definition of MSM in fact reached many men who would be considered "closeted" and others who did not self-identify as gay or bisexual, something not achieved by most other samples of MSM (Osmond et al., 1994; Winkelstein et al., 1987). Of the 2881 originally interviewed, 3.1% identified as heterosexual and another 3.5% as "don't know, don't use a label," or other than homosexual/bisexual/heterosexual. Twelve percent of the sample had not disclosed their homosexuality to their family, and another 14% had disclosed to less than half of their family. The limitation of convenience samples in representing MSM is seen in the percentages of UMHS participants who in the past year had not been to sites frequently used for MSM sampling: 95% had not been to an STD clinic, 71% had not gone to a sex club or a bathhouse, and 63% had not gone to a public cruising area (park, beach, etc.).

Telephone surveys with biological specimen collection may not work for all populations at high risk for HIV infection. Injecting drug users, for example, probably cannot be sampled effectively by telephone, but the general method could be extended to a number of other high-risk populations, such as heterosexual men and women of color, and could be used to obtain seroprevalence estimates for a variety of viral infections. By adding a follow-up interview and second home collection kit, it could also be used to obtain population-based seroincidence estimates. We believe it warrants consideration as an important addition to the methods available to us to monitor the HIV epidemic.

References

Anderson, J., Nelson, D., & Wilson, R. (1998). Telephone coverage and measurement of health risk indicators: Data from the National Health Interview Survey. *American Journal of Public Health*, *88*(9), 1392–1395.

Bayer, R., Stryker, J., & Smith, J. (1995). Testing for HIV infection at home. *New England Journal of Medicine*, *332*(19), 1296–1299.

Binson, D., Moskowitz, J., Mills, T., Anderson, K., Paul, J., Stall, R., & Catania, J. (1996). Sampling men who have sex with men: Strategies for a telephone survey in urban areas in the United States. Presented at the 51st Annual Conference of the American Association for Public Opinion Research, Salt Lake City, Utah, May 16–19.

Blair, J. (1999). A probability sample of gay urban males: The use of two-phase adaptive sampling. *Journal of Sexual Research, 36,* 39–44.

Campsmith, M. L., Goldbaum, G. M., Brackbill, R. M., Tollestrup, K., Wood, R. W., & Weybright, J. (1997). HIV testing among men who have sex with men—Results of a telephone survey. *Preventive Medicine, 26,* 839–844.

Gallo, D., George, J. R., Fitchen, J. H., Goldstein, A. S., & Hindahl, M. S. (1997). Evaluation of a system using oral mucosal transudate for HIV-1 antibody screening and confirmatory testing. *JAMA, 277*(3), 254–258.

Hahn, R. A., Magder, L. S., Aral, S. O., Johnson, R. E., & Larsen, S. A. (1989). Race and the prevalence of syphilis seroreactivity in the

United States population: A national seroepidemiologic study. *American Journal of Public Health, 79,* 467–470.

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Survey methods and theory. Vol. I: Methods and applications; Vol II: Theory.* New York: John Wiley & Sons.

Kalton, G. (1993). Sampling considerations in research on HIV risk and illness. In D. G. Ostrow & R. C. Kessler (Eds.), *Methodological issues in AIDS behavioral research*. New York: Plenum Press.

McQuillan, G. M., Khare, M, Karon J. M., Schable, C. A., & Vlahov, D. (1997). Update on the seroepidemiology of human immunodeficiency virus in the United States household population: NHANES III, 1988–1994. *Journal of Acquired Immune Deficiency Syndrome and Human Retrovirology, 14,* 355–360.

Mills, T. C., Paul, J., Pollack, L., Stall, R., Binson, D., & Catania, J. A. (1998). How are MSMs who reside in gay ghettoes different from other MSMs? Paper presented at the 1998 Annual Meeting of the American Public Health Association, Washington, DC.

Osmond, D. H., Page, K., Wiley, J., Garrett, K., Sheppard, H. M., Moss, A. R., Schrager, L., & Winkelstein, W. (1994). HIV infection in homosexual and bisexual men 18 to 29 years of age: The San Francisco Young Men's Health Study. *American Journal of Public Health, 84*(12), 1933–1937.

Siegal, D., Larsen, S. A., Golden, E., Morse, S., Fullilove, M. T., & Washington, A. E. (1994). Prevalence, incidence, and correlates of syphilis seroreactivity in multiethnic San Francisco neighborhoods. *AIDS Education and Prevention, 4*(6)*,* 460–465.

Sudman, S. (1976). *Applied sampling*. New York: Academic Press.

U.S. Bureau of the Census. (1997). Census of population and housing, 1990: Public use microdata samples, U.S. Washington, DC: U.S. Department of Commerce.

Winkelstein, W., Lyman, D. M., Padian, N., Grant, R., Samuel, M., Wiley, J. A.., Anderson, R. E., Lang, W., Riggs, J., & Levy, J. A. (1987). Sexual practices and risk of infection by the human immunodeficiency virus. *JAMA, 252(3),* 321–325.

The Methodological Implications of Conducting Web-Based Research

Elizabeth T. Miller

Introduction

The development of Internet-based technology has ignited a transformation in the methods and modes of current-day communication and data systems. We have yet to fully understand the implications that these systemic changes will produce. Some propose that computer technology will bring about changes comparable to an accelerated version of the Industrial Revolution (Evans, 1980; Toffler, 1981).

In 1998 more than 150 million adults worldwide used the Internet on a weekly basis from their business or home, and that figure is expected to rise to nearly 320 million Internet users—or 52.5 per 1,000 people worldwide—by the end of the year 2000 (Computer Industry Almanac Inc., 1999). This estimate does not include child and adolescent Internet use or those who use the Internet on an irregular basis. This number will continue to grow exponentially each year and, as access increases, boundaries of Internet use will change as well. Consumers are pushing the functionality of the Internet more than for any other technology in our history, and research participants represent a very important consumer group.

Research is predicated on the exchange of information (communication) and processing of that information (data). Health care research in particular requires direct communication between researchers and participants, typically in the form of surveys, questionnaires, assessments, diaries, and interview questions. Data collected are analyzed, and the results of those analyses build the foundation for a greater understanding of the human condition. Given the growth in electronic communication networks and data storage systems, the impact on alternative methods of conducting psychological research will likely be great. Utilizing web-based technology in data collection and management procedures offers an unprecedented opportunity to conduct cross-sectional and longitudinal research studies in a cost-efficient manner (Miller & Marlatt, 1999). This paper is aimed at providing an overview of the methodological implications of conducting web-based research.

Web-Based Research Defined

Web-based research techniques fall into three categories: data collection, data entry, and other.

Elizabeth T. Miller is in the Department of Psychology at the University of Washington.

Data Collection

Data collection includes surveys, diaries, assessments, and other forms of collecting information directly from the research participants. Currently, the most appropriate populations of participants include employees, general adults, and students since they are the most likely to have access to and experience with the Internet. Other populations should not be excluded simply because they do not fit into one of these categories. Just as many employees who have access do not utilize the opportunity, many individuals without jobs engage in regular use. Providing Internet access is an option for your study and can vary with regard to cost and location, including libraries, Internet cafes, medical provider waiting rooms or offices, home trials, and other venues.

Data Entry

Data entry refers to techniques of accumulating data via research staff or other hired employees. Entering information from a chart review process or responses from paper surveys into a web-based data entry module are two examples of methods in which the Internet can be utilized to create a single data repository. Since the data entry module exists on a server, any computer with access to the Internet can function as a data entry station, as opposed to designing the research database on one designated data entry computer. The advantages of the single data repository are greatly increased within the context of a multi-site trial.

Other

The "other" category includes a range of techniques that are more broad in nature and can be utilized by the researcher and/or the participant. One example is the ability to merge different data sets related to the same participants (e.g., lab results, satisfaction surveys, and behavioral assessments). Another technique is the development of a project website as a means of keeping the research team and participants in closer contact and thus communicating the importance of their contribution to the success of the project. This project website could include announcements, preliminary findings, an invitation for feedback, and general updates.

Advantages of Web-Based Research

Conducting web-based research offers several advantages over the more traditional paper- or telephone-based methods,

including reduced costs in five primary categories: financial, technical, administrative, time, and general.

Financial

Direct financial costs include those associated with publishing and distributing paper surveys, mailing/telephoning study participant reminders, hiring and training staff to conduct data entry, error checking, and project coordination. The estimated costs to develop, publish, and maintain web-based surveys are significantly lower and require significantly fewer administrative/project coordination resources. Costs associated with traditional assessment methods, such as publishing and distributing paper surveys, mailing/telephoning study participant reminders, and data collection and entry are eliminated (Schmidt, 1997), as are the costs associated with an increased sample of research participants (Buchanan & Smith, 1999).

Technical

Automated scoring, restricting irrelevant questions through programmed skip patterns and algorithmically defined dynamic flow, and automated item reliability checks represent some of the important technical benefits of computerized data collection (or data entry) methods. These built-in logic techniques result in cleaner data. In addition, researchers can potentially conduct expedited pilot studies leading to expedited changes in final survey items. With the ability to examine data in real time, research can be redirected (e.g., alternative sampling techniques used) to create a more representative sample. Conducting web-based data collection (or data entry) maximizes all of the aforementioned benefits while also increasing accessibility of survey forms (or data entry forms), eliminating potential hardware memory issues related to other computerized methods (e.g., CD-ROM), and providing immediate availability of cleaned data from anywhere in the world, at any time of day.

With increased access through the development of web-based surveys, there is the potential for an increase in research study participation and multi-site research, resulting in a more demographically diverse sample.

Furthermore, in order to develop accurate web-based data collection forms, an iterative process between the research team and technical staff is required. This development and testing process leads to a more precise representation of the data fields and in fact enhances the quality of the data collected because the researcher must work to clearly define how he or she will analyze the data prior to collecting it. This may turn out to be the most beneficial aspect of utilizing web-based research methods.

Administrative

Increased data accuracy, more complete data, and access to clean data immediately upon submission reduce the time and energy necessary to conduct traditional administrative tasks.

Time

The savings in time for both staff and research participants are immense. The research staff doesn't have to spend time publishing and distributing paper surveys or making telephone calls, organizing returns, going through paper versions frantically trying to identify "extreme" cases for follow-up/referrals, or worrying about the other tedious tasks associated with the day-to-day administration of a research study. Participants benefit from using web-based surveys as well, since it provides increased accessibility and use of dynamic/interactive forms, which eliminates the viewing of irrelevant questions. Real-time individualized feedback can be presented to users based on the responses submitted. This is important due to the traditional limitations encountered in survey research, such as sample size, attrition rates, financial resources, design issues (cross-sectional vs. longitudinal), and accessibility.

General

A web-based survey also has the potential to reach hidden populations who might not otherwise participate in research focused on high-risk behaviors (Nicholson, White, & Duncan, 1998). Conceivably, those individuals who might not be willing to drop by to pick up a survey and return it via postal mail, attend an organizational meeting (e.g., class) to complete a survey, or participate in a telephone interview would be more willing to log on to the Internet at their convenience. Previous research indicates that the psychometric properties of computerized psychological assessments are not compromised (Skinner & Pakula, 1986) and that validity of the responses regarding high-risk sexual behaviors may in fact be enhanced (Turner et al., 1998).

Longitudinal studies are much more easily implemented and managed since all management can potentially be conducted online with the help of a single centralized database and customizable automated reporting features. With the costs associated with administrative and time resources minimized and accessibility increased, larger multi-site samples could easily be managed, *which is an important public health consideration.*

Considerations for Conducting Web-Based Data Collection/Entry

The potential problems with conducting web-based data collection/entry can be defined as either technical or practical in nature. Technical considerations include incomplete responses, unacceptable responses, multiple submissions, security and data integrity violations, reliability and validity of responses, browser incompatibility, and technical expertise. Practical considerations include access to computers, personal obstacles, computer literacy of participants, ethical

considerations, use of copyrighted/proprietary measures, and sample bias (Miller, 1999; Schmidt, 1997).

Technical

Incomplete responses are often due to fatigue, discomfort, embarrassment, insufficient instruction or question clarity, lack of time, and sloppiness. At times, researchers may be ethically bound to permit incomplete responses. Web-based surveys can be developed to require that specific answers be completed in order for users to move to subsequent survey sections. This is advisable in order to utilize dynamic flow and minimize the presentation of irrelevant questions, thus increasing data integrity. Unacceptable responses can be identified through software coding. For example, traditional paper survey items must be excluded when a user selects an answer not offered (e.g., writing in his or her own response), marks a circle in between the answers offered, or circles more than one answer. Using web-based surveys, singular response options can be enforced where necessary and multiple response options allowed only where appropriate. Multiple submissions can be made impossible by prohibiting access to the web-based survey once the user has successfully submitted the final form. This can be accomplished through password utilization. Security and data integrity violations can be obviated by requiring users to log on to a specified website and enter a personal identification number (PIN) composed of unique identifiers determined by the specific project needs. Additionally, accurate and precise web-based survey development is required to ensure the highest data integrity possible. To obviate browser incompatibility issues, web-based survey developers need to program with the least common denominator in mind (i.e., Netscape 2.0 or above and Microsoft Internet Explorer (MSIE) 2.0 or above). Specific directions for logging on to the Internet, downloading Netscape or MSIE, and setting up an e-mail account, and general definitions of the web can be provided to your participants via e-mail, mail, and/or telephone.

Finally, technical expertise is required to successfully develop, implement, and host the web-based data collection/entry forms. For surveys or assessments that are relatively short (e.g., in which it is reasonable to present all the questions to a user on one scrollable page), straightforward HTML code may be sufficient and is relatively easy to learn (see "Technical Information" below for more information). However, if the project needs are more complex (e.g., surveys include skip patterns, confidentiality requires password login, differences in assigned condition requires diverse survey administration), expert computer programming skills are highly advised. Access to a server (a secure server is typically necessary, particularly for confidential transactions) is essential to enable the hosting of the web-based forms. Your university (e.g., the information services department) may be willing to provide personnel to develop and administer a secure server. The necessary knowledge, skills, and experience should be assessed early on in order to minimize unexpected and substantial time delays. Alternative options include utilizing a campus research lab (e.g., a psychology department research lab) or hiring a web-based data collection and management service provider (e.g., DatStat.com[1]). Assessment of the level of expertise applies to all potential assistance you seek, including developing the web-based data collection/entry forms, implementing and hosting the forms on a secure server, managing incoming data, troubleshooting and responding to user problems, and delivering the final dataset(s).

Practical

Access to computers can be an obstacle depending on the participant population. It is important to consider whether access is currently in place or will need to be put in place prior to adopting web-based data collection or entry techniques. If computer access is not already in place, the requirements of establishing connection(s) will have to be weighed with all of the other costs and benefits.

Farrell (1991) suggests that negative attitudes, lack of time, lack of skill, and lack of hardware/software all represent additional personal obstacles. Negative attitudes and lack of time, which are likely interrelated, are the most difficult to address and modify. Skills can easily be enhanced with clear directions and examples, thus increasing computer literacy for the long term, and the necessary hardware/software can be supplied directly or indirectly.

Skinner and Pakula (1986) advocate that researchers consider the professional responsibilities associated with conducting computerized psychological assessment, such as ethical considerations, access by unequal users, and premature use of unvalidated test interpretation systems. Conducting a research study on the web typically requires that participants sign a paper-based informed consent form prior to participation, although some human subjects committees and internal review boards allow an electronic signature or default consent to participate. Documenting security techniques is an important ingredient to obtaining approval. As with all research, it is impossible to delineate all of the potential effects that research participation might have, particularly with this relatively new method (Bier, Sherblom, & Gallo, 1996); however, consideration should be given to this topic at project conception. Inclusion of copyrighted/proprietary measures in a web-based survey requires obtaining the appropriate permissions. The process of obtaining permission is highly variable and may not be identical to requesting permission to use the paper-based survey forms. Experience in this process suggests educating decision-makers about the proposed methods of administration, and confidentiality precautions are important.

Sample and response bias have long been important considerations for both traditional and innovative research techniques. These will continue to be important issues to address and explore in the context of conducting web-based research. Demographic information on individuals with access to the Internet is not completely known, in part because it is changing rapidly and is therefore difficult to measure. In addition,

unknown response biases may exist due to contextual issues (e.g., privacy of computerized assessment, confidentiality concerns, question types supported by computer generator) regarding use of the Internet as a data collection tool. These biases deserve further investigation.

Whether these obstacles are unique to computerized, specifically web-based, research methods is unclear, but they are important considerations. Certainly, many of these obstacles will likely become obsolete as the Internet becomes more integrated into daily living and decision-makers become more familiar with online techniques.

Technical Information

Web-based data entry/collection involves the connection of a user (e.g., study participant, staff member) to a server. The web browser (Netscape, Internet Explorer, etc.) is instrumental in allowing the user to interact with a survey or data collection form, thus providing data to the server. The users' responses travel back to the server for storage via the Internet in an insecure fashion unless security measures are taken. Browsers and servers support an encryption technology called Public Key Encryption via the Secure Sockets Layer (SSL) protocol. Enabling this encryption technology allows web-based data entry/collection to be achieved while maintaining the security of the users' responses as they travel to and from the server.

One of the advantages of incorporating computers into data collection is the ability to cull invalid or incorrect responses while the user is entering data. However, many web-based data collection efforts incorporate the use of computer programming languages (e.g., JavaScript) that are "client-side" technologies, which can cause problems. For example, a mischievous user could edit the JavaScript on a web page and send dirty data to the server. Validation must be completed on the "server-side" to resolve these issues, and this can be done through the use of Common Gateway Interface (CGI) scripts. Employing this kind of "server-side" validation in addition to the JavaScript "client-side" validation leads to a nearly foolproof data collection session with users. Validating user identity and preventing multiple submissions can be achieved with further "server-side" programming based on project specifications (e.g., requiring student ID, birth date, or medical record number passwords combined with computer identification information), which in some cases may exceed the current paper validation standards.

Research Guidelines

Michalak and Szabo (1998) outlined several guidelines for conducting web-based research to address the potential skepticism and concern on the part of research participants or human subject review boards. These guidelines include identifying the nature and aims of the project, providing contact information, assuring confidentiality of data (e.g., encryption procedures), defining what constitutes consent to participate (e.g., completing the survey, signing and returning the consent form), conducting thorough pilot testing of the web-based survey, offering to post findings from research (e.g., on the designated website), utilizing newsgroups cautiously and considerately, offering a paper version of the survey, standardizing the environment as much as possible (e.g., restricting access at certain times of the day), considering the potential for increased diversity in sample characteristics, and obtaining appropriate copyright permissions prior to developing a web-based survey.

Further guidelines include identifying, screening, and enlisting the necessary technical expertise prior to undertaking a web-based research project (e.g., an experienced web developer/administrator or service provider), clearly defining measurement issues up front (e.g., skip patterns, logic algorithms), and developing web-based forms with the technical concerns of the least common denominator in mind (e.g., not all users have access to the most recent browser version). Finally, validation studies of the web-based version of the measures should be conducted to confirm that format of presentation and situational differences do not alter the results. The inclusion of a comparative paper-based sample would aid in this process.

Feasibility of Using Web-Based Data Collection Techniques

Several psychological web-based research studies have been conducted that resulted in the successful implementation of web-based data collection techniques and support for the feasibility of this research method. For web-based research to be declared a feasible alternative, an examination of the effectiveness of implementation and an analysis of the psychometric properties of web-based data collection techniques are necessary. In the web-based studies where comparisons were made between paper-based and web-based versions of psychological measures, results suggest that these differing administration techniques are equivalent.

Paper versus Web

In 1996, Miller and Marlatt conducted a longitudinal examination of New Year's resolutions utilizing web-based assessment methods with a small comparative paper-based sample. Participants did not differ significantly across assessment format with regard to choice of the #1 New Year's resolution, total number of resolutions made, number of start resolutions made, number of stop resolutions made, success of the #1 New Year's resolution, or gender (Miller & Marlatt, 1999). Smith and Leigh (1997) compared a paper-and-pencil measure to assess the nature and frequency of sexual fantasies among adults with an Internet-mediated version and found no reported differences in sexual orientation, marital status, ethnicity, education, or religious affiliation. There were, however, differences in age and gender between the two instruments.

Reaching Hidden Populations

Adult recreational drug use was examined via the web, and results suggest the web is a useful tool for reaching hidden populations, though a sample bias toward male, better-educated, and more computer-involved groups was revealed (Nicholson, White, & Duncan, 1998). A longitudinal effectiveness trial of two cost-efficient prevention programs aimed at decreasing harmful and hazardous alcohol use among college freshmen was conducted utilizing only web-based assessment techniques. Web-based data collection methods were applied in an effort to minimize resources and expedite the achievement of clean data and were found to be an effective alternative (Miller, 1999). Szabo, Frenkl, and Caputo (1996) conducted an Internet-based cross-sectional study on deprivation feelings, anxiety, and commitment to various forms of physical activity and were successful in contacting those who might not otherwise participate in such a study.

The general effectiveness of these studies supports the finding that web-based assessment techniques and web-mediated instruments are a feasible alternative; however, since no specific psychometric analyses were conducted, it would be premature to declare these techniques completely identical in nature and result to the more traditional methods.

Psychometric Properties

Some research has been conducted to examine the psychometric properties of web-based data collection forms specifically (Buchanan & Smith, 1999; Miller et al., 1999). Buchanan and Smith developed an Internet-mediated version of a popular measure of personality, Mark Snyder's Self-Monitoring Scale (SMS-R). They conducted reliability and validity analyses on both the Internet-based and paper-based versions and found no significant differences. Another personality inventory, the Self-Trust Questionnaire (STQ), was compared across samples completing a paper-based version with those completing a web-based version, and the results suggest that the psychometric properties of the scales were comparable across samples, with the single exception of variance among the web sample (Pasveer & Ellard, 1998). The authors posit that the advantages of web-based measures outweigh problems associated with generalizability.

In a study assessing the test-retest reliability of measures used in the addictive behaviors field, a comparison was made between web-based assessment techniques and traditional paper-based methods. No significant differences were found in terms of reliability estimates; however, analyses conducted to test for differences between subjective ratings of accuracy, convenience, and assessment format preference did reveal significant differences. Participants reported increased convenience and significant preferences for using web-based surveys (Miller et al., 1999). Anecdotal evidence from web-based research projects conducted at the University of Washington suggest that nonusers who developed computer skills based on instructions provided during research participation developed a sense of empowerment that allowed them to continue Internet use beyond the length of the study.

In summary, the results of these studies suggest that web-based data collection techniques are effective and do not statistically enhance or diminish the consistency of responses, lending further support to the feasibility of web-based techniques.

Conclusion

The purpose of this paper was to discuss both the advantages and considerations of conducting web-based research within the context of the methodological implications of this technique. Guidelines and technical information were included to describe the specifics involved in a project of this nature, and results from previous research were presented to provide an overview of current web-based research findings.

The possibilities for how the Internet will impact psychological research endeavors have not yet fully emerged. It is clear, however, that the potential for conducting data collection and data entry, providing individualized feedback, expediting the management of cleaner data, communicating with other researchers, data warehousing, and publishing results are maximized with web-based research methods.

Incorporating web-based technology into research is not a trivial task and should not be underestimated. Researchers interested in conducting web-based data collection/entry should consider the type of expertise necessary to conduct a successful study. Potential options range from learning web programming skills (e.g., HTML, JavaScript) to working with a research assistant with programming skills, an individual programming consultant, or a web-based service provider with combined programming, research, and web administration services (e.g., DatStat.com). Consideration of priorities, expertise, and expected timelines is an important first step in determining project needs.

The ultimate benefits of conducting web-based research include increased accessibility to participants, clean data, a single data repository, an expedited process, simplified data management, and more trustworthy outcomes. Web-based data collection is not currently appropriate for all populations and may never be the most appropriate technique. Strong consideration should be given to determine level of access, computer literacy, level of mental or physical disability, and the overall study design of the research project. For some populations, web-based data collection may prove to be more appropriate (e.g., high-risk behavior assessment).

The opportunities associated with providing mixed-mode options represent a new approach to participant satisfaction. Would the provision of a menu of survey technique options increase response rates and/or decrease attrition in longitudinal studies? Which measures are more/less reliable/valid when used in a web-based survey? Is there an interactional effect between mode, individual, measure, and context? Finally, further research is recommended to determine the

[1]DatStat.com (http://www.datstat.com) provides research consulting and web-based data collection, management, and analysis services to academic researchers and government agencies.

generalizability of the current trends in the effectiveness of web-based research in the health care field using other measures and populations.

References

Bier, M. C., Sherblom, S. A., & Gallo, M. A. (1996). Ethical issues in a study of Internet use: Uncertainty, responsibility, and the spirit of research relationships. *Ethics and Behavior, 6*(2), 141–151.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90,* 125–144.

Computer Industry Almanac Inc. (1999, August 18). http://www.c-i-a.com/199908iu.htm.

Evans, C. (1980). *The micro millenium*. New York: Viking Press.

Farrell, A. D. (1991). Computers and behavioral assessment: Current applications, future possibilities, and obstacles to routine use. *Behavioral Assessment, 13,* 159–179.

Michalak, E. E., & Szabo, A. (1998). Guidelines for Internet research: An update. *European Psychologist, 3,* 70–75.

Miller, E. T. (1999, under review). Preventing alcohol abuse and alcohol-related negative consequences among college students: Using emerging computer technology to deliver and evaluate the effectiveness of brief intervention efforts.

Miller, E. T., & Marlatt, G. A. (1999, under revision). Predicting successful self-initiated health-related behavior change in the context of New Year's resolutions: Utilizing the Internet for survey research.

Miller, E. T., Roberts, L. J., Neal, D. J., Cressler, S. O., Metrik, J., & Marlatt, G. A. (1999, under revision). Test-retest reliability of alcohol measures: Is there a difference between Internet-based technology and traditional methods?

Nicholson, T., White, J., & Duncan, D. (1998). Drugnet: A pilot study of adult recreational drug use via the WWW. *Substance Abuse, 19*(3), 109–121.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, and Computers, 30*(2), 309–313.

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, and Computers, 29*(2), 274–279.

Skinner, H. A., & Pakula, A. (1986). Challenge of computers in psychological assessment. *Professional Psychology: Research and Practice, 17,* 44–50.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, and Computers, 29,* 496–505.

Szabo, A., Frenkl, R., & Caputo, A. (1996). Deprivation feelings, anxiety, and commitment to various forms of physical activity: A cross-sectional study in the Internet. *Psychologia, 39,* 223–230.

Toffler, A. (1981). *The third wave*. New York: Bantam Books.

Turner, C. F., Ku, L., Rogers, S.M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science, 280,* 867–873.

Physician Response in a Trial of High-Priority Mail and Telephone Survey Mode Sequences

Danna L. Moore, Jim Gaudino, Pat deHart, Alan Cheadle, and Diane Martin

Introduction

The Department of Health and the Social and Economic Sciences Research Center (SESRC) at Washington State University conducted an experiment to determine the relative effectiveness of using U.S. Postal Service 2-day priority mail versus telephone-mode sequences in a survey of pediatric (PED) and family practice (FP) physicians on the subject of immunization. The trial was administered to the population of PEDs and FPs in Washington state. The overall objective of the survey was to measure attitudes and practices of physicians concerning immunization and the ways used to contact families of children who are due for vaccination.

Many researchers are plagued with the problem of trying to achieve acceptable response rates in surveys of physicians and other professionals. Response rates obtained for mail surveys of physicians that use less costly techniques and nonmonetary incentives are generally lower than the response experienced in mail surveys of physicians that use incentives. Similarly, encouraging participation is a key challenge in designing surveys of physicians and other professionals associated with establishment settings. While many techniques have been shown to be effective in increasing response to mail surveys for general population studies, and to a lesser extent for businesses, these same techniques have not been clearly demonstrated or evaluated for physician studies. The main venue explored in physician mail surveys for increasing response is the use of incentives. When one seeks to understand how physician surveys may differ from other populations primarily surveyed using mail techniques, there are three areas of literature to consult: literature on mail surveys, on interviewing professionals, and on establishment surveys. After reviewing this literature it is possible to gain an understanding of how one might improve physician response that is constrained on the use of incentives.

The problems associated with reaching a physician with a survey are very similar to problems experienced with business surveys in general. The barriers often associated with establishment surveys include reaching a correct or knowledgeable respondent for the business; getting past gatekeepers; and the ability to address problems associated with the survey questions or the professional's ability to respond at the onset of the interview.

For surveys of businesses to achieve a high level of response, it is generally recommended that more than one type of survey mode needs to be used. Paxson, Dillman, and Tarnai (1995) suggested that two elements of business surveys conducted by the Census Bureau since 1987 were consistently associated with higher levels of response. These elements were the use of telephone contacts as a means of follow-up for collecting data and the use of mandatory disclosure statements prominently stamped on mailing envelopes. Organizations outside of federal government are limited in their ability to use mandatory disclosure statements, and many government agencies are restricted in the use of financial incentives in survey efforts. However, organizations do have considerable flexibility in the mode sequence and design features they can implement for collecting survey data. Petrie, Moore, and Dillman (1997) provided empirical evidence from a large survey of manufacturing executives nationwide suggesting that use of a carefully designed set of procedures, in conjunction with reversing the order of the primary data collection mode from contact by mail survey to contact by telephone, can overcome some of the significant barriers inherent to business surveys or surveys requiring responses by a professional associated with an establishment.

Another important consideration in the survey process is the content and presentation of the survey questionnaire. When professionals such as physicians are needed for a response, there may be a higher requirement that the survey instrument presentation be professional and that the professional value of the survey be clear. Sudman (1985) describes the nature of the professional respondent (compared with the general population) as having a lower tolerance for ambiguity, stereotyping, and questions that don't make sense. Questionnaire and survey process design need to address the fact that professionals will quickly make a decision to participate based on the trade-off of the time required to be surveyed with the perceived value of the study, its benefits, and its professional relevance.

For surveys of physicians the survey challenge is complex and appears to be twofold. First, there is the need to penetrate

Danna Moore is the Research Coordinator for the Social and Economic Sciences Research Center at Washington State University.

Jim Gaudino and M. Patricia deHart are with the Office of Maternal and Child Health Community and Family Health Division of the Washington State Department of Health.

Alan Cheadle and Diane Martin are with the Department of Health Services, School of Public Health and Community Medicine, University of Washington.

the establishment or practice setting, and second, there is the need to gain cooperation from a professional who is very busy and often shielded from external contact. For physician studies, strictly experimental tests have not been done to evaluate and compare the response rate achieved and cost advantages associated with alternating the primary mode of administration for data collection that uses telephone and U.S. 2-day priority mail.

No published reports were found on the subject of using U.S. Postal Service 2-day priority mail as a strategy alone or in combination with other follow-ups such as telephone for reaching physicians or other professional populations. However, much has been written on the problem of maximizing response rates to mail surveys and the trial of some elements of mail techniques such as personalization, use of stamps of varying postage levels, and use of incentives.

The literature on nonmonetary and less costly methods of increasing physician response is not too extensive. Like studies for household mail surveys, studies of physician mail surveys show that attractive professional questionnaire design, personalization and higher classes of U.S. postage (first-class and certified mail vs. bulk) and inclusion of return envelopes have a payoff for response (Shosteck & Fairweather, 1979; Gullen & Garrison, 1973; Maheux, Legault, & Lambert, 1989; Shiono & Klebanoff, 1991; Urban, Andersen, & Tseng, 1993; and Del Valle, Morgenstern, Rogstad, Albright, & Vickrey, 1997). A prominent premise for most of these studies is that if nonresponse is reduced, then potential bias is reduced. Del Valle et al. (1997) found that this certified mail treatment in the final contact phase was more costly than the first-class mail by $0.43 per respondent, but that certified mail increased response rate in the final, third contact by 16.5% compared with the first-class mail group. For all contacts, the response rate reported for the certified mail group was 86.1%, compared with 82% for the bulk/first-class postage rate group. However, the authors concluded, using a sensitivity approach, that certified mail was not cost effective. Because strategies were mixed, one cannot compare a cumulative cost for all respondents through the successive contacts of the studies. Comparison of early to late responders showed the respondents had different distributions on characteristics of membership types, board certification, and practice settings. The authors also concluded that the third mailing was important to the quality of the sample data obtained, not only because it increased the sample size but also because it increased the representation of physicians with different characteristics.

Sample and Methods

The entire population of eligible physicians in Washington state was included, so there is no sample error in the study. The split-half experiment was conducted during January through October of 1998. Lists provided by state specialty association chapters (the American Academy of Pediatrics and the Washington Academy of Family Physicians) were the source for the 2,472 physicians (791 pediatricians and 1,681 family practitioners) in the study. The research reported is the result of the random assignment of physicians to either a telephone or a mail experimental treatment group for the primary mode of data collection.

For the listings of 2,472 physicians, 909 were missing phone numbers or addresses. Several resources were used to improve the sample contact information. These included the U.S. West Dex listing for physicians in Washington on the Internet, the American Medical Association for the state of Washington, directory assistance, and finally the Medicaid provider list for physicians currently receiving payments. Even with these efforts, a minimum of 10%—or 235 physicians—had some missing contact information.

A carefully designed set of procedures was used to achieve the highest response possible for both experimental groups in the assigned collection mode before the switch was made to the follow-up mode. That is, for the "telephone start" experimental group a notification letter was mailed prior to telephone contact, and up to 20 attempted telephone contacts were made over eight weeks before switching to high-priority mail follow-up. For the "mail start" experimental group, a U.S. Postal Service 2-day priority first mailing was used, with two additional follow-up contacts by mail (a postcard after one week, and a second questionnaire mailing after the fifth week) before the switch to a telephone follow-up contact after the eighth week. Physicians in both groups were contacted at their primary business establishments (the clinical settings in which they practice) to complete either a mail or telephone survey. The mail questionnaire was 12 pages long with a graphical cover, and the comparable telephone interview's average comletion time was 14.9 minutes. All 2-day priority mailer envelopes were bright red and blue cardboard mailers with the U.S. Postal Service logo printed on them. Another special mail design feature was a label added on the outside of the priority mailer envelope. This white label was located on the lower left of the envelope and read "Department of Health Information enclosed. If physician has moved or retired please contact 1-800-833-0867." The 800 telephone number was offered in both phone contacts and mail contacts.

Additionally, during the telephone contact phase for each experimental group, the script probed for new contact information if a doctor could not be reached or had moved. Because one of the goals of the overall study design was to define immunization practices in clinical settings, and it is well known that many physicians have nurses or other medical personnel administer the actual immunization under physician instruction, the interview also prompted for a proxy. After the 10th telephone contact, if the doctor was not going to be available, a substitute was asked for. Likewise, for the second questionnaire mailing, a proxy was asked for in the cover letter if the physician was not available.

Survey results are presented as a function of time, since the primary drivers of survey cost for each completed interview is the number of times an observation is contacted before it yields a disposition and the survey mode is utilized. Intuitively, the more times we contact a respondent before completing an interview or retiring an observation, the more it costs. The hypothesis tested is that the type of survey proce-

dure used may affect the time required to reach disposition. Response functions estimated provide the proportion of physicians reaching a completed interview at each stage. Cumulative cost estimates were made using a partial budgeting approach. Through this budgeting we account for interviewer and clerical staff labor needed at each stage and protocol, and the other variable costs for long-distance telephone as we go to more phone attempts, mail processing and data entry, stationery, and postage, which fluctuate based on numbers of observations entering each stage for recontact. Thus the results of this experiment allow a comparison of survey mode for: (a) the efficiency (rate of reply of data collection); (b) the overall response rate achieved at different stages and types of follow-up; (c) difference in response by sample characteristics; and (d) cost.

Results

Approximately 22% of physicians were ineligible for participation in the interview because they did not immunize children (n = 511), were retired (n = 45), or were deceased. Overall, the survey effort was successful, with 96% cooperation from the physicians reached. A final disposition for a physician reached is: a completed interview, an identification of ineligibility, or a refusal to participate. The response rate was 76% overall, as defined by a CASRO (Council of American Survey Research Organizations) response rate (Frankel, 1983). The CASRO response rate is the ratio of the number of completed interviews to the total number of sampling units. The estimated number of sampling units is determined using the eligibility factor to distribute the portion of the sample where eligibility was not determined. Twenty percent of the population was unreachable by phone or mail.

At the onset of the experiment, it was not known which treatment would work best—the sequence of telephone contacts to physicians with a follow-up switching to 2-day priority mail after week 6 (group 1 treatment), or the 2-day priority mail sequence with a follow-up switching to phone after week 6 (group 2 treatment). However, it was agreed with the Department of Health that whichever mode was working best at the eighth week would be the mode used for any additional follow-up contacts to raise the response rate. Physicians were randomly assigned to the two treatment groups. Table 1 provides a comparison of the two experimental groups for the cumulative CASRO response rate achieved at each stage. It can be seen in Table 1 at the end of week 6 of data collection, that if we had stopped contacts at this point there would have been better than a 20% difference between treatment groups. This is the stage to which many researchers carry the survey, at which point they then stop contact. Although the response rate at this stage for the group 2 treatment reached an acceptable level (62.3%), it is easily seen that switching to the opposite mode and extending the data collection period by two more weeks (to week 8) added just under 5% more response. For treatment group 1, the phone start intervention, the level of response at the end of week 6 was below 50%. Switching to the 2-day priority mail follow-up and extending data collection by two more weeks added just more than 17% response. Some conclusive statements can be made at that point. First, the 2-day priority mail treatment as the first survey contact to physicians with two additional mail follow-ups was more effective than the phone sequence. Second, extending the data collection period by two weeks and switching survey modes added response to both sides of the experiment. For group 1, switching to priority mail brought the response rate to just over 58%. As a final way to raise response rate overall in the study and for both sides of the experiment, a final "Hail Mary" priority-mail contact was made after letting all physician contacts rest for three months. The length of the rest period chosen was arbitrary and unintentional, and was actually the result of the time it took to get more survey funding in place. However, the long rest period had a surprising outcome in the level of response that was added to both sides of the experiment by this final, fifth contact by 2-day priority mail (week 24). For group 1 (phone start treatment) this brought the final CASRO response rate to 73.2%, and for group 2 (the mail start treatment) the final CASRO response rate was 76.8%. While the phone side of the experiment had a slightly lower response rate overall, it almost achieved the same level of response as group 1 after switching survey modes. However, this side never quite caught up with the other treatment group by the end of 24 weeks of data collection.

Table 2 demonstrates the changes that took place relative to cost of completed interviews achieved at various stages during the data collection period. The cost as presented includes a survey contractor's variable costs experienced during development of the two survey questionnaire versions (mail and telephone), CATI programming, and the changes in costs for labor, telephone long-distance, 2-day priority postage, other postage, and stationery relative to the number of recontacts that are made on each side of the experiment and at each stage. All other project costs, such as overhead and project management, are held constant and not included.

In this current study we found that pediatricians, as a physician speciality, tended to respond at a higher rate than family practice physicians and that response to questions tended to differ by both physician specialty and primary survey mode. Assignment of observations to the mail start treatment group were positively and significantly associated with physician response. The characteristic of practice location as rural or urban was not significantly associated with survey response.

Discussion and Conclusions

The reporting of response rate calculation in physician studies is inconsistent and makes comparison between studies difficult. Reported response rates fluctuate from as low as 30% to over 90%. How certain dispositions of unreachable respondents (return to sender, wrong telephone numbers, unpublished, and no-answer dispositions) are handled varies, and often they go unreported. In some physician studies it was noted that physicians not reached were excluded from the response rate calculation, and this tendency inflates reported response rates.

Table 1. Cumulative response rates by protocol stage and type of contact

Experimental Group 1: Phone Start (N = 1,235)			Experimental Group 2: 2-Day Priority Mail Start (N = 1,237)		Comparison
Timing and Protocol Stage	Response Rate[a] (%)		Protocol Stage	Response Rate[a] (%)	Group 2 to Group 1 Differences (%)
Week 2: Phone contacts (1 to 4)	20.7		First priority mailing	23.5	–2.8
Week 3: Phone contact (5 to 10)	37.3		Postcard reminder	46.3	–9.0
Week 6: Phone contact (11 to 28)	41.4		Second first-class mailing	62.3	–20.9
Week 8: Priority mail follow-up	58.6		Phone contacts (1 to 10)	66.8	–8.2
Week 24: Priority mail follow-up	73.2		Priority mail follow-up	76.8	–3.6
	χ^2	p-Value[b]			
Week 6: Response rate	87.3	.001			
Week 8: Response rate	17.4	.001			
Week 24: Response rate	4.72	.039			

[a] Response rate: CASRO response rate.
[b] Significant difference between group 1 and group 2 response rate as measured by χ^2.

Table 2. Cumulative variable cost per stage of follow-up and experimental group

	Experimental Group 1: Phone Start (N = 1,235)		Experimental Group 2:Priority Mail Start (N = 1,237)	
Timing	Protocol Stage	Cost ($) to Complete	Protocol Stage	Cost ($) to Complete
Week 2	Phone contacts (1 to 4 attempts)	128.33	First priority mailing	60.33
Week 4	Phone contact (5 to 10 attempts)	74.13	Postcard reminder	30.69
Week 6	Phone contact (11 to 28 attempts)	76.68	Second first-class mailing	26.55
Week 8	Priority mail follow-up	56.98	Phone contacts (1 to 10)	27.36
Week 24	Priority mail follow-up	48.98	Priority mail follow-up	31.53
Average cost over entire study: $36.61				

Use of financial incentives has been demonstrated in previous research to increase response rates to mail or telephone surveys of physicians even further. However, government agencies are often not able to incorporate such a strategy due to limitations in funding and the fear of public and organizational perceptions that they are wasting valuable resources.

In this study of a state population of family practice and pediatric physicians, even though the sample frame source was considered to be credible as a complete population listing of family practice and pediatric physicians, it was found that contact information for the respondent was deficient, incomplete, or wrong, and that extra effort was required to improve the quality of addresses and telephone numbers. Improvements in provider databases might be one of the best options for reducing sample nonresponse bias. Approximately 1,022 additional phone numbers for physicians were obtained through the use of multiple sources. As shown in this study, incorporation of procedures to improve contact information by accessing multiple sources of information for the same population, and adapting the telephone contact script such that the first level of contact screens for eligibility and updates case information, improves chances of contacting physicians and keeping observations in the study.

U.S. Postal Service 2-day priority mail in conjunction with telephone appears to be effective in increasing response among family practice and pediatric physicians. Given an 8- to 10-week data collection period and a multimode sequence of survey contacts incorporating 2-day priority mail, a response rate near 60% should be reached without the need for financial incentives. Unlike the manufacturing study reported by Petrie et al. (1997), the reversal of mode sequence utilizing telephone as the first and primary data collection mode reduced response for physicians. In the current study, U.S. 2-day priority mail as the first and primary mode sequence was significantly associated with higher response, and it also accelerated the rate of reply after the second week. Telephone contact in the first week is associated with about a 2% lead in determination of a final disposition (ineligible and refusals) for those physicians replying at that point in time. However, refusals were mostly obtained during the phone contact phase on both sides of the experiment, which suggests self-selection out of the survey by other practice employees. It can easily be seen from this study that a change to an alternative mode for follow-up is a valuable strategy and is recommended for interviewing physicians. For the phone start treatment (group 1), if the study had terminated after 28 phone attempts, the rate of reply would have been

47%, and the CASRO response rate would have been 41%. Thus the switch to U.S. 2-day priority mail follow-up resulted in an additional 18% reply level (final disposition as completed interview, a refusal, or ineligible) and 17% response rate. For the mail start treatment (group 2), the phone follow-up resulted in an additional 13.1% reply and 8.6% response. Switching modes has the consequence of increasing costs, especially if the mode is more costly to implement, such as a phone interview for collecting data. As more weeks of data collection and survey protocols are added, the cost per completed interview decreases as more completed interviews are achieved (see Table 2). At week 8 and week 24, for the mail side of the experiment, costs start to climb after the third stage of contact. This means that the marginal cost (the total variable cost associated with a survey protocol divided by the number of completed interviews at that stage) of a completed interview is now increasing instead of decreasing. This suggests that any more contact will significantly add to study costs.

One of the most surprising results is the yield in response gained by extending the data collection period three months. This step alone added 14% response to group 1 and 10% to group 2. This final, fifth contact almost brought both sides of the experiment even for the number of completed interviews. Overall, the survey results add to what is known about non-monetary survey protocol strategies, the impact of alternating survey modes as a way to optimize contact with physicians, and survey costs. It also suggests that one way to evaluate the value of tailored survey design or mode sequencing is to use an economic production approach and simultaneously look at efficiencies of response, declining variable cost, and the up turn of the marginal cost curve.

REFERENCES

Del Valle, M. L., Morgenstern, H., Rogstad, T., Albright, C., & Vickrey, B. G. (1997). A randomized trial of the impact of certified mail on response rate to a physician survey and a cost effectiveness analysis. *Evaluations & the Health Professions, 20*(4):389–406.

Frankel, L. (1983). The report of the CASRO task force on response rates. In F. Wiseman (Ed.), *Improving data quality in a sample survey.* Cambridge, MA: Marketing Science Institute.

Gullen, W., & Garrison, G. (1973). Factors influencing physicians' response to mailed questionnaires. *Health Services Reports, 88,* 510–514.

Maheux, B., Legault, C., & Lambert, J. (1989). Increasing response rates in physicians' mail surveys: An experimental study. *American Journal of Public Health, 79*(50), 638–639.

Paxson, M. C., Dillman, D., & Tarnai, J. (1995). Improving response to business mail surveys. In B. Cox, D. Binder, B. N. Chinnappa, A. Christianson, M. Colledge, & P. Kott (Eds.), *Business survey methods* (pp. 303–316). New York: Wiley.

Petrie, R., Moore, D., & Dillman, D. (1997). Establishment surveys: The effect of multi-mode sequence on response rate. Unpublished paper of the Social and Economic Sciences Research Center, Washington State University.

Shiono, P. H., & Klebanoff, M. A. (1991). The effect of two mailing strategies on the response to a survey of physicians. *American Journal of Epidemiology, 134*(5), 539–542.

Shosteck, H., & Fairweather, W. (1979). Physician response rates to mail and person interview surveys. *Public Opinion Quarterly, 43,* 206–217.

Sudman, S. (1985). Mail surveys of reluctant professionals. *Evaluation Review, 9*(3), 349–360.

Urban, N., Anderson, G., & Tseng, A. (1993). Effects on response rates and costs of stamps vs. business reply in a mail survey of physicians. *Journal of Clinical Epidemiology, 46*(5), 455–459.

Discussion of Papers on Mode Effects

Norman M. Bradburn

In the beginning there were two modes: face-to-face interviewing and self-administered questionnaires. Then God created the telephone, and a new day dawned. Simple telephone interviewing was quickly followed by computer-assisted telephone interviewing (CATI), which became so dominant in our thought, if not in practice, that many people forgot there was a distinction between the two. Before these innovations could be completely digested, there came in rapid succession computer-assisted personal interviewing (CAPI), computer-assisted self-administration (CASI), and its techno-glitzy offspring, audio-CASI. Now God is spinning the Web to entice us into new, little-known, and rapidly changing but irresistibly alluring territory. For a methodologically conservative researcher such as myself, this is a dangerous trend.

The organizers of this conference have invited the discussants to take a broad view of the topic and focus more on the general issues than on critiques of the individual papers. Without intending any discourtesy to the authors of the papers in this session, I would like to step back and reflect on the issues that need to be addressed in considering mode effects. Why, indeed, should we worry about mode effects? The simple answer is that in many instances we want to compare results from studies that have used different modes of data collection and, thus, we want to be reasonably confident that differences in results are not due to some modal artifact.

To some extent, this answer begs the question. Why should we expect a mode to have an effect? Or, more precisely, which modes (do we hypothesize) have what kinds of effects on what kinds of data? Many studies of mode effects have looked at a wide range of possible effects without any theoretical framework that might guide the research. Green and Krosnick note—referring to the bulk of the studies they have reviewed on differences between telephone and face-to-face interviewing—that "these studies have for the most part been atheoretical, looking for potential differences between modes with little conceptual guidance about what differences might be expected and why." Refreshingly, by contrast, several studies presented today lay out a series of theory-based hypotheses about possible mode effects.

What are the kinds of effects that one should worry about? I will discuss possible effects under three rubrics: (1) differences in the sample due to coverage or due to differential response rates, (2) differences in responses due to factors intrinsic to the mode of data collection, and (3) differences due to the social context within which the data are collected.

Factors Related to the Sample

The most obvious difference between telephone and household-based interviewing is in population coverage. Households without telephones cannot fall into the sample unless there is some mixed-mode provision to pick up the non-telephone households. Although there were some very early uses of telephone surveys (for example, I found that NORC did some telephone surveys during World War II), this obvious difference in coverage inhibited the development of telephone interviewing until a very high proportion of households had telephones. Surprisingly, the still-low proportion of households connected to the Internet does not seem to have the same inhibiting effect on the rush to do Web-based surveys. The lessons of the *Literary Digest,* which used sampling frames based partially on telephone and automobile ownership when these were far from being universal, have apparently been lost on many in the contemporary survey industry.

While not an effect intrinsic to the mode, differences in response rates are common between modes. Green and Krosnick, Moore et al., and Midanik, Rogers, and Greenfield document the common biases in sample surveys conducted by telephone and face-to-face. Even in the census, the mail-back rate from households produces severe biases.

While not a mode study, strictly speaking, the feasibility study by Osmond et al. demonstrates the problems that arise from differential response rates. Theirs is a particularly difficult problem because the probability of having the characteristic of interest is strongly correlated with the propensity to respond to the survey. If one wants to estimate rates for a disease and depends on different techniques for measuring prevalence, then one needs good quantitative estimates of the correlation between the measuring instrument and the propensity to respond in order not to underestimate the prevalence.

Factors Intrinsic to Modes of Data Collection

There are factors intrinsic to mode that either prevent some types of questions being used altogether in one mode, or that interact with modal characteristics to produce response differences. The most obvious factor is the inability to use visual materials in telephone interviews. A consequence of this limitation is that showcards cannot be used

either for complex response categories or as memory aids. This rules out such things as lists of organizations or magazines used in studies of organizational membership or readership, or pictures of medications in surveys of prescription drug usage. Web- and computer-assisted interviewing offer some very interesting potential for more innovative stimuli, such as pictures with motion and sound, but these have not been much exploited as of yet.

Less obvious are interactions between mode of presentation and response bias. Laboratory studies have established that serial order effects are different for different sensory modalities. Stimuli presented in a visual mode are more susceptible to primacy effects; stimuli presented in auditory mode are more susceptible to recency effects. Survey questions frequently offer a list of alternatives for respondents to choose from, and there is ample evidence to support the view that this interaction can cause significant modal differences in responses to questions involving lists for respondents to choose from or in multiple response categories (Sudman, Bradburn, & Schwarz, 1995).

Order effects are difficult to overcome in any form of interviewing that depends on verbal interaction because, whether over the telephone or face-to-face, questions are asked one at a time in a specific order. Paper-and-pencil self-administered questionnaires may be superior in this regard because respondents can see all of the questions before they answer any of them, and, as a result, question and response alternative order effects are lower in mail questionnaires. This effect might be the reason for the difference in measurement properties of the scales studied by Rockwood, Kane, and Lowry. One would have to know more about the relation of the order of the items in the scale and the factor structures found using the different modes to tell whether this is a viable hypothesis. Recency effects in the telephone mode appear to have been found in that study.

The order effect benefit in self-administered questionnaires, however, would not hold true for CASI because the computer generally presents the questions one at a time, and creates the same order effect that would be true with visually presented stimuli. Relatively little attention had been given to effects intrinsic to computer-assisted interviewing, perhaps because researchers are so focused on the effects of modal differences on respondents that they forget that there may be modal effects on the interviewers. In particular, different computer assistance programs have different functionality for the interviewer compared with paper-and-pencil questionnaires. For example, most computer assistance programs show only one question at a time on the screen. After the answer is entered, the question disappears and the next appropriate question appears automatically. Programs differ in the ease with which interviewers can go back to previous questions and what they can do to change answers in light of subsequent questions. In paper-and-pencil questionnaires, the interviewer typically sees a number of questions on at least two pages in a booklet and can easily flip back to previous questions. The interviewer has a greater feel for the totality of the questionnaire and the relation of questions to one another. This can be an advantage when there is a series of related questions such that respondents may anticipate questions and answer them before they are actually asked. In such situations, interviewers frequently enter the answers when given because they know where the question comes and they can find it easily. Thus they avoid having to ask the question a second time when they come to it. With CAPI, however, they cannot easily find the right place and have to stick to the rigid question order as it appears on the screen. Since interviewers also do not have to worry about the logic of skip patterns with CAPI, they concentrate on the individual questions and answers and may lose sight of the overall coherence of the answers. I do not believe that enough attention has been paid to the effects of different ways of constructing computer assistance programs on interviewer behavior.

Early work on Web interviewing suggests that the technical difficulties are more formidable than one would imagine given the technology. One unexpected finding is that different browsers may display the questions in different ways even though the programs are the same. Thus what the researchers think of as a constant visual presentation may in fact appear in quite different ways to respondents who are accessing the questionnaire through different browsers. Some of the differences affect the placement of the stimuli on the screen in ways that are known to produce order effects in paper-and-pencil questionnaires, such as the distance between response categories and the placement of questions or response alternatives on a page. In this regard, I think Miller vastly underestimates the amount of work and testing, and hence the cost savings, needed to construct a good Web-based survey instrument.

Factors Related to Social Context

The third major factor that differs among modes is the kind and quality of social interaction involved in the data collection. Face-to-face interviewing involves the closest contact between interviewer and respondent, with telephone being somewhat more distant and the self-administered mode being the most remote. The development of CAPI, and the use of CASI and especially audio-CASI within a CAPI survey, have reduced the social desirability bias of the traditional face-to-face interview. Early on, even before the various CASI emendations, CAPI was seen by respondents as more anonymous and offering greater protection of confidentiality, even though they could not know all the reasons that this was in fact true (Baker, Bradburn & Johnson, 1995).

A long line of research has shown that self-administered questionnaires reduce social desirability bias and produce more reports of negative behavior and unpopular opinions and fewer reports of positive behavior and socially approved opinions. The effect of telephone interviewing has not been consistent. Some earlier studies found the effects of the telephone mode to be between those of the face-to-face and self-administered modes in terms of the respondent propensity to give socially desirable results (Sudman & Bradburn, 1974). The studies reported here do not follow this pattern. Green and Krosnick found the telephone to produce slightly more

socially desirable attitudes and behavioral reports. Rockwood, Kane, and Lowry found the expected effect with mail questionnaires as compared with telephone. Midanik and her colleagues found few differences in mode that could not be accounted for by sample differences. What differences they did find indicated less social desirability bias in the face-to-face interviews. It is clear that we have more work to do before we understand the effects of telephone interviewing on social desirability bias.

It is difficult to predict what may happen with the Web survey. While it looks as if it should follow the pattern of the self-administered questionnaire, significant concerns about confidentiality in electronic transmissions may make respondents cautious about admitting to socially undesirable behavior or attitudes. The more one hears about the ease with which electronic transmissions can be traced to their source and the various warnings about insecure transmissions on the Internet, the more wary respondents may become about honestly answering questions. We need to know more about what respondents believe regarding our ability to preserve confidentiality in Web surveys before we can confidently predict what mode effects might occur.

Telephone interviews tend to be shorter than face-to-face interviews, and telephone interviewers do not have the nonverbal cues available to face-to-face interviewers to let them know when respondents may not understand a question or may have more to say. Thus, telephone interviews may be subject to more satisficing and result in less complete or well-thought-out answers. Green and Krosnick find some evidence for this hypothesis.

Conclusion

I suspect that surveys in the future will often be mixed-mode in their execution. Thus, we need to be able not only to show where there are mode effects but also to quantify their magnitude, so that the effects can be taken into account in the analysis. If there are known mode effects, combing data based on different modes of data collection will require some adjustment to take the mode effects into account. Estimating the size of these effects and developing models to handle these adjustments are major challenges for future research.

References

Baker, R., Bradburn, N. M., & Johnson, R. (1995) Computer-assisted personal interviewing: An experimental evaluation of data quality and costs. *Journal of Official Statistics, 11*(4), 394–415.

Sudman, S., & Bradburn, N. (1974). *Response effects in survey: A review and synthesis.* Chicago: Aldine.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1995). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Comparability of Data across Different Modes of Data Collection

Colm O'Muircheartaigh

General Framework

In any important area, whether of substantive science or of methodology, we need more than one source of information to be convinced of the truth or nontruth of a proposition. When we are in possession of those multiple items of information, we need a means of assimilating them in order to arrive at a conclusion. Among the key questions are the following: How do we evaluate the alternative sources? How do we combine them? If there should be only one, is it good enough to form the basis for a conclusion on its own? It is possible that the different sources will make contributions of different kinds to our understanding. This understanding itself may be multidimensional, and the different sources may each contribute to a different dimension.

The best hope for a superior understanding is in the accumulation of information. As our knowledge increases, it may be that our increased understanding arises from an appreciation of the specificity of our conclusions rather than their generalizability. It is possible in particular that the framework we have been using is not the most useful one for synthesizing our evidence.

This relates to, although it is not identical to, the issue of internal versus external validity. The title of this session focuses on differences between modes of data collection—a suitable comparison for a methodologist. The conference as a whole addresses issues of health survey research methods—on some dimensions a much broader field, on others narrower. In this discussion I would like to illustrate how the perspective we take may affect the conclusions we draw from the evidence presented.

For each of the papers in the session we should ask, What is its objective? Is it to find support for, or to test, a theory? Is it to identify a potential problem? Is it to find a solution to a problem? Furthermore, we should ask whether the evidence it produces is general or specific, and if specific, whether it is specific in terms of time, the context, or the particular survey from which it is drawn.

Below I categorize the six papers in the session according to a number of different criteria, the kinds a methodologist might use to label a piece of research. At the end I make some general remarks about the set of papers as a whole.

The author is at NORC, University of Chicago.

Methodology

Table 1 describes the papers according to the methodological issues they address. Two of them address what became an important issue for survey researchers in the 1980s, and remains so today—the conversion of an existing survey from face-to-face interviewing to telephone. Green and Krosnick use data from the 1982 National Election Studies to compare data quality for the same variables in parallel implementation using the two modes. Midanik, Rogers, and Greenfield use data from the U.S. National Alcohol Survey (NAS) from 1995 to estimate the impact on reports of alcohol consumption of the conversion from face-to-face to telephone this year. If the conversion were to affect estimates of consumption level, this would have serious implications for the use of NAS data for trend analysis.

Three other papers address less common mode-related issues. For Rockwood, Kane, and Lowry, the mode difference is that between telephone and self-completion, and the emphasis is on the psychometric properties of quality-of-life scales. Moore, Gaudino, deHart, Cheadle, and Martin contrast two sequences of modes—rather than two single-mode approaches—in surveys of physicians in Washington State. Osmond, Catania, Pollack, Canchola, Jaffe, MacKellar, and Valleroy are concerned with establishing whether or not a home collection test kit—a particular type of self-administration—is a feasible approach to biological sample collection in an HIV study as the final stage of a multimode recruitment strategy. Finally, the paper by Miller discusses the methodological implications of conducting Web-based research, mainly from the perspective of the Web as a substitute for lab-based data collection rather than in the context of large-scale survey research.

Research Design

The categorization in Table 2 looks at the papers from a different perspective—that of research design. If control of possible sources of confounding were our primary concern, only two studies—Green and Krosnick, and Moore et al.—provide any control through randomization. The Midanik et al. and Rockwood et al. papers must call on other assumptions to generalize their results beyond the cases in their samples. For Midanik et al., a possible issue is whether a follow-up

study yields results generalizable to independent administration; for Rockwood et al. a possible issue is whether the groups are sufficiently comparable. For Osmond et al., as the focus of the study is feasibility, establishing feasibility may be a first step toward the (later) objective of estimating the extent of feasibility in the relevant population as a whole.

I do not raise these topics as a criticism of any of the papers. For Rockwood et al. we might well argue that representation is less important in looking at structural properties. For Midanik et al., we might argue that there is no reason to believe that the earlier administration would contaminate the later. I wish simply to emphasize that there are many ways in which information may be gathered, and each of them is vulnerable to criticisms of different kinds. Research design (and the control of confounding variables) is only one of these dimensions.

Sample Size

A third categorization of the papers is given in Table 3. Here I use sample size, a favorite of statisticians. The range is substantial—from a total of 161 for Rockwood through 600 for Osmond, 1,000 for Midanik, and 2,000 or more for Green and Krosnick and for Moore. These samples have different implications not only in terms of power, but also in terms of representativeness, ranging from complete population coverage for Moore, with nonresponse as the dependent variable, through the very small samples used by Rockwood to compare structural effects.

Statistical Analysis

Finally, Table 4 shows a categorization by method of statistical analysis. This table shows considerable diversity also, with the emphasis on a direct comparison of proportions in two of the papers, explicit modeling used for the comparison in two others, and essentially the comparison of models in one.

General

The papers in this session address important themes; there is some commonality among the papers, but the motivations and the approaches taken differ substantially. The authors should be congratulated on the systematic way in which they tackled relevant practical problems and should not be criticized for failing to tackle problems they did not intend to address.

The terminology used in comparing modes of data collection, or indeed in evaluating any aspect of data quality, varies across disciplines. The statistician and the sociologist tend to use terms such as variance and bias to describe the quality of crucial outcomes of the research; the psychologist will tend to describe the same (or similar) aspects of data quality with such terms as reliability and validity. One welcome development is that survey methodologists are now more likely to use explicit modeling in their data analysis. In due course data analysts may also modify their methods to take into account the complexity of sample design.

Table 1. A methodological categorization of the papers

Green and Krosnick	Conversion of face-to-face to telephone (1982)
Midanik et al.	Conversion of face-to-face to telephone (1995)
Rockwood et al.	Mode comparison: telephone and self-completion
Moore et al.	Mode sequences: telephone and mail
Osmond et al.	Acceptability of mode: self-completion after telephone
Miller	Web-based general issues

Table 2. Categorization by design type

Green and Krosnick	2/3 group randomized
Midanik et al.	1-group follow-up comparison; 8-month interval
Rockwood et al.	2-group nonrandomized
Moore et al.	2-group randomized
Osmond et al.	1-group follow-up; random base
Miller	Various

Table 3. Categorization by sample size

Green and Krosnick	998 tel., 1,418 f/f
Midanik et al.	1,047 original (f/f) and follow-up (tel.)
Rockwood et al.	118 self-completion/47 telephone
Moore et al.	split-half whole population: 2,472
Osmond et al.	615 derived from 2,881 MSMs
Miller	Various

Table 4. Categorization by method of statistical analysis

Green and Krosnick	OLS regression
Midanik et al.	Logistic regression
Rockwood et al.	χ^2
Moore et al.	χ^2 or *t*-test
Osmond et al.	Comparison of proportions
Miller	Various

One general deficiency of the papers is the absence of explicit cost data, though Moore does provide some data on this. In some cases—as in the mode conversion papers—the (not unreasonable) assumption was made that the conversion was dictated by overwhelming cost issues. However, if we are to make substantial progress in methodology, we need to be systematic in trading off gains in precision and reductions in bias against increases in costs. Otherwise there will be no meaningful dialogue between methodologists and the more practically oriented data collection operations staff.

It may have been noticed by some that nowhere in this discussion have I made any substantive reference to the subject

matter of the surveys. This is another danger in conversations among methodologists (among whom I am generally pleased to be numbered). In the health field especially, and perhaps in all fields, it may be that we should be more sensitive to substantive concerns, acknowledge the specificity of different subject areas, and try to involve the clients and the end users of the data in the methodological experimentation we design and conduct. This would argue for a different grouping of papers than that in this session, and it would certainly provide a different basis for assessing their contributions.

Discussion Notes, Session 3

Mary Grace Kovar and Judith Lessler, Rapporteurs

Floor discussion of the session on mode effects touched on the following points:

- Importance of the differences in terms of the substantive and policy conclusions
- Mixed-mode designs and their use in future studies
- The need to consider both coverage and measurement issues
- Future of web-based surveys
- Future research issues

Policy Importance versus Significant Differences

Most of the mode effects discussed, while statistically significant, are small. We need to consider whether these statistically significant effects have important policy implications. In making this assessment, we should consider the situations that would produce the largest mode effects and determine if they have policy significance. If they do, then we should be concerned; otherwise, mode may not be an issue.

Mixed-Mode Designs

Mixed-mode designs are likely to be the wave of the future. Several issues arise in relation to their use. The best mix of modes and the order of their use will depend on characteristics of the population. If mixed modes are being used, guidelines must be developed as to the optimal time to switch from one mode to another based on cost and accuracy considerations. In some cases, the best choice is to offer the respondents a choice of modes. If this is done, it is necessary to retain information on their choices; mode effects may be different if they are due to respondent choice rather than investigator assignment. In addition, modes used for initial response may differ from modes used for tracing. For example, e-mail may be used to track respondents and encourage them to respond in another mode.

Coverage and Measurement

Estimation of mode effects must include both measurement and coverage issues, and document the interaction between them. Net effects do not separate these two and do not necessarily provide guidance for how to make changes. For example, the telephone mode may place higher cognitive demands on respondents, and mail questionnaires have high literacy demands. Respondents with different types of characteristics—lower education, language difficulties, and so forth—may be differentially affected by these increased cognitive demands. These may be the same respondents who are poorly covered in telephone surveys.

Web-Based Surveys

This is an exciting new technique that is not yet viable for general population-based surveys because not enough is known about coverage in the general population. It is already being used for special closed populations, for example, the employees of a particular company. Discussants wondered whether or not web-based response will track self-administered questionnaires that show enhanced reporting of sensitive issues due to increased privacy, or whether data security concerns will cause effects similar to interviewer-administered questionnaires.

Research-Recommendations

Discussants mentioned a number of additional future research needs:

- Studies of interviewer behavior should be included in studies of mode differences.
- Studies need to be expanded to include more modes and use of multiple modes.
- We need to distinguish “reluctance to report” and “forgetting,” which may be particularly prevalent in a telephone survey. There is an assumption that forgetting results in underreporting. It was noted that providing appropriate recall cues can reduce both underreporting and overreporting.

Validity of Results

All six previous conferences have addressed the issue of validity of survey data, and the seventh continued that trend. Three of the six featured papers in this session (Pascale; Harter, Kuby, and Moore; Blumberg and Cynamon) deal explicitly with the validity of reporting of health insurance information in surveys, using a variety of methodological approaches (both within and across these papers), including detailed comparisons of measurement approaches across different surveys, comparisons with administrative records, and methodological field tests of alternative measurement approaches. In combination, they provide a fairly detailed and sophisticated picture of how complicated it is to collect accurate data on this obviously quite critical measure, and why estimates of the numbers of uninsured or those on Medicaid can differ so greatly across surveys and among various data sources.

The other three papers do not focus on health insurance, but two use a fairly classic records-check approach to evaluate the validity of respondent reports—one to "calibrate" a survey battery designed to identify children with special needs (Fowler, Gallagher, and Homer) and the other to demonstrate the value of using behavior coding to diagnose and improve survey invalidity (Belli, Lepkowski, and Kabeto). The final paper (Burt) uses contextual records data at an aggregate, geographic level to assess the validity of data reported by physicians' offices in the National Ambulatory Medical Care Survey (NAMCS).

Methodological Issues in Measuring the Uninsured

Joanne Pascale

Introduction

The health care system in the United States has undergone massive and rapid changes over the past decade. While these changes have profoundly affected the way health care is delivered and accessed, the changes have also presented major measurement challenges to survey researchers. National surveys on health insurance coverage over the past few years produced estimates of the uninsured that range from a low of about 8% up to a high of almost 18% (Bennefield, 1996; Lewis, Ellwood, & Czajka, 1998; Rosenbach & Lewis, 1998). The inconsistency in these estimates is taking on particular importance as the official rate of uninsured continues to rise through the 1990s (U.S. Census Bureau, 1997), new federal and state health insurance initiatives are introduced, and researchers and policymakers struggle to make informed decisions regarding the distribution of resources targeted for the uninsured and underinsured populations. In spite of these new challenges and demands, however, some commonly employed methods of measuring the uninsured are not yet well understood. This paper will present an analysis of three particular survey design features that are integral to major national surveys used to measure health insurance coverage: reference period, unit of measurement, and verification of the uninsured. The goal of this research is not to suggest the "real" number of uninsured but rather to enable a better interpretation of national surveys that rely heavily on these design features, and to inform future research that sets out to measure the uninsured. The analysis will use data from the Census Bureau's 1999 Questionnaire Design Experimental Research Survey (QDERS), which was designed explicitly for experimental testing of alternative survey design features.

Within the health survey research literature, the reference period is intertwined with the definition of the uninsured. Some surveys (e.g., the Current Population Survey, or CPS) ask respondents in March whether they had health coverage at any time during the previous calendar year, and the uninsured are defined as those lacking coverage throughout the entire year. Other surveys (e.g., the Community Tracking Study Household Survey, or CTS) ask respondents if they are currently covered, and the uninsured are defined as those lacking coverage at a particular point in time. While these two surveys are clearly measuring two different concepts, and one would expect the point-in-time estimate of the uninsured to be higher than the estimate of those uninsured throughout the calendar year, the estimates derived from these surveys present a confusing picture: they are roughly the same (Rosenbach & Lewis, 1998). Still other surveys (e.g., the National Medical Care Expenditure Survey [NMCES] and the Survey of Income and Program Participation [SIPP]) ask respondents if they were insured during the previous three or four months (respectively), and data can be aggregated to represent a full calendar year. One might expect these calendar- year estimates of the uninsured to come close to the CPS estimate. However, research has shown the difference to be striking: The CPS estimate of the uninsured is one and a half times higher than the NMCES estimate (Schwartz, 1986) and two times higher than the SIPP estimate (Bennefield, 1996).

Meaningful comparisons of estimates across surveys, however, are threatened by a host of differences in design features across surveys—such as question wording and sequencing, skip patterns, interviewer training, survey context and sponsor, sample design, weighting and imputation—any of which could be responsible for observed differences or similarities in the estimates. Some researchers (Schwartz, 1996; Rosenbach & Lewis, 1998) have attempted to control for these design differences post-hoc by isolating particular features (e.g., sample design, weighting and imputation, instrumentation), scrutinizing their similarities and differences across surveys, and/or manipulating them in analysis to gauge their impact on the estimates. While this type of research is informative, it is virtually impossible to control all of the subtle design differences across surveys, and without that control only tentative conclusions can be made. One such conclusion often cited in the literature concerns the reference period. Many researchers believe that a major source of measurement error in health insurance estimates is the previous-calendar-year reference period, because estimates of the uninsured are very similar whether derived from a survey that asks about point-in-time coverage status or one requesting previous-calendar-year status. Researchers speculate that respondents either don't attend to the calendar-year reference period and instead simply report current coverage status, or they fail to recall coverage they may have had during part of the 12-month reference period. Either of these behaviors could lead to underreporting of coverage during the previous calendar year. In order to bring some evidence to bear on this hypothesis, the QDERS survey was designed to experimentally manipulate the reference period while controlling for all other survey design features—half the QDERS respondents were

The author is at the Center for Survey Methods Research, U.S. Bureau of the Census.

asked about current coverage, and the other half were asked about coverage during the previous calendar year.

A second design feature, unit of measurement (person vs. household level), has received little attention in the survey methodology literature generally, but could have implications for the accuracy of health insurance estimates, respondent burden, and nonresponse. In many households all or most members are covered under the same health plan. For these households, asking about coverage person by person is a lengthy undertaking and can be tedious if one respondent has to proxy for other household members. An alternative, more efficient design for many households—employed by several major national surveys—uses a household screener approach. Specifically, a household informant is asked if anyone in the household is covered by various types of plans and, if the answer is yes, names of individuals covered by each plan type are collected. This design generally shortens the survey in households where multiple people are covered by the same plan, which reduces burden, length, and general tedium. Little is known, however, about how the estimates derived from each of these methods compare. In QDERS, half of the sample households were asked health insurance coverage questions at the household level, and half received person-level questions. Each of these sample halves was then further divided between the current and calendar-year versions (as described above). That is, altogether the QDERS experimental survey involved four different questionnaire treatments: point-in-time coverage (at the household and person levels) and previous-calendar-year coverage (also at the household and person levels; see Exhibit 1). Analysis will focus on a comparison of the estimates of the uninsured derived under each of these four questionnaire treatments.

Finally, more recent health survey research has focused on a design feature that has not yet been extensively studied: the method of identifying the uninsured. Until recently, most major health surveys identified the uninsured by asking a series of questions about different types of insurance and calculating the uninsured as those who say no to all types of insurance. But several recent surveys (e.g., the 1996 CTS, the 1997 Maine Health Insurance Survey, and the 1998 North Dakota Health Insurance Survey) have shown that including a direct question to verify the uninsured captures a nontrivial number of people who are actually insured but failed to report their coverage in the initial series of questions on plan types. The percentage of sample members who appear to be uninsured and for whom coverage is reported only in response to the verification question varies across surveys—from 6% to almost 13% (Rosenbach & Lewis, 1998; Carlson, 1998). However, thus far the verification question has been examined only in the context of household-level point-in-time coverage. That is, there appears to be no evidence yet on the number of insured people that would be captured by the verification question if it followed a series of calendar-year health coverage questions, or a series of person-level questions on coverage. This research will present an analysis of the effectiveness of the verification question (see Exhibit 2) under each of the four different questionnaire treatments (displayed in Exhibit 1).

Methods

Overview

In the spring of 1999, the Census Bureau fielded its first survey designed for the sole purpose of experimentally testing survey methods for general research purposes (vs. survey-specific applications). The 15-minute telephone survey included questions on four main topic areas: demographics, disability, health insurance, and nonwage income receipt (e.g., Social Security income from assets). A single household respondent was asked to report for him/herself and up to five other household members. Four independent random-digit-dial (RDD) samples were drawn (one for each of the four different questionnaire treatments), and each sample was designed to be nationally representative of the civilian noninstitutionalized U.S. population (excluding Alaska and Hawaii) living in households with telephones. The survey

Exhibit 1. General question wording and cell sizes across treatments

	Person Level	Household Level
Current coverage	The next questions are about health insurance coverage. **Are you** covered by [plan type]?	The next questions are about health insurance coverage. **The questions apply to ALL persons of ALL ages. Is anyone in this household** covered by [plan type]?
	IF YES: Who is the policyholder for this plan?	IF YES: Who in this household is a policyholder? Anyone else? In addition to you, who else in this household is covered by your plan? Anyone else?
	n = 324 households (818 people)	*n* = 332 households (869 people)
Calendar-year coverage	The next questions are about health insurance coverage **during the calendar year 1998. At any time during 1998 were you** covered by [plan type]?	The next questions are about health insurance coverage **during the calendar year 1998. The questions apply to ALL persons of ALL ages. At any time during 1998 was anyone in this household** covered by [plan type]?
	IF YES: Who was the policyholder for this plan?	IF YES: Who in this household is a policyholder? Anyone else? In addition to you, who else in this household was covered by your plan? Anyone else?
	n = 307 households (765 people)	*n* = 316 households (776 people)

Exhibit 2. Verification question wording across treatments

	Person- and Household-Level Questionnaires
Current coverage	I have recorded that (you do/NAME does) not have health care coverage of any kind. (Do you/Does NAME) have health insurance or coverage through a plan I might have missed?
Calendar-year coverage	I have recorded that (you/NAME) did not have health care coverage of any kind at any time during 1998. Did (you/NAME) have health insurance or coverage through a plan I might have missed?

was conducted by 22 Census Bureau interviewers with a range of experience at the Hagerstown, Maryland, telephone facility from April 26 to May 15. Initial training was held just prior to interviewing and lasted 5–6 hours. The response rate varied, depending on the treatment of cases of unknown eligibility. According to AAPOR guidelines (American Association for Public Opinion Research, 1998) the near minimum response rate (i.e., including partial interviews as respondents and including all cases of unknown eligibility in the base) was 36%, and the "maximum" response rate (again including partial interviews as respondents but *excluding* all cases of unknown eligibility from the base) was 46%. When eligible noncontact cases were excluded from this same base, the cooperation rate was 52%. No refusal conversion efforts were made, and refusals accounted for about half of the nonresponse. The final number of households that completed interviews was 1,304, covering 3,288 people. Due to missing data, however, only 1,291 households, covering 3,228 people, were available for this analysis.

Exhibit 3. Listing of person-level current coverage questions

1. The next questions are about health insurance coverage. (Are you/Is name) covered by a health insurance plan provided through a current or former employer or union?
   PROBE: Include COBRA and health insurance plans provided by colleges and universities to students.
   PROBE: Do not include military health insurance here; that will be covered later in another question.
   1A. [IF YES] Who is the policyholder for this plan?
2. (Are you/Is name) (also) covered by a health plan that was PURCHASED DIRECTLY, that is, not related to current or past employment?
   PROBE: Include insurance plans purchased through a professional association or trade group.
   PROBE: Do not include military health insurance here; that will be covered later in another question.
   2A. [IF YES] Who is the policyholder for this plan?
3. (Are you/Is name) (also) covered by the health plan of someone who does not live in this household?
4. (Are you/Is name) (also) covered by Medicare?
   PROBE: Medicare is the health insurance for persons 65 years old and older or persons with certain disabilities.
5. (Are you/Is name) (also) covered by Medicaid or any other type of government assistance program that pays for health care?
6. (Are you/Is name) (also) currently covered by CHAMPUS, CHAMPVA, Tricare, VA, military health care, or Indian Health Service?
   6A. [IF YES] Which plan (are you/is name) covered by?
7. (Are you/Is name) (also) covered by any (other) type of health plan?
   7A. [IF YES] Which type of insurance (do you/does name) have?

Instrumentation

In keeping with the general structure of most national surveys on health insurance coverage, respondents were asked about coverage through seven main sources of health insurance: employers/unions, direct purchase, someone outside the household, Medicare, Medicaid, military or Indian Health Service, or any other plan (see Exhibit 3 for a display of the exact question wording in the current person-level questionnaire version).

Interviewer Pools and Training

In an attempt to avoid contamination and interviewer conditioning effects, interviewers were divided into two pools. One pool was trained on the person-level questionnaires and the other was trained on the household-level questionnaires. Within each pool, interviewers alternated their work on any given shift between the current and calendar-year versions of the questionnaire. Halfway through the field period interviewers switched treatments (from person to household level, or vice versa) and a follow-up 2½-hour training was conducted on the treatment that was new to each pool of interviewers. Due to some attrition, only 7 (vs. 11) interviewers who were initially trained on the person-level questionnaires were trained on the household-level questionnaires, and 10 (vs. 11) interviewers who were initially trained on the household version were trained on the person-level questionnaires. After the switch, data collection continued for nine more days. However, all samples were released at the beginning of data collection, and due to the rate at which interviews were completed during the first 11 days, the majority of cases had been completed before interviewers switched treatments.

Treatment of Missing Data

The household-level versions of the questionnaires included a household-level "screener" to determine if anyone in the household was covered by that plan type, and, if so, the screener was followed by person-level questions on who was covered. The paper questionnaire included columns for up to six people in the household, and each of these columns included person-level response categories for "yes," "don't know," and "refused." That is, there was no explicit "no" response category at the person level; missing data on any

one person indicated no coverage of that type. Due to this design, and the very low numbers of "don't know" and "refused" responses in the household-level questionnaires, it was necessary to use dichotomous coding of coverage: A "yes" response indicated coverage, while a "don't know," "refused," or missing response indicated no coverage.

Plan Type Coding

Following six questions on main types of insurance, all questionnaires contained a catch-all question on coverage through "any other" plan, and a follow-up item captured the type of plan. After this catch-all question, for people with no reported coverage a final question verified insurance status by asking if the individual had coverage through a plan that might have been missed earlier, and a follow-up item captured the type of plan. Any plan captured in either the catch-all or verification question was back-coded to its appropriate category for analysis. Furthermore, if individuals were reported to have coverage through multiple different plan types, these data were maintained in analysis. That is, no attempts were made to develop a "plan hierarchy" so that each insured person would be covered by only one "main" plan.

Findings

Overview

Several features of the QDERS data threaten its generalizability to the target population. First, the low response rate could be responsible for high nonresponse error. Second, the RDD sample design, with no coverage for non-telephone households, could underrepresent the uninsured. Furthermore, comparisons between the QDERS data and other national surveys could be problematic because of (among other factors) decisions discussed above regarding treatment of missing data and coding, and because the QDERS data set is unweighted and includes no imputation for missing data. For all these reasons, comparisons of absolute estimates to other national surveys could be misleading. However, of analytical interest in this study are *relative* comparisons of estimates across treatments. An evaluation of the basic demographics across treatments revealed no statistically significant differences, providing some evidence that these relative comparisons may not be severely threatened by nonresponse or coverage error.

In order to assess relative differences across treatments, findings were analyzed under a "more is better" model consistent with much of the survey methods literature on reporting patterns. Specifically, with regard to health insurance, the model assumes that all reports of coverage are valid and that any relative differences in the uninsured rate (within similar definitional categories, i.e., uninsured at a point in time vs. uninsured throughout the year) are due to underreporting. Clearly the assumption that all reported plans are valid could be problematic—respondents may be in fact reporting plans they don't actually have—and record-check and other validation studies could shed some light on any overreporting patterns. However, the QDERS data only enable analysis of one side of the coin (underreporting)—hence the "more is better" model.

Rate of the Uninsured

Current versus Calendar-Year Questionnaire Comparisons

Among the person-level questionnaires, the overall percentage of the uninsured was lower, as expected, in the calendar-year version than in the current version (see Table 1)—6.9% versus 10.3% (a difference that is statistically significant; chi-square = 5.581, p = .018). Again, the calendar-year version is essentially estimating those covered *at any point* during the entire previous calendar year, not just those covered at the *current* point in time. Thus one would expect the calendar-year version to produce more reports of coverage than the current version and to generate a lower rate of uninsured. So the QDERS results seem reasonable: As the reference period moves from date of interview to a full calendar year, sample members' chances of being insured at some point during the year go up, resulting in a calendar-year rate of uninsured that is 3.4 percentage points lower than the rate of uninsured at a point in time.

In the household-level versions of the questionnaires, somewhat surprisingly, there was no difference in the rate of uninsured between the calendar-year and current treatments; the uninsured rate was 12.0% in both versions. It's unclear why the expected difference in calendar-year and current estimates would manifest in the person-level questionnaires but not the household-level questionnaires. However, these household-level results are not entirely inconsistent with the literature. The QDERS calendar-year household-level questionnaire was based on the CPS, and the QDERS current household-level questionnaire was based on the CTS. Rosenbach and Lewis (1998) compared 1996 rates of uninsured from these two surveys and, after adjusting the data for certain known design differences between the two surveys, found the estimates to be quite similar—16.3% (CTS) versus 17.7% (CPS), a difference that is not statistically significant.

Household-Level versus Person-Level Questionnaire Comparisons

In comparing person-level to household-level questionnaires, there is little empirical evidence in the survey method-

Table 1. Rates of uninsured people

	Person	Household	Difference
Current	10.3% (n = 84)	12.0% (n = 104)	1.7%
Calendar year	6.9% (n = 53)	12.0% (n = 93)	5.1%*
Difference	3.4%**	0%	

*Significant at .01 level
**Significant at .05 level

ology literature to guide expectations. Intuitively, we may expect more reports of health plans (and a lower rate of uninsured) in the person-level version than the household-level version because respondents may focus more clearly on individuals within the household when presented with a series of questions explicitly asking about individual household members by name. Indeed, the SIPP uses a person-level approach and results in an estimate of the uninsured that is roughly half that of the CPS, which uses a household-level approach. However, the SIPP also employs only a four-month reference period (compared to the CPS's calendar-year reference period), so enhanced recall on the part of SIPP respondents (not the person-level approach) could be driving the observed difference in estimates.

The QDERS data suggest that, in general, the person-level approach results in improved reporting of health insurance coverage. Among the current questionnaires, the person-level estimate of the uninsured (10.3%) is somewhat lower than the household-level estimate (12.0%), but the difference (1.7%) is not statistically significant. Within the calendar-year versions, however, the difference is much larger in magnitude—6.9% versus 12.0%—a difference of 5.1 percentage points that is statistically significant (chi-square = 11.484, p = .001).

Effectiveness of the Verification Question

Evaluation of the verification question in the QDERS survey is compromised because the question was not properly administered in many cases. Given the paper-and-pencil design, and the complexity of the health insurance questions in general, interviewers may have found it difficult to track individual household members' insurance status as they went through the seven main questions on plan types. The person-level questionnaires contained a check item that required interviewers to look back over seven questions on the previous two pages. The household-level questionnaire contained an "insured" checkbox below each person's name on the grid of household members. Interviewers were to check this box whenever coverage was reported for a household member. Perhaps due to the complexity of tracking coverage on a paper-and-pencil questionnaire, there were several cases where the verification question should have been asked for apparently uninsured individuals but was not (see Table 2). This occurred much more frequently in the household-level questionnaires than in the person-level questionnaires. In the household versions, the question was not asked of 24.9% of eligible sample members; in the person-level versions, the question was not asked of 13.4% of eligible sample members. When the verification question was asked, it had a trivial effect. In the person-level questionnaires, it generated only five reports of coverage, and in the household-level questionnaires, no plans were reported in response to the verification question. Given the small sample size of household members eligible for the verification question, the improper administration of the question, and the low numbers of plans reported in response to the verification question, further analysis of these data is not warranted.

Table 2. Verification question administration and yield

	Person Level		Household Level	
	Q Not Asked	Yield	Q Not Asked	Yield
Current	13.8% (12/87)	3	24.0% (25/104)	0
Calendar year	12.7% (7/55)	2	25.8% (24/93)	0
Total	13.4%	5	24.9%	0

Discussion

Preliminary analysis has been conducted on several different aspects of the data in an attempt to explain why the expected difference between current and calendar-year estimates manifested in the person-level but not the household-level questionnaires, and to explain why there was such a pronounced difference between uninsured rates in the person- and household-level calendar-year treatments. While most of this analysis is still ongoing, there is fairly strong evidence that much of the problem may lie in underreporting of employer-based plans in the household-level calendar-year treatment (see Table 3). In examining this table we would expect to find more reports of employer-based coverage in the calendar-year version relative to the current version. Indeed, among the person-level questionnaires the expected difference is observed—employer-based coverage was reported for 73.1% of respondents in the calendar-year version versus 70.7% in the current version (though this difference is not statistically significant). However, among the household-level version a very different pattern is observed. The calendar-year estimate of respondents with employer-based coverage is actually *lower* than the current version, and the difference is statistically significant at the .10 level (chi-square = 2.86; p = .091). More striking are the household- versus person-level comparisons. Within the calendar-year versions the person-level survey picked up far more employer-based plans than the household version—73.1% versus 64.2%—a difference of almost 9 percentage points, which is statistically significant (chi-square = 14.153; p = .001). Within the current versions, the same pattern is observed—the person-level version picked up more employer-based plans (70.7% vs. 68.1%)—but the magnitude of the difference is much lower and it is not statistically significant.

One possible explanation for these results could be the complexity of the reporting task in the household-level calendar-year version. In the simpler person-level calendar-year version, respondents are asked to perform three main tasks: focus on a particular individual in the household, think back over the previous calendar year (while omitting the past 4–5 months—i.e., from January through the date of interview in April or May of the current year), and then report whether that individual was covered by a particular type of health insurance at any time during the calendar year. These tasks alone could be rather taxing, especially if the household respondent is not very knowledgeable about the person for whom he/she is reporting. However, an even more demanding

Table 3. Employer-based plans

	Person	Household	Difference
Current	70.7%	68.1%	
	(n = 578)	(n = 592)	2.6%
Calendar year	73.1%	64.2%	
	(n = 559)	(n = 498)	8.9%**
Difference	2.4%	–3.9%*	

*Significant at .10 level
**Significant at .01 level

cognitive task is presented in the household-level version, where rather than focusing on one individual at a time, a household respondent is asked to think of all household members at once and report their coverage status. Similarly, while the household-level *current* version of the questionnaire poses some complexity, at least respondents are asked only to report coverage status for household members as of the interview date. In the calendar-year version, respondents have the added task of searching their memories over the full previous calendar year. Given this complexity, it could be that respondents are overtaxed by the request and can't adequately recall all plans for all household members, or for other reasons (e.g., respondent fatigue) simply underreport plans.

If indeed the complexity of the household-level calendar-year version does pose reporting problems for respondents, those problems may manifest most in the first question in the series on health coverage, before respondents have had a chance to adjust to and focus on the general topic of health insurance. The question on employer-based coverage happens to be the first in the series, and since employer-based coverage is the most prevalent type of health insurance, reporting problems associated with this plan type would have the greatest impact on the overall estimates of the uninsured. So the general complexity of the household-level calendar-year version, combined with the sequencing of plan types (i.e., the employer-based question coming first) could be responsible for underreporting of this plan type. This, combined with the prevalence of employer-based coverage, could be driving the overall observed differences in the uninsured rates across the four treatments.

As mentioned above, however, several other aspects of this research are not yet complete but could also help explain the findings. Interviewer behavior is one aspect still under investigation. Specifically, it is yet to be determined whether interviewers adequately probed for the insurance status of all household members in the household-level version relative to the person-level version. In the household-level version, after establishing that at least one person in the household was covered by a certain plan type, interviewers were to probe "anyone else," and continue repeating the probe until the respondent said "no [no one else]." If interviewers did not use this probe exhaustively (i.e., until all household members were accounted for), then respondents would not have received adequate prompts for each household member's coverage status, which may have led to underreporting. In the person-level version, by contrast, a separate questionnaire covering health insurance was administered for each household member. If all questions were read for all household members (which is likely, given the questionnaire design), then respondents in the person-level treatment *would* have received adequate prompts for each household member.

Using forthcoming behavior coding data, analysis will be conducted to determine the extent to which equivalent stimuli for each household member's coverage status (i.e., probes in the household-level version and questions in the person-level version) were provided to respondents. If results show that respondents were *not* provided with adequate prompts in the household-level version relative to the person-level version, the paper-and-pencil questionnaire formatting could be partially responsible. In the paper-and-pencil design the instrument does not assist interviewers in tracking household members' coverage status or in adequate probing. Computerized instruments, on the other hand, generally display the full household roster and the "anyone else" probe at each plan type question. Interviewers therefore may be more likely to track answers for each household member and to read the "anyone else" probe for any household member not yet accounted for. Furthermore, interviewers are forced to enter the code for "no [no one else]" before they can move on to the next question. So to the extent that differences in the adequacy of probing *may* be observed between the household-level and person-level versions, those differences could be an artifact of the paper-and-pencil QDERS design.

Other future research planned for the QDERS data include an examination of differences in the demographics of the uninsured across treatments, and an analysis of reliability via re-interview data.

Conclusions and Future Research

Preliminary findings seem to indicate that the household-level calendar-year questionnaire results in underreporting of health plans relative to the current and person-level questionnaires. Failure to report employer-based plans in the household-level calendar-year questionnaire seems to be driving the overall higher estimates of the uninsured observed in this treatment. Possible reasons for underreporting include sequencing effects, complexity of the reporting task, recall failure, respondent fatigue, and other factors not yet fully explored. However, further analysis of the QDERS data set itself, and forthcoming behavior coding data and re-interview data, is necessary before any firm conclusions can be drawn.

These preliminary results suggest other areas for future research as well. First, with regard to the identical rate of uninsured in the current and calendar-year household-level questionnaires, cognitive research could provide insights into respondents' perceptions of the household-level calendar-year questions and their recall and reporting strategies for household members' health insurance status. Specifically, probing could focus on the time period for which respondents report searching their memories, the individual household members they think of, and the particular plans that come to mind. This research, in turn, could suggest methods to

improve recall and reporting. Second, with regard to the verification question, there seems to be enough evidence in the literature in general to justify further study of the effectiveness of this type of followup question to verify the uninsured. Further research could be especially valuable if conducted on large samples in a computerized environment, where tracking coverage status would be automated and, likely, more reliable than in a paper-and-pencil version. A larger sample size would enable more meaningful analysis of people who are reported as covered only in response to the verification question.

References

American Association for Public Opinion Research. (1998). *Standard definitions: Final dispositions of case codes and outcome rates for RDD telephone surveys and in-person household surveys.* Ann Arbor, Michigan.

Bennefield, R. (1996). A comparative analysis of health insurance coverage estimates: Data from CPS and SIPP. Presentation at the 1996 Joint Statistical Meetings of the American Statistical Association.

Carlson, B. (1998). Improving survey estimates of the uninsured using computer assisted interviewing logic. Paper presented at the annual meeting of the American Statistical Association.

Lewis, K., Ellwood, M., & Czajka, J. (1998). *Counting the uninsured: A review of the literature.* Washington, DC: Mathematica Policy Research.

Rosenbach, M., & Lewis, K. (1998). *Estimates of health insurance coverage in the Community Tracking Study and the Current Population Survey.* Document No. PR98-54. Washington, DC: Mathematica Policy Research.

Schwartz, K. (1986). Interpreting the estimates from four national surveys of the number of people without health insurance. *Journal of Economic and Social Measurement, 14,* 233–242.

U.S. Census Bureau (1997). Health insurance historical table 1: Health insurance coverage status and type of coverage by sex, race and Hispanic origin: 1987 to 1997. March Current Population Survey data.

Methodological Differences in Measuring Health Care Coverage

Rachel Harter, Alma Kuby, and Whitney Moore

Introduction

Many surveys in recent years have attempted to measure the uninsured and underinsured population in the United States. In advancing such studies, researchers and policymakers recognize their relevance to a host of critical public policy concerns. For example, changes within the medical insurance industry, governmental health care reform, welfare reform, and immigration policies all determine to some extent who receives what types of health care and coverage. With welfare reform, for example, many persons who leave the welfare rolls are unaware that they still qualify for Medicaid, and this affects the types of health care they seek. There is concern that increased Hispanic immigration, particularly illegal immigration, is leading to a higher percentage of uninsured persons in the United States. Furthermore, evidence shows that a disproportionate number of uninsured individuals are low-income employed persons who do not qualify for Medicaid and are either not offered coverage or are unable to pay for insurance premiums. Studies have suggested that lack of insurance coverage reduces consumption of medical care services, which could result in adverse health effects.

Insufficient health care coverage for large segments of the population not only poses risks for the well-being and health of individuals and the general public as a whole, but also entails increased health care costs from care that has been delayed. Recent legislation has attempted to address the growing concerns related to inadequacy of health insurance coverage, especially among certain subgroups of the population. The 1996 Health Insurance Portability and Accountability Act (HIPAA), while not increasing access to health insurance for persons who are currently uninsured, does provide availability and portability for employees who change jobs and to people who have exhausted coverage under COBRA. The Child Health Insurance Program (CHIP) is intended to help reverse the decline in health care coverage for children. However, accurate measurement of health care coverage for these and other subpopulations is essential for determining the success of these programs.

Most major studies of the uninsured and underinsured have been population-based studies rather than provider surveys or analyses of administrative records. Provider surveys will naturally miss that portion of the population who do not seek medical care. Administrative records may be complicated by regulations that vary from state to state. Population-based studies bypass these issues and have the additional advantage of flexibility in asking about related items for more in-depth analysis.

Survey methodologists have long recognized that different survey designs and related methodological considerations can influence survey findings. One critical finding from studies of the uninsured and underinsured is the estimate of the numbers of such persons in the population. Yet this figure is seen to vary widely across some key government and privately supported studies, including the Current Population Survey (CPS), the Survey of Program Participation, and the National Medical Expenditure Survey. Monheit (1994) provides a thorough comparison of these well-known studies. Additional estimates of the uninsured and underinsured populations are provided by the Kaiser Survey of Family Health Experiences (K-SOFHE), a major foundation-supported project in which the present authors were senior survey researchers.

A number of methodological factors can influence the measurement of the uninsured and underinsured population. This paper will explain how differences in family definition, eligibility, choice of respondents, reference periods, question wording, and general context affect the estimates of uninsured families and individuals across two important population-based studies, the K-SOFHE (first round), conducted in the fall of 1995, and the March 1996 CPS. Descriptions of their methodologies and comparisons of their health care coverage estimates will illustrate their differences and help to provide possible reasons for the widely differing estimates of the uninsured arrived at by studies today. In addition, the analytical approach used by the authors will offer a model for data users and analysts to determine the most appropriate estimates for their purposes.

Family Definition

Giovannini, Kasper, Hoffman, & Lee (1999) showed that the insurance status of individual family members can affect the entire family's attitudes and behavior toward seeking health care. It is useful, therefore, to estimate the uninsured families—the families with at least one uninsured member. In estimating uninsured families, the definition of "family" will invariably affect the estimates. To begin with, individual interpretation of family composition may lead to definitional differences. For example, in some households, it is unclear what constitutes a "family" because the household composition is in

The authors are at the National Opinion Research Center.

Table 1. Weighted numbers of families of size *n* (in thousands)

	Family Size						
	1	2	3	4	5	6+	Total
CPS	28,830	23,260	15,730	14,260	6,160	3,880	91,620
K-SOFHE	18,260	28,510	13,750	15,920	7,100	2,670	86,210

flux. Adult children, elderly parents, significant others, and roommates can blur the boundaries of the family unit. Furthermore, procedural aspects of a survey may influence definitions even more directly. In K-SOFHE, for example, unrelated and unmarried adults living together are considered a family if they refer to themselves as family or use terms such as "wife," "husband," or "spouse." In CPS, however, unrelated and unmarried adults living together are not considered one family. Counting such an arrangement as two separate families increases the total family count and increases the estimate of uninsured families because single adults account for a high percentage of the uninsured population. (Of course, these differences in family definition should not affect the estimates of uninsured persons.)

Table 1 displays the weighted estimates of families by size for the two surveys. While the estimates of larger families are generally comparable, the overall family counts and the estimates of families with one or two people are widely disparate. The CPS has far more families of size 1 and fewer of size 2 than the K-SOFHE. Further, the CPS has more families overall than the K-SOFHE. These differences may be explained by the definitional variations between the two surveys described above. The fact that the CPS counts an unmarried couple living together as two families and the K-SOFHE counts them as one will almost surely increase the CPS total family count and the count of single-person families, and may also contribute to the CPS's smaller estimate of two-person families.

Table 2 breaks the CPS and K-SOFHE family estimates down by insurance type. As just illustrated, the CPS has larger numbers of families overall and larger numbers of single-person families than the K-SOFHE. The implications of these differences are realized in the larger estimates of uninsured families in Table 2. It is generally known that unmarried people account for a large proportion of uninsured individuals (Bennefield, 1996). As a result of the CPS procedures involving family definition, uninsured couples who are living together but not married would constitute two uninsured families in the CPS. The same couple would count as only one uninsured family in the K-SOFHE.

Table 2. Weighted estimates of families by insurance type (in thousands)

Insurance Type	CPS		K-SOFHE	
	Count	Percent	Count	Percent
Uninsured	19,720	21.5	12,050	14.0
Medicaid	12,450	13.6	8,170	9.5
Private	59,450	64.9	65,990	76.5
Total	91,620	100.0	86,210	100.0

Eligibility

Studies may differ as to which families or persons are considered in-scope. In the K-SOFHE, an in-scope family must have at least one person under age 65. In our comparisons with CPS, we restricted CPS families to those with at least one person under 65 for comparability. The K-SOFHE excluded families who did not speak English. The CPS does not exclude non-English-speaking families, and we had no way of identifying and excluding them from our analyses. Although we expect the non-English-speaking population to be small, they may have a greater probability of being underinsured, contributing in some small way to the discrepancies between CPS and K-SOFHE.

Respondents

For some studies, any adult in the household may be able to answer at least the basic screener questions. For insurance matters, typically only one or two of the resident adults has sufficient knowledge to respond satisfactorily. Thus, the selection of a respondent within a household can affect the completeness and quality of the results. The problem of selecting a knowledgeable respondent is compounded in households with multiple or extended families.

Reference Periods

Reference periods vary widely among surveys, ranging all the way from point-in-time, in which questions refer to present circumstances, to retrospective, in which the respondent is asked to recall information from sometime in the past, usually more than six months before the survey. Just as definitional differences can greatly affect estimates, the reference period for insurance questions can also have a major impact on the outcome. For example, the number of persons who were uninsured at any time in the past year far exceeds the number uninsured at a specific point in time, which in turn exceeds the number that were uninsured throughout the entire year. The K-SOFHE is a point-in-time survey; its questions refer to insurance status specifically at the time of the interview. The CPS actually has two reference periods. In its

employment benefits section, respondents are asked whether household members were covered by health insurance *at any time* during the prior year. In 1996 the CPS expanded its survey to include an additional reference period. In this new health insurance variables section, it asks about insurance coverage for household members during the week preceding the survey.

The reference period can also affect outcome by the phenomenon known as "telescoping," in which the respondent's memory of experiences does not coincide exactly with the reference period (Sudman, Bradburn, & Schwarz, 1996). The respondent may incorrectly recall a period of no insurance coverage as being within the last 12 months when in fact it happened prior to that, or vice versa. Point-in-time questions do not have problems with telescoping.

The questions presented by the CPS span from a nearly point-in-time reference period to a retrospective reference period, exacting widely disparate responses, as illustrated in Tables 2 and 3. However, the differences due to reference periods are confounded with the effects of other methodological factors discussed below.

Question Wording

The wording of a question may also influence the outcome. For example, questions that list a variety of providers from which the respondent may have received health insurance can help with recall. Both the original health insurance questions in the CPS employment benefits section and the new insurance variables introduced in 1996 probe the respondent about health care coverage from a variety of providers, including employer or union insurance, private insurance purchased directly, insurance coverage provided by someone outside of the household, Medicare, Medicaid, and military health care. Specifically, the CPS questionnaire asks, "At any time during 1995, (were you/was anyone in this household) covered by (insurance type)?" The questionnaire later reads, "These next questions are about *current* health insurance coverage, that is, health coverage last week. (Were you/Was anyone in this household) covered by *any* type of health insurance plan last week?" It then asks who was covered and by what type of plan.

The first round of K-SOFHE, conducted in late 1995, collected insurance coverage at the time of interview for each individual family member through a series of questions that addressed employment-related private insurance, individual (non-employment) private coverage, Medicaid, Medicare, and CHAMPUS/VA coverage. Specifically, the screener asks, "Is anyone in the household covered by (insurance type)?" It also asks questions such as "Whose insurance is covering (each person)?" and "From what I have recorded, it appears that (person) is not covered by private health insurance, Medicare, Medicaid, or health insurance through the military. Is that correct?" Persons for whom no coverage was reported for any of these were considered uninsured.

According to Monheit (1994), Kronick (1991) pointed out that CPS insurance questions prior to 1988 were worded in such a way that older children living at home, dependents covered by nonresident policy holders, and children under 15 with their own insurance were considered uninsured, inflating the estimates. It is unclear whether the differences in question wording between the 1996 CPS and K-SOFHE caused any significant differences in outcome, but the potential exists.

Context

Finally, the context of the insurance questions may influence responses. The CPS primarily covers employment issues, and insurance coverage is a secondary topic. The responses may be conditioned to different thought patterns and memories than in surveys such as K-SOFHE that primarily target health care. According to Sudman, Bradburn, and Schwarz (1996), the content of preceding questions may affect any of the steps of the question answering process and involve several different psychological processes. For instance, supporting questions that specifically target children's coverage by a nonresident parent's employment may aid a respondent's recall and trigger different thought patterns, thereby reducing the estimates of uninsured children. Thus, in the context of employment issues, recall of health insurance coverage may elicit responses skewed toward employer coverage.

Comparison of Person Estimates

A weighted tabulation of persons by insurance type is presented in Table 3. Two estimates are given for the CPS, corresponding to the two sets of questions regarding insurance coverage in the March 1996 survey. The K-SOFHE estimates are poststratified to CPS population totals; thus, the estimated totals of persons across age, race, and gender will be equal.

Though we would expect the K-SOFHE estimates to resemble most closely the CPS estimates for coverage during the prior week, as these are closest to point-in-time estimates, Table 3 shows just the opposite. The K-SOFHE indicates that approximately 12% of individuals are uninsured, while the CPS prior-week estimate shows that 29% are uninsured. The CPS prior-year estimate of uninsured persons is closer to the K-SOFHE estimate, at 17%. The K-SOFHE estimate of Medicaid recipients falls between the two CPS estimates (at 9% versus 7% for the CPS prior-week estimate and 13% for the CPS prior-year estimate), and K-SOFHE yields the largest percentage with private insurance (close to 77%, compared to 62% and 69%, respectively, for CPS prior-week and prior-year estimates).

Table 4 compares the weighted CPS insurance coverage estimates to the weighted K-SOFHE estimates by age group. As seen above, the CPS shows a higher percentage of uninsured persons overall than the K-SOFHE. Table 4 illustrates that this difference holds for all of the age groups younger than 65 years fairly consistently. Again, the CPS prior-week

Table 3. Weighted estimates of persons by insurance type (in thousands)

	CPS		CPS		K-SOFHE	
	March 1996: Prior Year Coverage		March 1996: Prior Week Coverage		Late 1995: Point-in-Time	
	K-SOFHE In-Scope[1]		K-SOFHE In-Scope[1]		K-SOFHE In-Scope[1]	
Insurance Type	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Uninsured	40,463	17.1	68,931	29.1	27,845	11.7
Medicaid	29,953	12.6	16,287	6.9	21,201	8.9
Private	163,034	68.7	146,943	62.0	181,589	76.6
Other	3,759	1.6	5,049	2.1	5,762	2.4
Refused / DK	0	0.0	0	0.0	814	0.3
Total	237,210	100.0	237,210	100.0	237,210	100.0

[1]Persons in-scope for K-SOFHE are those who are members of families with at least one person under the age of 65. Thus, all CPS persons in such a family are considered to be in the K-SOFHE in-scope universe.

Table 4. Weighted numbers and percentages of persons by age group and insurance type (in thousands)

		CPS (Prior-Year Coverage)		CPS (Prior-Week Coverage)		K-SOFHE	
Age Group	**Insurance Status**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
0–6	Uninsured	3,824	13.6	8,607	30.6	2,183	7.8
	Medicaid	8,276	29.4	4,663	16.6	7,130	25.3
	Private	15,916	56.6	14,570	51.8	18,474	65.6
	Total	28,138	100.0	28,138	100.0	28,138	100.0
7–17	Uninsured	5,971	14.2	11,884	28.2	3,895	9.2
	Medicaid	8,248	19.6	4,506	10.7	5,466	13.0
	Private	27,802	65.9	25,348	60.1	32,640	77.4
	Total	42,171	100.0	42,171	100.0	42,171	100.0
18–49	Uninsured	25,923	20.7	39,125	31.2	18,564	14.8
	Medicaid	10,370	8.3	5,500	4.4	6,643	5.3
	Private	88,392	70.4	79,335	63.2	97,927	78.0
	Total	125,451	100.0	125,451	100.0	125,451	100.0
50–64	Uninsured	4,563	13.6	7,792	23.2	3,042	9.0
	Medicaid	2,163	6.4	1,148	3.4	1,006	3.0
	Private	26,363	78.4	23,767	70.7	28,847	85.8
	Total	33,640	100.0	33,640	100.0	33,640	100.0
65+	Uninsured	182	2.3	1,522	19.5	162	2.1
	Medicaid	896	11.5	470	6.0	955	12.2
	Private	4,561	58.4	3,922	50.2	3,701	47.4
	Total	7,808	100.0	7,808	100.0	7,808	100.0

coverage estimates of the uninsured are dramatically higher than both the CPS prior-year and the K-SOFHE estimates. For Medicaid coverage, the CPS prior week estimates and the K-SOFHE estimates are more similar (K-SOFHE estimates are consistently slightly higher than CPS prior-week estimates), and both fall below the CPS prior-year estimates. Finally, private insurance coverage rates are highest for the K-SOFHE, and lowest for the CPS prior-week, for persons aged 64 and younger.

Summary and Conclusions

As shown here as well as in several other studies, estimates of uninsured families and persons can vary widely across surveys. In the case of the CPS and K-SOFHE comparison, several methodological differences contribute to their disparate estimates by insurance status. The issue is not right versus wrong, but rather which is more appropriate for the individual user's purposes.

We conclude with a list of questions that researchers should ask to determine which estimates are most relevant to their purposes given the methodologies of the studies that produced the estimates.

1. How are key terms defined, and how might these definitions affect the estimates of uninsured?
2. What are the eligibility criteria for the surveys? Will these result in greatly differing populations for which sample estimates will be made?
3. How were respondents selected?
4. What is the reference period, and how would one expect it to affect estimates of uninsured?
5. How are the questions worded? Is the language ambiguous or is it targeted to the measure most relevant to your study?
6. What is the general context of the questionnaire? Is health insurance coverage the primary focus, or a secondary topic?

References

Bennefield, Robert L. (1996). Health insurance coverage: 1995. *Current Populations Reports,* U.S. Census Bureau, Pub. No. P60-195.

Current Population Survey. (1997). *March 1997 technical documentation.* Prepared by Administrative and Customer Services Division, Microdata Access Branch, Bureau of the Census. Washington: The Bureau.

Giovannini, T., Kasper, J. D., Hoffman, C., & Lee, Y. (1999). Health insurance coverage in American families. Presented at the 16th Annual Meeting of the Association for Health Services Research, June 1999, Chicago.

Monheit, Alan C. (1994). *Underinsured Americans: A review.* Agency for Health Care Policy and Research, U.S. Department of Health and Human Services, Pub. No. 94-0088, 461–485.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers*. San Francisco: Jossey-Bass.

Identifying Children with Special Needs

Floyd Jackson Fowler, Jr., Patricia M. Gallagher, and Charles J. Homer

Introduction

In doing surveys of children designed to describe or assess medical care experiences, it is often important to include survey questions that identify children with special needs. Because their needs for care are, by definition, greater than average and because the care they receive may make a particularly important difference in their lives, describing the experiences of such children is seen by many as critical to the assessment of the quality of care that is delivered.

In theory, medical records might seem to be the best way to identify children with special needs. However, many surveys are done when it is impossible, or at least not easy, to use such records. For general population surveys and for surveys of children when good diagnostic data are not readily accessible, a set of questions to identify children with special needs is needed.

Several different approaches can be used to characterize children as having special needs. One standard on which virtually everyone agrees is that we should be considering a condition or health state that is long-term, rather than short-term. For adults, the National Health Interview Survey has a long-standing standard that a "chronic condition" is one that lasts at least 3 months. Three months seems a minimum standard (Perrin et al., 1993). Some researchers recommend a 12-month standard for children (Stein, Westbrook, & Bauman, 1996).

In addition to being "chronic," several other characteristics often are used in the definition of a child with special needs:

1. Requires a significant amount of medical care
2. Limited in ability to do certain activities
3. Requires services that most children don't need

The Consumer Assessment of Health Plans (CAHPS®) project has, among its various goals, the measurement of the care received by children with special needs. In order to accomplish this, one necessary step is the development of a set of questions to identify such children. Minimizing respondent burden is a fundamental standard for CAHPS® instruments. Hence, one of our goals was to find the most parsimonious set of questions that would meet our objective (Crofton, Lubalin, & Darby, 1999).

This paper reports on the results of a methodological study done with the Medicaid population of Massachusetts that was designed to evaluate the contribution that different combinations of questions would make to identifying children with special needs in a survey.

Methods

The Sample

The sample frame consisted of children 17 years old or younger who were covered by the Primary Care Clinical (PCC) plan of MassHealth for at least six continuous months as of December 1997. There are four programs by which children can be made eligible for MassHealth by virtue of having a particular health condition or problem; the largest of these programs is SSI (Supplemental Security Income). A probability sample of 800 children was drawn from those children who were eligible for MassHealth through one of these four programs. Another sample of 800 children was drawn from the remainder of the MassHealth population in the PCC program.

The Survey Instrument

The survey instrument was the basic CAHPS® core of questions designed to capture people's experiences with access to care, interactions with providers, and interactions with plans (Weinberger, 1999). For half the sample, the instrument also included additional questions designed to capture the experiences of people who have special health care needs, and hence whose use of services is more intense than average.

The basic series designed to identify individuals with a chronic condition includes the following three questions:

1. Does your child now have any medical conditions that have lasted or are expected to last for at least 12 months?
2. (If YES) In the last 12 months, has your child seen a doctor or other health provider more than twice for any of these conditions?
3. (If YES) Has your child been taking prescription medicines regularly for any of these conditions?

The authors are at the Center for Survey Research, University of Massachusetts Boston, and the Children's Hospital, Boston.

The research reported here was supported by a cooperative agreement with the Agency for Health Care Policy and Research. We also gratefully acknowledge the assistance and cooperation of the Division of Medical Assistance, Commonwealth of Massachusetts.

Table 1. Returns for children sent 95-item (chronic condition) questionnaire by data collection phase and program enrollment

	Ineligible	Eligible Sample	Mail Survey Completed	Phone Interview	In-Person Interview	Total Returns	Refusal	Other Non-Interview	Response Rate
Cross-section of MassHealth	13	387	130	42	109	281	24	82	73%
Enrolled in SSI or other special program	11	389	137	32	94	263	13	113	68%
Total	24	776	267	74	203	544	37	195	70%

Any child reported to have a condition that has lasted (or is expected to last) 12 months and with a "yes" answer to at least one of the two follow-up questions could be classified as having a special need for medical care.

In addition, the following questions in the survey could be used to indicate a child with special needs:

1. Does your child have a physical, emotional, or mental condition that seriously interferes with your child's ability to do the things that most children that age can do?
2. (If enrolled in school) Does your child have health care needs that require any *special help* from teachers, nurses, or staff at your child's school?
3. In the last six months, did your child have any health problem that required you to get or replace any special medical equipment or devices such as a walker, wheelchair, nebulizer, feeding tubes, or oxygen equipment?
4. In the last six months, did your child have any health problems that needed special therapy, such as physical, occupational, or speech therapy?
5. In the last six months, did you need someone to come into your home to give home health care or assistance for your child?
6. In the last six months, did you need respite services for your child?
7. Does your child have any kind of emotional, developmental, or behavior difficulty now for which he or she has received treatment or counseling?

The Data Collection Protocol

Data were collected using a combination of mail, telephone, and in-person interviewing strategies. First, a copy of the survey instrument, printed in English and in Spanish, was sent to the parents or guardians of the sampled children. Along with the instrument was a fact sheet, answering commonly asked questions about the survey, and instructions to fill out the survey instrument for the specific child who was listed on the fact sheet. Subsequently, a reminder postcard and another mailing of the survey instrument to nonrespondents was carried out. After the mail protocol had been completed, an effort was made to find telephone numbers and to conduct telephone interviews with responsible adults who were informed about the sampled children. Finally, after the telephone protocol was complete, interviewers were sent to the addresses of nonrespondents and an effort was made to conduct interviews in person.

Field Results

Since the analysis will focus on that half of the sample assigned to answer the supplemental questions for children with special needs, we discuss the results for those children. After the mailing phase of the project had been completed, survey returns had been obtained for about 34% of sampled children. The telephone data collection effort raised the response rate to about 44%. Finally, in-person interviewers obtained returns from about 26% of sampled individuals, making the final response rate 70%. Only about 5% refused. The others were nonrespondents for miscellaneous other reasons, including inability to speak either English or Spanish. About 10% could not be located at all. In all, there were 544 responses for those who were asked the full set of questions designed for children with chronic conditions. Of those, 263 were in SSI or another program indicating the presence of a serious chronic condition, while the balance (281) were not known to have a condition meeting the eligibility criteria for special programs in MassHealth (see Table 1).

Analysis Plan

The main goal of the analysis is to evaluate the implications of using the various candidate survey questions to identify a group of children likely to have special needs. To do this, we carried out four different kinds of analyses.

1. For the cross-section of MassHealth members (whose Medicaid eligibility did not depend on the presence of a known chronic condition) we looked at the extent to which individual additional candidate questions identified additional children who might have special needs, over and above those identified by the basic three core questions.

2. One group of children known to have some kind of chronic condition were those who met eligibility requirements for SSI or another special program. Thus, a second analysis was carried out to evaluate the productivity of each of the questions in helping to identify those children who were known to be eligible by virtue of having a condition.
3. Claims data were examined for the two years prior to data collection. The five most common diagnoses appearing in those claims were identified. Then, the diagnosis codes were compared with a list of ICD-9 codes considered to be indicative of the presence of a chronic health condition in children (see Appendix). The third analysis examined the productivity of the candidate questions in identifying those children whose claims indicated a chronic condition.
4. The costs associated with caring for children were also examined. The total cost of services covered by MassHealth during the prior two-year period were calculated. To adjust for differing lengths of time in the MassHealth program, the MassHealth costs recorded in that period were divided by the number of months during that two-year period in which a child had been enrolled in MassHealth, producing an average expenditure per month. Using that variable, we divided covered children into quartiles. The fourth analysis used the questions to see how effective they were in identifying children who had distinctively high costs for medical care.

Results

Table 2 shows the number of children identified by the various questions as having a chronic health condition. In the first column, the three basic questions (asking about the presence of a condition that has lasted for at least 12 months, and whether or not the child has seen a doctor more than two times or regularly takes medication for the condition) are used as the basic identifying mechanism. The responses are broken down for the cross-section of MassHealth children and those who have been made eligible for SSI or some other program that indicates the presence of a chronic condition. It can be seen that 65% of those eligible for SSI or another program were identified by the three questions; 14% of the cross-section of the remainder of MassHealth were also identified as having a chronic condition.

In the second column, we calculated the percentage of children who would have been further identified by each of the seven questions in the survey instrument that could be used to identify children with special needs. The right-hand column shows the total that would be identified by using the three basic questions plus each of the individual seven questions alone. For the children enrolled in SSI or some other special program, the two most productive additional questions dealt

Table 2. Percentage of sampled children identified as having a chronic condition using three basic questions plus individual additional questions by program of enrollment

	Additional Possible Questions to Identify Chronic Conditions or Special Needs	Using Three Basic Questions*	Plus Each Additional Question	Combined Total
Cross-section of MassHealth (n = 281)	Has physical, emotional, or mental condition interfering with ability to do what most kids that age can do	14%	5%	19%
	Required special help at school for health care needs	14%	8%	22%
	Needed special medical equipment	14%	4%	18%
	Needed physical, occupational, or speech therapy	14%	8%	22%
	Needed home health care	14%	2%	16%
	Needed respite services	14%	1%	16%
	Needed treatment or counseling for an emotional, developmental, or behavior difficulty	14%	11%	26%
Enrolled in SSI or other special program (n = 263)	Has physical, emotional, or mental condition interfering with ability to do what most kids that age can do	65%	17%	82%
	Required special help at school for health care needs	65%	11%	76%
	Needed special medical equipment	65%	2%	67%
	Needed physical, occupational, or speech therapy	65%	11%	76%
	Needed home health care	65%	2%	67%
	Needed respite services	65%	2%	66%
	Needed treatment or counseling for an emotional, developmental, or behavior difficulty	65%	16%	80%

* Has condition that has lasted (or is expected to last) 12 months and *either* saw a doctor three or more times in past year *or* takes prescription medicine regularly for condition.

Note: Data are reported only for those who responded to the questionnaire that included CAHPS® chronic condition supplemental questions.

Table 3. Percentage of children identified using various combinations of questions as having a chronic condition by program of enrollment

	Using Three Basic Questions	Plus Limitation of Activity	Plus Need for Counseling	Using All Questions
Cross-section of MassHealth (n = 281)	14%	19%	27%	35%
Those enrolled in SSI or other special program (n = 263)	65%	82%	86%	90%

Note: Data are reported only for those who responded to the questionnaire that included CAHPS® chronic condition supplemental questions.

Table 4. Percentage of children identified using various combinations of questions as having a chronic condition by presence or absence of significant diagnoses in claims data

Claims Indicate Diagnosis of Significant Chronic Condition*	Using Three Basic Questions	Plus Limitation of Activity	Plus Need for Counseling	Using All Questions
Yes (n = 55)	36%	40%	47%	55%
No (n = 244)	13%	19%	27%	35%

* Defined by matching five most common diagnosis codes in claims for past 24 months against a list of ICD9 codes that signify the presence of a significant chronic condition.

Note: Data are reported only for those who responded to the questionnaire that included CAHPS® chronic condition supplemental questions.
For claims analysis, those enrolled through SSI or other special programs are weighted down to adjust for their higher probability of selection.

Table 5. Percentage of children identified using various combinations of questions as having a chronic condition by average health care expenditures per month

Average Expenditures per Month in Past 24 Months	Using Three Basic Questions	Plus Limitation of Activity	Plus Need for Counseling	Using All Questions
Bottom quartile: $0–48.68 ($n$ = 71)	4%	10%	17%	21%
Second quartile: $48.69–89.59 ($n$ = 78)	6%	9%	14%	19%
Three quartile: $89.60–197.06 ($n$ = 74)	22%	23%	30%	40%
Top quartile: $197.07 or more ($n$ = 75)	37%	48%	61%	74%

Note: Data are reported only for those who responded to the questionnaire that included CAHPS® chronic condition supplemental questions.
For expenditure analysis, those enrolled through SSI or other special programs are weighted down to adjust for their higher probability of selection.

with limitation of activities or having needed counseling. The addition of either one of those questions would have brought the percentage of children identified to 80% or higher. In the cross-section of the remainder of MassHealth children, the most productive additional question asked if the child had a condition that required counseling.

Table 3 shows what would happen if one used some optimal combinations of questions. The first three columns show results from the three basic questions plus the two most productive questions, limitation of activity and need for counseling, using a positive response to indicate the presence of a chronic condition. The last column shows the results when all the candidate questions are used. With that approach, it can be seen that 90% of those enrolled in SSI or another special program would have been identified; 35% of all children who are in the cross-section of the balance of MassHealth would have been so identified.

Table 4 presents a parallel analysis, except that the target is to identify those children who had a diagnosis code in the previous two years of claims that indicated a chronic condition. It can be seen that only 36% of those whose claims indicated the presence of a chronic condition are identified by using the three basic questions. Adding the two most productive questions increases the number to 47%; using all the questions moves the total up to 55%. Meanwhile, it is important to note that a good number of those who had *no* such diagnosis code in their claims data would have been identified by the survey questions as having a significant chronic condition. If all the questions were used, 35% of those children would have been identified as having a chronic condition. Notice that the data for that group are very similar to the data for the cross-section of MassHealth members who have not been made eligible for a special program (see Table 3).

In Table 5, we look at the association between average expenditure per month for two years and whether or not children would be identified as having a chronic condition by the survey questions. Focusing on the top quartile, the three basic questions identify only 37% of the high-cost children. The two other most productive questions move that figure to 61%. If all the questions are used, about 74% of those in the top

quartile would have been identified as having a significant chronic condition. At the other end of the table, for those in the bottom two quartiles, who have lower-than-average expenditures, about 20% of these children would have been identified as having a significant condition if all the questions in the survey instrument were used. In contrast, if only the three basic questions were used, about 5% of children in the low-expenditure quartiles would have been identified as having a significant condition.

Discussion

Any discussion of an optimal set of questions depends on goals and priorities. Asking more questions with the potential to identify children with special health care needs or chronic conditions will flag more children as potentially falling in that category. Each new question increases survey cost and respondent burden. It also brings some potential risk of false positives, identifying children who are not really among those we want to identify.

In this analysis, probably the best standard against which to measure the productivity of questions is the knowledge that a child is enrolled in MassHealth by virtue of being eligible for SSI or some other special program that requires certification of a health need or condition. Using that as a standard, the three basic questions plus the questions on limitation of activity and need for counseling would identify 86% of that population. The additional five questions only move that 86% figure up to 90%. Based on that analysis, one might argue that those initial five questions are close to optimal.

Those five questions identify only a minority of those who have a diagnosis of a significant chronic condition among the top five diagnosis codes in their claims over the last two years. However, adding the other five questions still only gets us to 55%. There clearly are some reasons to be concerned that the diagnosis codes in claims files are not a reliable way to identify children with special needs.

We had hoped that the diagnosis codes from the claims files would provide a basis for assessment of false positive rates—the rates at which questions identified children who did not have significant chronic conditions. However, the rates at which respondents gave answers that indicated chronic conditions were the same for those with *no* chronic condition diagnosis codes as for the cross-section of MassHealth children. This suggests that not having such a code provides almost no information about the "absence" of a chronic condition. Others have found that claims files are not a very good source of diagnostic information. The diagnoses that are coded depend on the kinds of services that are provided and, to some extent, physician discretion.

We expect to do some further analysis of the claims files, to see if we can make them more sensitive and improve the case for the validity of the results for achieving our goals. The NACHRI (National Association of Children's Hospitals and Related Institutions) classification, for example, is a much more developed system than the one we used, and it also is much more complicated to use (Gay, Muldoon, Neff, & Wing, 1998). We think the approach we used in this analysis produces a questionable standard against which to assess the value of the survey questions to validly identify children with special needs.

Turning to the expenditure data, obviously people can have high expenditures for acute problems, such as injuries, that we would not want to identify with these questions. On the other hand, there are some quite serious long-term health conditions that do not necessarily use a great deal of medical care on an ongoing basis. Thus, one would not expect a perfect correspondence between a good set of questions about the presence of conditions and expenditure data; one would expect a strong positive correlation between expenditures and measures of the presence of conditions, and we found that.

In this analysis, using all seven questions in addition to the three basic questions proved to have a significant positive effect on the number children identified as having special needs in the high-expenditure category. Those results might argue for the value of keeping more questions in a survey instrument designed for this purpose. On the other hand, this increase may reflect the identification of acute conditions that caused high expenditures.

The choice of questions, finally, comes down to the purposes for which people are using screening questions in their surveys. The issues are both conceptual and pragmatic. For the CAHPS® surveys, the principal reason to identify children with special needs is so that plans can be compared in the way they manage children who require more than the average medical care and who place special demands on the medical care delivery system. For that purpose, a reliable set of questions, even if it misses some people, may serve quite well. The five-question series (three basic questions plus limitation of activities and use of counseling) identify 86% of children in a special needs program and may be an optimal series.

In contrast, if the goal is to get a good estimate of the number of children with special care needs in a population, then asking more questions that cover more of the various ways in which special needs might appear would be valuable. In the cross-section of MassHealth children, the number of children identified as having special needs was increased by one-third when all seven supplemental questions were asked; in identifying children with special needs among those with high expenditures, the increase was over 20%.

The further question of which children should be counted is, of course, key. Asking about the presence of specific conditions has the benefit of enabling one to be more specific about who has a special need, but it has other serious limitations including low prevalence rates and problems with validity of reports (Jabine, 1987). Most observers agree that a focus on the impact of conditions is the right approach (Stein et al., 1997; Perrin, et al., 1993; Newacheck, et al., 1996). However, that still leaves open the extent to which the following need to be identified: (a) developmental problems (b) cognitive problems (c) behavioral problems (d) conditions that do not limit the child or require extra medical care and (e) chronic acute problems, such as earaches. Thoughtful researchers would not be comfortable with a

narrow, medically defined approach, but where the lines should be drawn is not always clear.

If failing to identify some children with special needs is an important cost, the additional questions will surely pay off. Identifying exactly which questions will require further debate and research. Moreover, there are details about the design and wording of the specific questions tested that warrant review. However, we think the basic concepts covered by the questions are sound. If identifying a very large percentage of children with special needs in a population will suffice, the three basic questions (covering duration of conditions and use of medical services and prescription medications) plus questions about conditions that limited activity or required counseling will serve those purposes very well.

References

Crofton, C., Lubalin, J., & Darby, C. (1999). Forward. *Medical Care*, *37*(3), Supplement, MS1–MS9.

Gay, J., Muldoon, J., Neff, J., & Wing, L. (1998). Profiling the health service needs of populations: Descriptions and uses of the NACHRI classification of congenital and chronic health conditions. *Pediatric Annals, 26*(11), 655–663.

Jabine, T. (1987). Reporting chronic conditions in the National Health Interview Survey: A review of tendencies from evaluation studies and methodological tests. *Vital and Health Statistics* (ser.2, no. 105). Washington, DC: Government Printing Office.

Newacheck, P., Stein, R., Walker, D., Gortmaker, S., Kuhlthau, K., & Perrin, J. (1996). Monitoring and evaluating managed care for children with chronic illnesses and disabilities. *Pediatrics, 98,* 953–958.

Perrin, E., Newacheck, P., Pless, I., Drotar, D., Gortmaker, S., Levinthal, J., Perrin, J., Stein, R., Walker, D., Weitzman, M. (1993). Issues involved in the definition and classification of chronic health conditions. *Pediatrics, 91,* 787–793.

Stein, R., Westbrook, L., & Bauman, L. (1997). The questionnaire for identifying children with chronic conditions: A measure based on a noncategorical approach. *Pediatrics, 99,* 513–521.

Weinberger, M. (Ed.). Consumer Assessment of Health Plans (CAHPS®), Appendix 1. (1999). *Medical Care, 37*(3), Supplement.

Appendix

Condition	ICD-9 Code(s)
Infectious diseases	030–030.9, 040.2, 052, 046–046.9, 135, 136.3
Malignancies	140–208.9
Neurofibromatosis	237.7–237.9
Thyroid	240–246.9
Diabetes	250–250.9
Other endocrine	252–252.9, 253–253.9, 255–255.9
Nutritional deficiencies	260–262, 268–268.1
Metabolic disorders	270–273.9, 275, 275.1, 279–279.9
Cystic fibrosis	277–277.9
Blood disorders	281, 282–283.9, 282–284, 284–284.9, 286–286.9, 288–288.9
Psychosis	290–299.9
Mental retardation	317–318.2
Neural degeneration	330–330.9, 331–331.4, 334–334.9, 335–335.9
Multiple sclerosis and other CNS disorders	340, 341-341.9, 344-344.9, 356-356.9
Cerebral palsy	343–343.9
Epilepsy	345–345.9
Muscular dystrophy	359–359.9
Glaucoma and blindness	365.14, 369
Hearing loss	389, 389.7
Heart disease	424.1, 424.3, 425, 446–446.7
Asthma	493–493.9
Other lung disorders	516
Ulcer	531–534, 532, 533, 534
Enteritis	555–555.9, 556
Other digestive disorders	671, 571.4–571.49, 571.6, 577.1, 579–579.1
Renal disorders	581–581.9, 582–582.9, 583–583.9, 585–586, 588.0–588.1
Lupus	695.4
Connective tissue disorders	710–710.9, 714–714.9
Joint disease	720–720.9, 728
Osteomylelitis	730.1, 732–732.9
Spina bifida	741–741.9
Congenital conditions	742–742.9, 745–745.9, 746–746.9, 747–747.9, 748–748.9, 749–749.25, 750.3, 751.61, 751.62, 752.7, 753, 754.3, 755.2–755.29, 755.3–755.39, 755.55, 756–756.9, 758–758.9, 759.5,759.81–759.89
Mental disorders	300.4, 307.1, 307.5, 309.1–309.9
Premature birth	765
Developmental delay	783.4

Misreporting Medicaid Enrollment: Results of Three Studies Linking Telephone Surveys to State Administrative Records

Stephen J. Blumberg and Marcie L. Cynamon

Are surveys of health insurance coverage accurately reporting the number of children enrolled in Medicaid? Comparisons of survey estimates to totals from administrative records suggest that the answer is no. For example, the number of children reportedly enrolled in Medicaid at any time during 1995 on the Current Population Survey (CPS) was 22.9% lower than the number of Medicaid enrollees under age 18 in administrative databases maintained by the Health Care Financing Administration (HCFA) (Lewis, Ellwood, & Czajka, 1998).

Researchers have speculated on the causes of this underreporting bias. The stigma associated with public assistance may lead respondents to avoid mentioning enrollment in such programs; alternatively, complex eligibility rules and guaranteed eligibility programs may mean that respondents are not aware that their children have been enrolled (or are still enrolled). Sometimes parents become aware of their children's enrollment only when medical services are needed, but medical services may not have been needed recently. Even when medical services are needed, Medicaid managed care plans may operate like, and appear to be, private insurance companies. Then again, respondents simply may not hear the reference period referred to in the question: CPS questions ask about coverage at any time in the past year, whereas respondents may be responding with current coverage information. (Point-in-time estimates will necessarily be lower than "ever-enrolled-last-year" estimates.) Missing the reference period is possible in CPS, considering that the questions come late in the interview at a time when respondents may suffer from fatigue.

To measure the magnitude of this underreporting bias, most researchers have been forced to rely on the validity of HCFA's administrative data. Yet, the number of Medicaid enrollees in these databases will be overstated because children residing in two states during the past year are counted twice, and because the administrative data include institutionalized children not in the survey universe. Thus, comparing annual survey estimates to administrative totals may not paint an accurate picture of the degree of Medicaid underreporting.

The authors are with the Division of Health Interview Statistics at the Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland. We would like to thank Joan Cornett and Jeannine Volcjak for their assistance with data entry, and Pradip Muhuri, Lorayn Olson, and Linda Tompkins for their analytical support.

This paper presents three studies that use an alternative approach to assess the degree of Medicaid underreporting. In studies 1 and 3, state administrative records were used to create sampling frames consisting of children currently enrolled in Medicaid. In study 2, data on children selected during a random-digit-dial (RDD) telephone survey were linked back to state Medicaid enrollment records. This paper will focus on the magnitude and effect of an underreporting bias and on the difficulties encountered when trying to link telephone surveys and administrative records.

Questionnaire

All three studies were part of the State and Local Area Integrated Telephone Survey (SLAITS), conducted by the National Center for Health Statistics (NCHS). From October 1998 to August 1999, SLAITS pilot-tested a new questionnaire: the Child Well-Being and Welfare (CWBW) module. This questionnaire investigates child well-being indicators from the perspective of program participation. These indicators include education, child care, family functioning, household stability, neighborhood characteristics, social development, public assistance program participation, and health insurance. The majority of the questions were drawn from national surveys.

The health insurance questions, however, were newly created for this survey. In contrast to the CPS, which asks about insurance coverage in the past year, the CWBW health insurance questions ask about current insurance coverage. In contrast to the National Health Interview Survey (NHIS), which first asks one global question about health insurance coverage, a series of questions were created that ask specifically about each of the major types of coverage. Each type of health insurance coverage was described in order to avoid confusion between types of coverage (e.g., Medicaid vs. Medicare). For example, Medicaid was described as "a health insurance program for low-income families." Medicaid was the first type of coverage included in the list (see Table 1). The health insurance questions were asked early in the 30-minute interview, preceded only by the household rosters, identification of the focal child, demographic information on the focal child and his or her biological mother, and the relationship of the respondent to the focal child. In all cases, the respondent for this survey was the parent or guardian who

Table 1. Health insurance questions in the SLAITS Child Well-Being and Welfare module

Intro. The next few questions are about health insurance.

1. (MN)[a] Is CHILD covered by Medical Assistance (MA), that is, Minnesota's Medicaid program, a health insurance program for low-income families?
   (TX)[a] Is CHILD covered by Medicaid, a health insurance program for low-income families?
2. Is CHILD covered by Medicare, a health insurance program for the elderly and persons with disabilities?
3. [IF #2 = YES] Is CHILD coverage by Medicare Supplemental Coverage, also known as Medi-Gap plans?
4. Is CHILD covered by the Indian Health Service?
5. Is CHILD covered by Military Health Care, CHAMPUS, CHAMP-VA, or TRICARE?
6. Is CHILD covered by Private Health Insurance—that is, health insurance obtained through employment or unions or purchased directly?
7. (MN)[a] Is CHILD covered by any other kind of health insurance or health care plan that pays for hospital and physician services such as MinnesotaCare or General Assistance Medical Care?
   (TX)[a] Is CHILD covered by any other kind of health insurance or health care plan that pays for hospital and physician services?
8. [IF NO AFFIRMATIVE ANSWERS FOR #1–#7] It appears that CHILD does not have any health insurance coverage to help pay for services from hospitals, doctors, and other health professionals. Is that correct?
9. [IF #8 = NO] What kind of health coverage does CHILD have?

[a]State-specific program names were used in Minnesota, but not in Texas.

knew the most about the focal child's health care, child care, and education.

Study 1: Medicaid Enrollee Sampling Frame in Minnesota

Method

The target population for this study was all children under age 18 in Minnesota currently enrolled in Medicaid. Two weeks prior to the scheduled start of data collection, the Minnesota Department of Human Services provided NCHS with a list of Medicaid enrollees under age 18, including their birth dates, Social Security numbers (SSNs), sex, counties of residence, and ZIP codes. Because data collection was scheduled to last 12 weeks, there was some concern that Medicaid enrollees who were just shy of their 18th birthday would turn 18 (and thus be out of scope for the survey) before their households were contacted. Therefore, children whose 18th birthday would occur within 8 weeks of the list construction were removed from the list. In addition, because the list was created prior to the start of data collection, very recently enrolled children were also effectively excluded from the population.

A sample of 1,360 children was randomly selected from the list of Medicaid enrollees (N = 215,469). This sample was necessarily larger than the survey sample desired because we anticipated that some children on the list would have missing or inaccurate telephone numbers. When telephone numbers were missing or inaccurate (nonworking, nonresidential, fax/modem, temporarily not in service, or no children in the household contacted), new telephone numbers were sought through the use of directory assistance, the National Change of Address Service, CD-ROMs of addresses and telephone numbers, and Internet searches.

The telephone numbers were then called by interviewers who were blind to the Medicaid status of this population. Once an adult aged 18 or older residing in the household was contacted, and after an explanation of the survey that included the authorizing legislation, confidentiality, and voluntary and burden statements, a roster of children under age 18 in the household was obtained. At this point, the computer selected the child whose first name and birth date matched the name and birth date selected from the Medicaid enrollment list. When the birth date and name did not match, an algorithm selected the child that best matched on month of birth, year of birth, and first initial of first name. If no match was possible, but the Medicaid enrollment list indicated a child enrolled in Medicaid at that telephone number, two children were randomly selected for interview. Full names and SSNs were then obtained for these children at the end of the interview. If a match still was not made, the sample interview was omitted from the evaluation. Interviewers and respondents were blind to this selection process.

Results

From the sample of 1,360 children, 640 interviews were completed (47.1%). Of a preliminary data file of 621 completed interviews, 128 interviews (20.6%) were omitted from this evaluation because of failure to match the focal child in the interview to the sample child from the Medicaid enrollment list. These omissions were generally due to inaccurate telephone numbers for these sample children; new telephone numbers were not located.

Among the 493 completed interviews with the correct sample child in the preliminary data file, 392 parents (79.5%) reported current Medicaid coverage as the child's type of health coverage. Of the 101 parents who did not report current Medicaid coverage, 22 reported that the child did not have any current coverage, 41 reported private health insurance coverage, 9 reported Medicare coverage, and 35 reported some other type of health coverage (parents could report multiple types of coverage for their children).

Estimates of Bias

Given a sample of Medicaid enrollees, the Minnesota data suggest that a survey of that sample will underestimate the number of current Medicaid enrollees by 20.5%. This estimate is similar to Lewis, Ellwood, and Czajka's (1998) estimate of a 22.9% underreporting bias in the 1995 CPS. Verdier (1998) also reported similar results: 18.8% of household respondents in Maine who were living with at least one Medicaid enrollee (according to state enrollment files) failed to indicate this coverage on the 1998 Maine Health Insurance Coverage Survey. It is important to note, however, that this Medicaid underreporting bias in Minnesota would lead to only slight bias in estimates of the uninsured. Of those parents misreporting Medicaid coverage, 78.2% still reported some other type of health insurance coverage for their children, suggesting that uninsurance would be underestimated by only 4.5%.

Study 2: RDD Sampling Frame in Texas

Of course, given that children's health insurance coverage may have changed in the time period between the creation of the enrollment file and the interview, some of the "incorrect" responses in study 1 may have been accurate. Thus, 20.5% might best be considered an upper bound on the magnitude of the bias. A better estimate of the degree of underreporting could be obtained using enrollment data accurate on the date of the interview. This was not available in Minnesota, but arrangements were made to have such enrollment data available in Texas.

In addition, the results of the Minnesota study do not indicate the proportion of children not currently enrolled in Medicaid whose parents might erroneously report current Medicaid coverage. For example, Verdier (1998) suggested that "overreporting" of Medicaid coverage was sufficiently high to offset underreporting errors. To explore both underreporting and overreporting errors, a RDD survey of households with children was conducted in Texas using the same questionnaire used in Minnesota.

Method

The CWBW module was implemented in Texas using the SLAITS mechanism. The SLAITS mechanism is typically conducted as a RDD telephone survey that builds off the large sampling frame needed for the National Immunization Survey (NIS) (Ezzati-Rice, Zell, Battaglia, Ching, & Wright, 1995). Because the NIS must screen a large number of telephone numbers to reach its relatively sparse target population of children between 19 and 35 months of age, it provides a cost-effective opportunity to survey other populations. The target population for the SLAITS survey was all households with children under 18 years of age in Texas, and children under 18 years of age in these households. The sample was therefore selected by screening a subsample of Texas households from the NIS sampling frame, which of course excludes households without telephones.

Once an adult aged 18 or older residing in the household was contacted, and after an explanation of the survey, respondents were screened for eligibility for the NIS. Respondents in NIS-eligible households completed the NIS before proceeding with the CWBW screening protocol and the CWBW questionnaire. The CWBW screening protocol included the number of household members, the number of children under age 18, and a general income question. Because the focus of this study was on Medicaid enrollment, a larger sample of low-income households was needed than would be obtained through a simple random sample of households. To accomplish this, respondents were asked whether their household income in the last calendar year was above or below a dollar amount determined to be 200% of the federal poverty level (FPL) based on 1998 DHHS poverty guidelines and the household's size. All households with children and with reported income below 200% FPL were included in the sample. Initially, households with reported income above 200% FPL were subsampled at a rate of 56%. This subsampling rate was later eliminated because the actual number of households reporting income below 200% FPL was greater than anticipated from 1997 CPS data.

Respondents in households screened into the sample then completed a roster of children under age 18 in the household. Children in each household were stratified into two age groups: 0–5 and 6–17. If children were present in both age groups, then one child was selected at random from each age group. If there were children in only one age group, no more than two children were randomly selected from that household.

At the conclusion of the interview, the children's full names, birth dates, and SSNs were obtained to permit links to the Texas Medicaid enrollment database. Prior to asking for this information, interviewers informed parents that the information would be used "to conduct health-related research by linking your survey data with Texas Department of Health coverage information." Parents were reminded that providing this information was voluntary, and interviewers noted if the parents explicitly refused to permit this link. The Texas Department of Health provided us with access to the Texas Medicaid Network (TexMedNet) electronic eligibility verification system, which enabled us to check Medicaid enrollment status for a single individual for any given day. For security reasons, this system enables the user to obtain Medicaid enrollment information only if two of the following three fields are completed: SSN, birth date, and name (first five letters of last name and first letter of first name are sufficient). Parents who refused to provide two of these three fields implicitly refused to permit their survey data to be linked to this enrollment system.

Results

Interviews were completed with 1,265 households, for a total of 2,009 sampled children (686 aged 0–5 years; 1,323 aged 6–17 years). Of these children, 60.2% lived in households with income below 200% FPL. Spanish-speaking interviewers completed 266 of these interviews (21.0%) using a

Table 2. Number of children for whom Medicaid coverage was correctly and incorrectly reported, Texas RDD sample, study 2

	Criterion Standard (Enrollment Records)	
Survey Answer	**Medicaid**	**Non-Medicaid**
Medicaid	196	110
Non-Medicaid	29	1,022

Note: Sample restricted to children whose parents permitted their survey data to be linked to enrollment records. Frequencies are unweighted.

Spanish version of the questionnaire. In 39 households, interviews could not be attempted because a language other than English or Spanish was spoken. The interview completion rate was 87.5%, the screener completion rate was 87.2%, and the household resolution rate was 90.7%. Thus, the overall response rate was 69.9–70.2% (American Association for Public Opinion Research, 1998; Massey, 1995).

Of the 2,009 sampled children, 368 were reported as currently enrolled in Medicaid. Estimates of the proportion of children covered by Medicaid in Texas require survey sampling weights to adjust for the unequal probabilities of selection (because of the oversampling of low-income households and the age stratification), unit nonresponse, and noncoverage of non-telephone households (Frankel, Ezzati-Rice, Wright, & Srinath, 1998). These child-level weights were then adjusted to the known totals of children obtained from 1998 Census projections by age, sex, race, and ethnicity. Using these weights, 16.3% (standard error [SE] = 1.12) of children in Texas were estimated to be enrolled in Medicaid at the time of the survey.

Linking to Medicaid Enrollment Records

To examine both underreporting and overreporting of Medicaid coverage, attempts should be made to link all 2,009 sampled children to the TexMedNet enrollment records. This effort was hindered when 566 parents (28.2%) explicitly refused to permit the link, and 86 additional parents (4.3%) did not provide sufficient information to permit the link. The children of parents prohibiting the link were more likely to be older (aged 6–17), non-Hispanic Caucasian, and living in households with incomes above 200% FPL (all χ^2 [1] > 33, $p < .001$). These children were also less likely to be reportedly enrolled in Medicaid, χ^2 (1) = 50.1, $p < .001$, which may be accurate given their age and family income level. Data for these children were excluded from all subsequent analyses. Without data for these children, the weighted estimate of Medicaid-enrolled children in Texas rose to 19.6% (SE = 1.44).

Of the 1,357 children whose parents permitted the link, 443 (32.6%) were linked, suggesting that these children were currently enrolled in Medicaid or had been in the past. Failure to link indicates one of four possibilities: (1) the child had never been enrolled in Medicaid, (2) the identifying information reported by the parent was not accurate, (3) the identifying information recorded by the interviewer was not accurate, or (4) the identifying information in the TexMedNet system was not accurate. When SSN was reported, linking was more likely: Of the 596 children for whom SSN was reported, 267 (44.8%) were linked, compared with 176 (23.1%) of the 761 children for whom SSN was not reported. Though the uniqueness of SSN undoubtedly provides a better means for linking survey data to enrollment records, the differential linking rate for children with SSNs versus children without SSNs also would have occurred if parents of children on Medicaid were more likely to be aware of and provide their children's SSNs. This possibility is supported by data indicating that parents with lower household incomes ($12,500 or less) were much more likely than parents with higher household incomes (greater than $46,000) to provide the requested SSNs (51.6% compared with 18.9%). Still, the differential linking rate raises some questions about the accuracy of the linking effort.

Despite these inaccuracies (which we attempt to quantify in study 3), data from the TexMedNet enrollment records were used as the criterion standard for diagnostic discrimination analyses of the survey data. Because these analyses assume that the sample represents the population, weighted frequencies were used. The CWBW question on Medicaid coverage had a sensitivity of 0.85 and a specificity of 0.91; the positive predictive value was 0.63 and the negative predictive value was 0.97. Table 2 presents the number of children (unweighted) in each diagnostic discrimination category.

Underreporting Medicaid Coverage

Medicaid coverage was underreported for 29 children whose parents permitted their data to be linked. Using sampling weights, these 29 children represent 14.7% of all children enrolled in Medicaid at the time of the survey. Of these false negatives, 15 children (51.7%) were first enrolled in Medicaid during the past three months, suggesting that some of the underreporting errors were due to recall problems or lack of knowledge about their child's new coverage. Alternatively, some parents may be confusing Medicaid coverage with other types of insurance coverage. For example, Medicaid managed care plans have insurance cards and procedures that closely resemble those of private managed care plans, which could lead to confusion about the child's coverage. Of the 29 false negatives, 10 (34.5%) did report private insurance coverage, though none were enrolled in managed care plans according to the enrollment record. No insurance coverage was reported for 18 of the remaining 19 false negatives.

Overreporting Medicaid Coverage

Medicaid coverage was overreported for 110 children whose parents permitted their data to be linked. Using sampling weights, these 110 children represent 8.5% of all children not enrolled in Medicaid at the time of the survey. Only 19 of these false positives (17.3%) had Medicaid coverage in the past, and for 12 this coverage had been terminated more than six months prior to the survey. This suggests that overre-

porting errors were probably not due to recall problems or lack of knowledge that the coverage had lapsed.

Estimates of Bias

By correcting the false negatives and false positives so that the survey data accurately reflect the Medicaid enrollment status in the linked enrollment files, the sampling weights can be used to estimate the extent to which overreporting of Medicaid coverage offset underreporting of Medicaid coverage. With the corrected survey data, 14.3% (SE = 1.28) of children in Texas are estimated to have been enrolled in Medicaid at the time of the survey. This new estimate indicates that the original estimate (19.6%) was inflated by 36.3%.

These estimates of the magnitude of Medicaid reporting bias are accurate only to the extent that linkage of the survey data to the enrollment records was accurate. When this was not the case, children actually enrolled in Medicaid would not be considered Medicaid enrollees. Reports that these children were enrolled in Medicaid would incorrectly be considered overreports. In addition, reports that these children were not enrolled in Medicaid would be treated as correct responses when in fact they are underreports. In both cases, the extent to which Medicaid appears to be underreported in the survey would be decreased, and the extent to which Medicaid appears to be overreported in the survey would increase.

Study 3: Medicaid Enrollee Sampling Frame in Texas

To explore the extent to which linking was accurate in study 2, study 1 was replicated with a sample of known Medicaid enrollees drawn from a list provided by the Texas Department of Health. If this list was accurate, all children in this sample should link to the enrollment records following the survey. Aside from drawing its sample from a different state, study 3 differed from study 1 in that the list of known Medicaid enrollees was restricted to children enrolled in managed care programs. This targeted sample provided the opportunity to further explore misreporting by parents of children enrolled in Medicaid managed care.

Method

The target population for this study was all children under age 18 in Texas currently enrolled in Medicaid managed care plans. Ten weeks prior to the scheduled start of data collection, the Texas Department of Health provided NCHS with a list of Medicaid managed care enrollees under age 18, including their birth dates, SSNs, sex, counties of residence, and ZIP codes. This list was generated from a database maintained separately from the TexMedNet system. The lengthy delay between the production of the list and the start of interviewing increased the possibility that the children's insurance coverage may have changed. However, this delay was not considered problematic because (1) links to the enrollment records would still occur provided that the child was covered by Medicaid at any time in the past year, and (2) the enrollment records would indicate if the child's coverage had lapsed.

As with study 1, the computer selected the child whose first name and birth date matched the name and birth date selected from the Medicaid enrollment list. The interviews were conducted by the same interviewers and at the same time as the interviews in study 2. Interviewers were blind to the Medicaid status of this population, and they were never told that study 3 was occurring along with study 2.

Results

A sample of 750 children was randomly selected from the list of Medicaid managed care enrollees (N = 287,829). Of the sample selected, 66 cases (8.8%) did not have telephone numbers. When initial calls were made, an additional 303 cases (40.4%) did not have accurate telephone numbers. Correct telephone numbers for 22 of these cases (6.0%) were later identified.

From the 403 remaining numbers, 246 interviews were completed (61.0%). This low response rate was due primarily to difficulties in determining the accuracy of 135 of these telephone numbers. Household rosters were not completed in these households due to lack of contact (36), unsuccessful callbacks (22), refusals/breakoffs (69), or language barriers (8). In only 22 cases were accurate numbers identified, but interviews were not completed (12 due to refusals).

Using the information provided by respondents on the roster of children in the household, the first name and birth date of 123 children (50.0%) perfectly matched those of the sample children drawn from the Medicaid enrollment list. Fifty-six children matched only on month of birth, year of birth, and/or first initial of first name; their data were removed from analyses of linking accuracy (but not from analyses of bias) due to some uncertainty about whether the correct child was selected. Additionally, data from 67 children were omitted from all analyses because the focal child in the interview did not match the sample child from the Medicaid enrollment list.

Linking to Medicaid Enrollment Records

Parents of eight children whose names and birth dates perfectly matched the Medicaid enrollment list (from which the sample was drawn) explicitly refused to permit a link to the TexMedNet enrollment records; one additional parent did not provide sufficient information to permit the link. Of the "perfect matches" for whom linking was permitted, 96 (84.2%) did indeed link to the enrollment records, suggesting that efforts to link the survey data to the TexMedNet enrollment records were often, but not always, successful. As noted earlier, failing to link may artificially increase the number of overreports and decrease the number of underreports in study 2. Assuming that the linking accuracy with this sample is equivalent to the linking accuracy for the RDD sample, the actual underreporting bias for the RDD sample may be more severe than that reported in study 2.

Table 3. Summary of Medicaid coverage underreporting for children enrolled in Medicaid at the time of the interview

Study	Sample	Degree of Underreporting of Current Medicaid Coverage
1	Medicaid enrollees in Minnesota (full sample)	20.5%
2	RDD sample in Texas (permitted links only)	14.7%[a]
3	Medicaid managed care enrollees in Texas (full sample)	13.8%
	Medicaid managed care enrollees in Texas (linked perfect matches only)	0.0%

[a]When overreporting of Medicaid coverage was considered, the resulting survey estimate overstated the number of Medicaid enrolled children by 36.3%.

Estimates of Bias for Linked "Perfect Matches"

Among the 96 "perfect matches" who were linked to the TexMedNet system, these records indicated that 14 children had coverage that lapsed between the creation of the enrollment list and the interview date. All parents of the remaining 82 children correctly reported their children's current Medicaid coverage. Of the 14 children whose coverage had lapsed, however, 4 were reported to still be enrolled in Medicaid. This result suggests that, as in study 2, overreporting may occur more often than underreporting.

Overall Estimates of Bias

Considering instead all 179 completed interviews about focal children who matched (perfectly or partially) the sample children originally selected from the enrollment lists, the TexMedNet system indicated that 20 children (11.2%) had coverage that lapsed between the creation of the enrollment list and the interview date. Four of these children (20.0%) were incorrectly reported to be currently enrolled in Medicaid. In the remaining 159 interviews, 137 parents (86.2%) reported current Medicaid coverage as the child's type of health coverage, 6 reported private coverage, 2 reported "other" coverage, and 15 reported no coverage. Thus, given a sample of Medicaid enrollees, these data suggest that a survey of this sample will underestimate the number of current Medicaid enrollees by 13.8%.

Why did the larger sample reveal underestimates, whereas the sample restricted to "linked perfect matches" did not? One possibility, of course, is that the larger sample included many focal children who were not the children sampled from the enrollment lists. To the extent that these children were, in fact, not enrolled in Medicaid and their parents provided correct answers, underreporting of Medicaid would be falsely observed. Alternatively, the larger sample may have included children whose parents used nicknames or made errors in reporting their children's birth dates. If these parents failed to provide accurate identifying information (i.e., information that was the same as information in TexMedNet), linking would not be possible and coverage that had lapsed would not be identified. Correct reports that their children were not covered by Medicaid would then be mistaken for underreports. Of the 22 parents who did not report current Medicaid coverage, only one linked to the Medicaid enrollment records. (This child was found to be enrolled at the time of the interview.) Had the remaining 21 children been linked and found to have coverage that lapsed, Medicaid underreporting would have dropped to 0.007% (1 out of 138). In sum, the accuracy with which respondents report identifying information about their child may influence whether focal children are matched to the sample children, whether lapses in coverage are correctly identified, and whether underreporting of Medicaid coverage is accurately measured.

Discussion

The reporting of Medicaid coverage in population-based surveys is lower than the number of persons enrolled in Medicaid, according to administrative data. While there is no agreement on the exact level of underreporting, many health care research experts suspect that about 20% of the parents of Medicaid recipients do not report that their children currently receive Medicaid. The three studies reported here observed levels of underreporting that ranged from 13.8% to 20.5% (see Table 3).

Because of this underreporting, survey organizations have logically imputed Medicaid coverage to persons likely to be enrolled in Medicaid. For example, the Urban Institute uses its Transfer Income Model (TRIM2) to assign Medicaid coverage in the CPS to some children who would appear to qualify for Medicaid based on household income levels. For its part, the Census Bureau has imputed Medicaid to children receiving AFDC (Aid to Families with Dependent Children) or "other public assistance" (which typically meant AFDC). Children receiving SSI (Supplemental Security Income) in states with mandatory Medicaid coverage for SSI recipients are also assigned coverage. These logical imputations increased the CPS estimate for Medicaid enrollment by 19% in 1995 (Lewis et al., 1998).

Further analyses of the present data are planned to explore the accuracy of these adjustment strategies and to suggest new adjustments based on comparisons between children with accurate and inaccurate coverage data. There is reason to believe, however, that these adjustments may not be entirely appropriate. In the present research, much of the perceived underreporting of Medicaid coverage may be an artifact of difficulties in matching survey focal children to Medicaid enrollment lists and/or records. The enrollment lists from both Minnesota and Texas were fraught with missing or inaccurate telephone numbers, making contact with the sample children difficult. We consider it likely that, even when contact was made and interviews completed, some of the interviews did not focus on the correct sample children. In studies 1 and 3, at least one in five focal children proved to be the

incorrect sample child, and the number of imperfect matches suggests that some of these may have been incorrect as well. Study 2 was not hindered by an imperfect enrollment list, but study 3 suggests that a significant portion of study 2's sample may have been enrolled in Medicaid yet remained unlinked to the enrollment records due to poor identifying information from the parent or imprecise record keeping by the state. The effects of these inaccuracies were profound when compared to the "linked perfect matches" in study 3. No underreporting was observed for this sample that perfectly matched the enrollment list and accurately linked to the TexMedNet enrollment records.

We are therefore left in a quandary. There are some suggestions that recall and knowledge problems affect Medicaid reporting: Half the false negatives in study 2 were recently enrolled, and parents may not have been aware of this coverage. There are other suggestions that recall and knowledge problems are not affecting Medicaid reporting: Very few parents of Medicaid managed care enrollees in studies 2 or 3 incorrectly reported private coverage instead of Medicaid. The perceived stigma of Medicaid enrollment may still play a role in Medicaid reporting, although the present research could not shine any light on this topic. So, while continued research including known Medicaid enrollees in surveys, or linking survey respondents to Medicaid records, would seem warranted, we fear that future researchers will find that these gold standards are tarnished by difficulties in connecting focal children with these enrollment lists and records.

References

American Association for Public Opinion Research. (1998). *Standard definitions: Final dispositions of case codes and outcome rates for RDD telephone surveys and in-person household surveys*. Ann Arbor, MI: Author.

Ezzati-Rice, T. M., Zell, E. R., Battaglia, M. P., Ching, P., & Wright, R. A. (1995). The design of the National Immunization Survey. In *1995 Proceedings of the Section on Survey Research Methods* (pp. 668–672). Alexandria, VA: American Statistical Association.

Frankel, M. R., Ezzati-Rice, T. M., Wright, R. A., & Srinath, K. P. (1998). Use of data on interruptions in telephone service for noncoverage adjustment. In *1998 Proceedings of the Section on Survey Research Methods* (pp. 290–295). Alexandria, VA.: American Statistical Association.

Lewis, K., Ellwood, M., & Czajka, J. L. (1998). Counting the uninsured: A review of the literature. In *Assessing the new federalism*, Occasional Paper No. 4. Washington, DC: Urban Institute.

Verdier, J. M. (1998). *Measuring, monitoring, and reporting on state children's health insurance programs: A primer for state officials*. Washington, DC: American Public Human Services Association.

The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits

Robert F. Belli, James M. Lepkowski, and Mohammed U. Kabeto

Introduction

The quality of survey results are influenced by many factors, including those that are part of the survey interviewing process. Both cognitive processing of a survey request and the conversational rapport established during the interview appear to have important influences on the quality of the information obtained in an interview.

Theoretical models of the survey interview typically involve several cognitive processes, including comprehension of the question and its objectives, information retrieval strategies and processes, respondent and interviewer judgment of whether retrieved information is relevant to question objectives, respondent edits of responses to meet social desirability concerns, and respondent and interviewer edits of responses to meet question objectives (Sudman, Bradburn, & Schwarz, 1996; Tourangeau, 1984). Difficulty in any one of these cognitive processes may reduce the quality of survey reports and resulting estimates from the survey data.

For retrospective reports specifically, there is substantive evidence that a number of factors influence the effectiveness of the cognitive processes respondents use while reporting on their pasts. Difficulties in cognitive processing emerge if the target of the report is misinterpreted; if the events occurred long ago, frequently, and are nondistinctive in memory; if respondents fail to use rate-based information for regularly occurring behaviors; and if survey researchers fail to provide adequate cues (e.g., Belli, 1998; Menon, 1997).

The role of conversational rapport in survey interviewing has been more elusive to study than the role of cognitive processing. Some definitions of rapport address interviewing conditions conducive to meeting research goals, whereas others emphasize developing harmonious relationships between the survey participants (Goudy & Potter, 1975). Operationally, measures of rapport have included such measures as interviewers' assessments of their relationships with respondents (Weiss, 1968), experimental manipulation of interviewer instructions (Dijkstra 1987; Henson, Cannell, & Lawson 1976), and observations concerning the types of statements that characterize friendliness or digressions from the objectives of the survey questions (Belli & Chardoul 1997; Houtkoop-Steenstra 1997). Equivocal patterns of relationship between rapport and data quality have been found and are due in part to variation in definitions and measures. For example, Dijkstra (1987) found that rapport improves response quality, while Weiss (1968) found the opposite.

Conversational rapport may increase the motivation to cooperate with the survey request and thereby influence the accuracy of responses. Theory suggests that if optimal conversational rapport could be created, it would represent a situation in which respondents are motivated to try hard to answer even the most cognitively taxing of questions (Dijkstra, 1987; Henson et al., 1976) or are comfortable revealing potentially embarrassing information (Weiss, 1968; Williams, 1968). Some have argued that rapport has a curvilinear relationship with response quality, in which too little or too much rapport is detrimental (Dijkstra, 1987; Williams, 1968).

For purposes of reducing response errors in surveys, the survey interviewing task is typically standardized. Interviewers are required to ask questions as written, to recognize inadequacies in responses by probing respondents nondirectively, and to encourage the motivation of respondents by providing appropriate and behavior-reinforcing feedback (Beatty, 1995; Brenner, 1985; Cannell, Miller, & Oksenberg, 1981). Yet there is concern that standardization inhibits certain beneficial conversational processes (Clark & Schober, 1992; Suchman & Jordan, 1990). Whereas conversational flexibility allows the reframing of intent and the clarification of meaning, in standardized survey interviewing, interviewers are discouraged from reducing ambiguities in question content, since by so doing they could influence their respondents to extract certain interpretations over other ones (Schober & Conrad, 1997).

Despite the constraints on the ordinary processes of conversation that are imposed by standardization, interviewers and respondents in practice use conversational processes as a context for the interview. Interviewers and respondents develop a rapport concerning their relationship to one another and to the standardized nature of the survey instrument to which they devote their attention (Belli & Chardoul, 1997; Clark & Schober, 1992). At times, conversational rapport may be expressed in a manner that is at odds with the goals of standardization (Belli & Chardoul, 1997; Houtkoop-Steenstra, 1996).

In the present research, we are interested in detecting the global influences of cognitive difficulty and rapport on response accuracy within the natural context of the survey interview. Our methodology is based on coding the audiotaped verbal behaviors of interviewers and respondents that indicate the respective presence of respondent cognitive difficulty in answering questions and conversational rapport

The authors are at the University of Michigan.

between participants. Although our methodology is unable to advance a greater understanding of the roles of specific cognitive and conversational processes on response accuracy, we are able to determine in the context of an actual survey interview whether global expressions of cognitive difficulty or rapport are associated with the accuracy of responses.

Methods

Our investigation of the influences of cognitive processing difficulty and rapport on the response accuracy of retrospective reports of doctors' office visits involves the combination of several procedures associated with data collection and analysis. In this section, we describe the data collection procedures of conducting the survey interview and our coding of verbal behaviors.

Survey and Respondents

The Survey Research Center at the University of Michigan administered a survey about health and health care utilization in 1993 using standardized survey interviewing techniques in face-to-face mode to a probability sample of 2,006 members of an HMO in the Detroit metropolitan area. Interviews lasted approximately one hour, with questions about hospital stays, number of visits to health care providers, health insurance coverage, health expenditures, and the presence of medical conditions. African Americans, individuals who were between the ages of 14 and 17 years, and individuals 65 years old or older were selected at higher rates to provide adequate representation within the sample for comparative purposes. A total of 1,834 respondents gave permission for audiotaping and the release of medical records information from their HMO.

Following data processing, a sample of audiotapes were selected for coding of verbal behaviors during the survey interview. Controlling for respondent age, gender, race, and the date of interview, a systematic sample of 317 taped interviews were selected for a coding operation. After the deletion of 21 taped interviews for which the number of audible questions was less than 30, a sample of 296 audiotape-recorded interviews were available for coding.

Coding of Verbal Behaviors

We employed a small staff of survey interviewers (different from those who conducted the interviews) to code a variety of verbal and other behaviors from the audiotaped interviews. Table 1 summarizes the coding scheme to which each question-answer exchange between interviewer and respondent was assessed. The coding process was designed to record the presence or absence of these behaviors for each question-asking exchange. For convenience of reference, behaviors are grouped on the basis of interviewer question-asking behaviors (including repeating questions), respondent answering behaviors, interviewer probing and feedback behaviors, and the conversational behaviors of laughter and digressions.

Table 1. Verbal behavior codes

	Interviewer Question-Asking Codes
Q-E	Exact: Reads exactly as written or makes insignificant changes.
Q-S	Significant changes: Makes wording changes that can affect written question meaning.
Q-O	Other changes: Verifies, states, or suggests an answer; reads nonapplicable question; skips applicable question.
RQ-E	Exact repeating of question.
RQ-S	Significant changes in repeating question.
	Respondent Answering Codes
R-I	Interruption: Interrupts question with an answer.
R-C	Clarification: Expresses uncertainty, requests question repetition, or seeks clarification.
R-Q	Qualified response: Qualifies answer with phrases such as about, I guess, maybe, etc.
R-U	Uncodable/inadequate response: Response does not meet question objectives.
R-DK	Expressions of "don't know" that occur before a final codable response is given.
R-CR	Respondent corrects a response to a previous question.
	Interviewer Probing Codes
P-A	Adequate probing: Probing is nondirective and sufficient.
P-D	Directive probing: At least one probe is directive.
P-U	Underprobing: Failure to probe in situation that requires it.
	Interviewer Feedback Codes
F-AS	Acceptable short: Neutral and appropriate short phrase (1–3 words) such as "Thank you."
F-AL	Acceptable long: Neutral and appropriate longer phrase such as "Thanks. That's useful information for our study."
F-US	Unacceptable short: Offers short phrase that may indicate approval for the content of the response.
F-UL	Unacceptable long: Offers longer phrase that may indicate approval for the content of the response.
F-UR	Unacceptable reward: Approval for a "don't know" response, refusal, digression, interruption, or inadequate final answer. Includes a digression that follows a respondent digression.
	Interviewer Conversational Codes
I-D	Interviewer introduces digression: Digressions are verbal comments that are not directly related to satisfying question objectives.
I-L	Interviewer laughs.
	Respondent Conversational Codes
R-D	Respondent digresses.
R-L	Respondent laughs.

Regarding cognitive processing, many of the respondent answering codes reflect difficulties in comprehension, retrieval, judgment, and response formatting (Fowler & Cannell, 1996) (see Table 1). Interruptions may indicate that

respondents are uncertain about the appearance of response options. Requests for clarification indicate the presence of interpretive problems with questions or response options. Inadequate answers also may be the result of interpretive problems, but also may more extensively signal the inability to retrieve information, difficulty in judging which information is relevant to the objectives of the question, or difficulties in formatting one's internal answer to response requirements.

Interviewer probing and repeating questions are also indicative of cognitive problems, as interviewers either probe or repeat questions whenever they notice that respondents are having difficulty understanding a question or developing a response. Yet directive probing may be more indicative of a communicative agreement between interviewer and respondent that lies outside of the direct survey-related tasks of the interview, and less of a verbal behavior that indicates cognitive difficulty with the question (Houtkoop-Steenstra, 1996). Directive probing can have desirable communicative effects, as its use can simplify respondent burden. At times, respondents will provide a range of numbers (e.g., "five or six") as a response. By offering a single option (e.g., "well, let's say six") as a directive probe, an interviewer may be tacitly communicating his or her understanding that the respondent is willing to fulfill the spirit of the interview but is also not motivated to work too hard.

As for conversational rapport, the interviewer feedback and the interviewer and respondent conversational behaviors are most noteworthy (see Table 1). The feedback behaviors follow responses and are considered aids toward motivating respondents to work hard at the task when they are presented in a neutral manner, that is, in a manner acceptable to standardized survey interviewing (Cannell et al., 1981). These acceptable feedback phrases typically thank respondents for answering. Feedback phrases that are unacceptable to standardized interviewing are those that may bias responses because they indicate approval of the content of the response. For example, an interviewer who expresses positive feelings toward a respondent's good health may encourage the underreporting of health visits, and one who empathizes with poor health may encourage overreporting. Conversely, unacceptable feedback may also help to motivate respondents to try hard to answer questions. Respondents who sense that interviewers are generally concerned with their personal situation may be motivated to reciprocate by doing their best in responding to the task requirements of the survey.

The interviewer and respondent digression and laughter behaviors are most indicative of the establishment of a personal attachment between interviewers and respondents. Many verbal digressions appear as attempts to build a personal relationship between the participants, and laughter may indicate that the survey interviewers and respondents are getting along well (Belli & Chardoul, 1997). Although developing a personal attachment may motivate respondents to perform at their best, it may also bias responses if there is an attempt to ingratiate interviewers or if an orientation is developed toward the interviewing task that is less than optimal from the perspective of the survey researcher (Dijkstra, 1987). For example, Houtkoop-Steenstra (1996, pp. 220–221) observed respondent laughter following an interviewer directive probe ("I'll just put six times") concerning the number of visits to a museum during elementary school. According to Houtkoop-Steenstra, laughter in this instance was communicating agreement with the interviewer about a "little secret" that no one other than the participants "needs to know."

Five coders, all with interviewing experience, were trained in group sessions to code the behaviors of interest. Group sessions combined with practice coding exercises were used to improve coding reliability. Follow-up group sessions concentrated on resolving coding differences. All audiotapes were coded independently by a single coder.

Results

Analyses focused on determining the prevalence and reliability of code assignments, determining a measure of agreement between survey reports of office visits and those recorded in an HMO database, simplifying the coding scheme into factors that represent cognitive difficulty and conversational rapport, and conducting regression models that attempt to control for factors that might confound the relationships of interest and are a direct product of how survey interviews are conducted.

Frequency and Reliability of Code Assignments

The frequency of each behavior obtained in the coding process for all 296 respondents across all questions asked is shown in the last two columns of Table 2. Some behaviors occur very infrequently, such as significant changes in question wording during a repeated asking of the question. Others, such as exact question reading, occur very often. The frequencies shown in Table 2 are similar to those observed in other studies of interviewer and respondent behaviors in surveys conducted by the Survey Research Center.

For purposes of measuring the reliability of code assignments between coders, 24 audio tapes were coded twice by two different coders. Codes from these double-coded tapes were compared to assess intercoder reliability. Kappa statistics (κ), which control for chance agreement in comparisons between coders, provide an estimate of intercoder reliability (see Fleiss, 1973). Values of κ between 0.21 and 0.40 are considered to be an indication of a fair level of agreement; values between 0.41 and 0.60 indicate moderate agreement; values between 0.61 and 0.80, substantial agreement; and values between 0.81 and 1.0, almost perfect agreement (Landis & Koch, 1977). Table 2 presents κ values for the coded behaviors in the sample of 296 respondents. Almost all of the codes reached what would be considered fair or moderate agreement. Only the RQ-S behavior failed to reach an acceptable agreement level of $\kappa = 0.21$, and this behavior was excluded from further analysis.

Accuracy Assessment

From the entire survey, reports on approximately 25 health care utilization and health characteristics could be compared

Table 2. Kappa values for double-coded cases (n = 24) and the frequency of behaviors (n = 296)

Item	Kappa-Value	ASE*	Average	SD
Q-E	0.511	0.028	0.946	0.141
Q-S	0.432	0.032	0.043	0.121
Q-O	0.467	0.099	0.012	0.042
RQ-E	0.722	0.033	0.051	0.090
RQ-S	0.181	0.157	0.002	0.016
R-I	0.652	0.031	0.037	0.080
R-C	0.783	0.023	0.096	0.127
R-Q	0.538	0.035	0.065	0.106
R-U	0.390	0.033	0.115	0.190
R-DK	0.643	0.071	0.013	0.062
R-CR	0.769	0.092	0.011	0.038
P-A	0.676	0.022	0.179	0.242
P-D	0.283	0.050	0.025	0.075
P-U	0.303	0.066	0.008	0.032
F-AS	0.700	0.016	0.282	0.241
F-AL	0.441	0.064	0.041	0.083
F-US	0.314	0.045	0.138	0.216
F-UL	0.548	0.046	0.015	0.061
F-UR	0.430	0.050	0.018	0.057
I-D	0.323	0.093	0.008	0.039
I-L	0.574	0.040	0.016	0.056
R-D	0.371	0.044	0.047	0.132
R-L	0.639	0.031	0.028	0.069

*Approximate standard error

with data contained in HMO records. The remaining analyses are limited to one item of the survey asking about doctor's office visits. This item is the culmination of a series of questions asking about health care received from a medical doctor or assistant at hospital emergency rooms, urgent care centers, doctor's offices, and any other health care facility. Within this series, respondents were also asked if the health visits made to emergency rooms or urgent care centers were because of injury, accident, or poisoning. In an experimental manipulation, one-half of the respondents in the survey were asked about doctor's visits during the previous 6-month period beginning with the first of the month 6 months prior to the date of the interview. The other half were asked about doctor's office visits for the previous 12-month period. The interviewer summed the number of reported visits and confirmed that the sum was correct. Responses to the health visits summary question were compared against HMO records, and the absolute values of the differences were used as a measure of response accuracy. Since 12 respondents did not provide their number of medical visits, analyses that involved reports were based on 284 respondents.

The medical records are known to be incomplete, failing to capture visits made to facilities outside of the HMO (Jay, Belli, & Lepkowski, 1994). A sequence of questions involved detailed queries concerning the last visit to a medical care provider, including where care was received. Of the entire sample of 296 respondents, 15% reported having last visited a medical facility that was outside of the HMO. As a result, we expected respondents to report visits that would not appear in the medical records. Thus, the measure of accuracy is not a complete assessment of the quality of reported doctor's office visits relative to the true number. However, it serves as a useful measure of the quality of the survey reports. At a minimum, it can be interpreted as a difference between survey reports and an often-used measure of the frequency of visits, administrative records.

The average absolute deviation between survey reports and HMO records on the total number of doctor's office visits was 4.32 (standard deviation = 11.54). The average absolute difference varied by the length of the reference period: 2.60 (standard deviation 4.30, n = 147) for the six-month reference period and 6.17 (standard deviation 15.83, n = 137) for the 12-month period. Of the 284 respondents, 21% had perfect agreement between their survey report and their HMO record on total number of doctor's office visits, and 25% had a difference of only 1. However, some large differences were observed, with one individual having a difference of 133 visits.

Code Reduction

The large number of verbal behaviors and the availability of coded measurements of each behavior for every survey question complicated assessments of the levels of cognitive difficulty and conversational rapport. The investigation was limited to verbal behaviors associated with questions immediately prior to the survey question on the number of doctor's office visits during the last 6 or 12 months. Each of the behaviors was coded for each question in a series of health visit questions: visits to hospital emergency rooms, urgent care centers, doctor's offices, and any other health care facility. Due to skip patterns, respondents were asked a different number of questions in the sequence. For each interview, the total number of times a code was assigned was divided by the number of questions that were asked in the series. This provided a code assignment average for each case across the sequence of questions in the series of interest.

An a priori partitioning of behaviors was made into two groups: those that were expected to correspond to conversational rapport and those related to cognitive processes. The question-asking behaviors of Q-E, Q-S, and Q-O were excluded from the process since they were not considered to be direct indicators of either personal rapport or respondent cognitive difficulty, as question asking is the most highly constrained verbal behavior in the context of survey interviewing. An exploratory factor analysis with an orthogonal rotation yielded two factors with eigenvalues greater than 1.0, which were labeled "rapport" and "cognitive difficulty," respectively (see Table 3 for estimated factor loadings). For the most part, the behaviors in each factor confirmed the a priori expectations. As expected, feedback, laughter, and digression behaviors loaded on the rapport factor (factor 1). Also as expected, many of the respondent answering behaviors did load on the cognitive difficulty factor (factor 2), with the exception of respondent interruptions. It is also

Table 3. Estimated rotated (orthogonal) factor loadings based on maximum-likelihood estimated factor analysis

Variable	Factor 1	Factor 2	Communality
RQ-E	0.09	**0.40**	0.17
R-I	**0.28**	0.04	0.08
R-C	0.08	**0.54**	0.30
R-Q	–0.10	0.08	0.02
R-U	0.20	**0.63**	0.43
R-DK	0.10	0.13	0.03
R-CR	–0.03	**0.32**	0.10
P-A	0.25	**0.71**	0.56
P-D	**0.40**	0.15	0.18
P-U	–0.08	0.12	0.02
F-AS	–0.01	0.04	0.002
F-AL	0.13	0.12	0.03
F-US	**0.32**	0.08	0.11
F-UL	**0.31**	–0.06	0.10
F-UR	**0.59**	–0.02	0.35
I-D	**0.39**	–0.03	0.15
I-L	**0.46**	0.03	0.21
R-D	**0.47**	0.07	0.22
R-L	**0.52**	0.06	0.27

Bold: ≥0.20 in absolute value

important to note that directive probing signaled more the presence of conversational rapport than cognitive difficulty, as it too loaded on the rapport factor. In summary, the rapport factor consisted of interruptions (R-I), directive probing (P-D), unacceptable short feedback (F-US), unacceptable long feedback (F-UL), unacceptable reward (F-UR), interviewer digression (I-D), interviewer laughter (I-L), respondent digression (R-D), and respondent laughter (R-L). The cognitive difficulty factor included the exact repeating of a question (RQ-E), respondent requests for clarification (R-C), uncodable or inadequate responses (R-U), a correction to a previous response (R-CR), and acceptable probing (P-A).

Confirmatory factor analysis was used to test further whether the observed factor structure fit the data. Initial fit of a simple two-factor structure provided a promising but inadequate fit. Modification indices indicated that lack of fit was due to substantial covariance between errors of behaviors within a factor, as well as a modest correlation between the two latent factors (0.24). When a revised model incorporating the larger error covariances and the correlation between factors was used, measures of fit improved to the acceptable range (goodness-of-fit index 0.96, adjusted goodness-of-fit index 0.94, and normed fit index 0.97). A cognitive and rapport code was computed for each case as the simple sum of the behaviors associated with each latent factor. The mean cognition score was 0.45 (standard deviation = 0.49), with a range of 0 to 3.43, and the mean rapport score was 0.21 (standard deviation = 0.33), with a range of 0 to 2.57. Of the 296 respondents, 68 (23%) and 88 (30%) had cognition and conversational rapport scores of 0.

Statistical Methods

Since the absolute values of the differences between survey reports and HMO records were a function of count measures, a Poisson regression model seemed appropriate for the statistical analysis. This model assumes that the mean and variance of outcomes are identical. The Poisson regression model did not fit well because of overdispersion. A negative binomial regression, which models count data that violate the Poisson regression assumption of equality of mean and variance, was used instead. Model fit as assessed with a dispersion index was excellent. Using a generalized linear model with a log link, models were estimated for the log expected absolute difference. Results were summarized in terms of a 95% confidence interval for an odds ratio.

Interviews were clustered by interviewer assignment as well as by coder assignment. Behaviors were expected to be somewhat homogeneous within interviewer or coder assignment, violating standard statistical assumptions for the negative binomial model of independent observations. A modified jackknife variance estimation procedure was employed to examine the extent to which the standard negative binomial results were biased by the lack of independence of sample selections. Interviews were grouped by interviewer as well as by coder. For the interviewer grouping analysis, a separate model was estimated for the sample remaining after each interviewer's interviews were dropped. (A similar analysis was done by coder grouping, but the interviewer grouping produced the largest increases in variance and was the method used here.) The variances of the estimated coefficients were computed as the variability of the jackknife estimated coefficients relative to the estimate computed for the sample of 284 respondents. Variances computed using the interviewer jackknife procedure were larger than those obtained from standard statistical software, confirming suspicions of correlations among behaviors within interviewer assignments. Standard errors of final results were adjusted upward to account for the increased variance due to within-interviewer correlation.

Statistical Model Findings

Using a negative binomial model, we first tested separate models that regressed the accuracy measure on the cognitive difficulty and conversational rapport scores, respectively. When cognitive difficulty was the only predictor, a one-standard-deviation increase from the mean cognition score significantly increased the absolute mean difference between survey reports and HMO records nearly twofold (odds ratio [OR] = 1.87, 95% confidence interval [CI] = 1.65–2.13). When rapport was the only predictor, there was a nonsignificant decrease in absolute differences as rapport scores increased by one standard deviation from the mean (OR = 0.91, 95% CI = 0.79–1.05).

We also regressed the accuracy measure on the cognitive difficulty and rapport scores controlling for respondent age, gender, and level of education; number of medical visits; and an indicator for the 6- and 12-month reference periods. Table 4

Table 4. Parameter estimates and odd ratios from negative binomial models

	Model 1		Model 2	
Variable	**OR**	**95% CI**	**OR**	**95% CI**
Cognition	1.91	1.42–2.58	2.74	1.82–4.13
Rapport	0.98	0.70–1.38	0.96	0.68–1.35
Reference period				
12-month (referent)	1.00		1.00	
6-month	0.72	0.54–0.97	1.02	0.69–1.50
RefPeriod × cognition			0.47	0.28–0.82

presents the odd ratios and 95% confidence intervals for a model that consider only the main terms (model 1) and for a model that includes an interaction between the cognition and an indicator of the length of the reference period (6- or 12-month, model 2). As can be seen for model 1, an increase in cognitive difficulty scores is significantly associated with an increase in the absolute difference between reports and HMO records, whereas the rapport scores are nonsignificantly associated with accuracy. Table 4 reports odds ratios in terms of unit changes. However, reporting the odds ratios in terms of standard deviation units is more appropriate for cognition and rapport scores, since the range of values is narrow for both variables. Accordingly, for a one-standard-deviation increase from the mean cognition score, the absolute mean difference between survey reports and medical records increased by 38% (OR = 1.38; 95% CI = 1.19–1.59), holding all other variables constant. That is, as cognition problems increased, there was less agreement between survey reports and HMO records. Adjusting for other variables, a one-standard-deviation increase from the mean rapport score produced a nonsignificant mean difference between survey reports and HMO records that decreased by 1% (OR = 0.99, 95% CI = 0.86–1.14).

An unadjusted comparison, not shown in Table 4 between the 6-month and 12-month reference periods reveals an odds ratio of 0.42 (95% CI = 0.32–0.55), which means that the absolute difference between survey reports and medical records was reduced by 58% among respondents in the 6-month reference period in comparison to those in the 12-month reference period. As shown in model 1 in Table 4, adjusting for other variables, we also found that the 6-month visits had greater correspondence between survey reports and medical records in comparison to the 12-month visits. Interestingly, the interaction term between the length of the reference period and cognition scores in model 2 shows that the association between cognitive difficulty and the absolute difference was not the same within the 6-month and 12-month reference periods (OR = 0.47, 95% CI = 0.28–0.82). For the 6-month reference period, a one-standard-deviation increase from the mean cognition score increases the absolute difference between survey reports and HMO records by 14% (OR = 1.14). For the 12-month reference period, however, a one-standard-deviation increase from the mean cognition score increases the absolute difference between survey reports and HMO records by 69% (OR = 1.63, 95% CI 1.34–2.00). That is, verbalizations indicating cognitive difficulty have a larger effect on absolute differences for a 12-month reference period than for a 6-month reference period, which should be expected given that longer reference periods place greater memory and cognitive demands on survey respondents.

Importantly, we also tested whether conversational rapport might be curvilinearly associated with accuracy, in which either too little or too much rapport decreases data quality but just the right amount improves it (see Dijkstra, 1987; Williams, 1968). A model that included a curvilinear term for rapport did not find a significant association between this term and the absolute difference between records and reports.

Discussion

Within the context of a standardized survey interview, we found that whereas verbal expressions of respondent cognitive difficulty are indications of poorer accuracy in the retrospective reports of doctor's office visits, verbal expressions that indicate the presence of conversational rapport between interviewers and respondents are not associated with response accuracy. Interviews that are marked by a high level of rapport are characterized by a cluster of verbal behaviors that are discrepant from the ideals of standardized interviewing, including inappropriate interviewer feedback, directive interviewer probing, and digressions and laughter from both interviewers and respondents. Given the emphasis among many survey methodologists regarding the biasing dangers of verbal behaviors that are discrepant from the ideals of survey interviewing (Beatty, 1995; Brenner, 1985; Cannell et al., 1981), particularly providing inappropriate feedback and directive probing toward response quality, our rapport results are rather surprising. One possibility is that the presence of conversational rapport introduces processes that are simultaneously beneficial and detrimental to response quality. Rapport may increase the motivation of respondents to answer challenging retrospective questions correctly while simultaneously introducing biases (Dijkstra, 1987; Goudy & Potter, 1975). Another possibility is that within the constraints of standardized interviewing and interviewer training, the verbal behaviors that represent rapport are not severe enough to warrant a decrease in response quality.

The association that we found between cognitive difficulty and the accuracy of retrospective reports is consistent with prior work showing that respondent problem behaviors, such as expressions of uncertainty and inadequate responses, are significant indications of poorer data quality (Belli & Lepkowski, 1996). Particularly noteworthy is the finding of an interaction between cognitive difficulty and the length of the reference period. The association of cognitive difficulty and the accuracy of retrospective reports was significantly more pronounced in the longer (12-month) reference period than in the shorter (6-month) one. There is no doubt that longer reference periods introduce a more difficult cognitive task than shorter ones. Apparently, those cognitive problems that are encountered with more difficult retrieval tasks are more unresolvable than those associated with easier tasks.

Our results have implications regarding the most effective directions that should be pursued by survey methodologists toward improving the quality of retrospective reports. In the confines of a structured survey interview, our results point to the need to focus attention on reducing the cognitive difficulty that is encountered by respondents in providing retrospective reports. The quality of retrospective reports will likely benefit by improving the clarity of the questions and with provision of effective cues. Regarding the latter, benefits can accrue by imposing decompositional and event-historical cues that are tailored to the idiosyncratic experiences of respondents (Belli, 1998; Means & Loftus, 1991; Menon, 1997).

References

Beatty, P. (1995). Understanding the standardized/non-standardized interviewing controversy. *Journal of Official Statistics*, *11,* 147–160.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, *6*, 383–406.

Belli, R. F., & Chardoul, S. A. (1997, May). The digression of survey actors in a face-to-face health interview. Paper presented at the annual conference of the American Association for Public Opinion Research, Norfolk, VA.

Belli, R. F., & Lepkowski, J. M. (1996). Behavior of survey actors and the accuracy of response. In *Health Survey Research Methods: Conference Proceedings* (pp. 69–74). DHHS Publication No. (PHS) 96–1013.

Brenner, M. (1985). Survey interviewing. In M. Brenner, J. Brown, & D. Canter (Eds.), *The research interview: Uses and approaches* (pp. 9–36). London: Academic Press.

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

Clark, H. H., & Schober, M. F. (1992). Asking questions and influencing answers. In J. M. Tanur (Ed.), *Questions about questions* (pp. 15–48). New York: Russell Sage Foundation.

Dijkstra, W. (1987). Interviewing style and respondent behavior: An experimental study of the survey-interview. *Sociological Methods & Research, 16,* 309–334.

Fleiss, J. L. (1973). *Statistical methods for rates and proportions*. New York: Wiley.

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

Goudy, W. J., & Potter, H. R. (1975). Interview rapport: Demise of a concept. *Public Opinion Quarterly, 39,* 529–543.

Henson, R., Cannell, C. F., & Lawson, S. (1976). Effects of interviewer style on quality of reporting in a survey interview. *Journal of Psychology, 93*, 221–227.

Houtkoop-Steenstra, H. (1996). Probing behaviour of interviewers in the standardised semi-open research interview. *Quality and Quantity, 30,* 5–230.

Houtkoop-Steenstra, H. (1997). Being friendly in survey interviews. *Journal of Pragmatics*, *28*, 591–623.

Jay, G. M., Belli, R. F., & Lepkowski, J. M. (1994). Quality of last doctor visit reports: A comparison of medical record and survey data. In *American Statistical Association 1994 Proceedings of the Section on Survey Research Methods* (pp. 362–372).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical agreement. *Biometrics, 33*, 159–174.

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

Menon, G. (1997). Are the parts better than the whole? The effects of decompositional questions on judgments with frequent behaviors. *Journal of Marketing Research*, *34*, 335–346.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*, 576–602.

Suchman, L., & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*, 232–241.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Weiss, C. H. (1968). Validity of welfare mothers' interview responses. *Public Opinion Quarterly, 32,* 622–633.

Williams. J. A. (1968). Interviewer role performance: A further note on bias in the information interview. *Public Opinion Quarterly 32*, 287–294.

Use of Geographic Contextual Variables in Examining Survey Item Validity

Catharine W. Burt

Introduction

The National Center for Health Statistics has used several methods to explore the validity of its survey data. For the National Ambulatory Medical Care Survey (NAMCS), one of the Center's records-based provider surveys, the methods have tended to rely on various forms of consistency checks to assess the quality of the data. The consistency checks have been in the form of either data edits (e.g., no pregnant males), consistency of estimates from year to year for items (e.g., percentage of visits by females the same from 1997 to 1998), consistency of data processing (e.g., 10% sample quality control on coding), robustness of the estimator (e.g., total estimates of visits similar from year to year despite different sample of providers), and consistency of the abstract procedure (e.g., sample of medical records reabstracted). Most of the methods are really assessing the reliability of the survey data rather than the validity. Validity of the data goes beyond consistency, although from a theoretical view an unreliable estimate cannot have much validity. This paper examines some of the methods that have been used to assess the validity of the NAMCS in the past few years, highlighting the latest study, which uses geographical contextual variables to test hypotheses about the relationship between NAMCS estimates and similar existing data elements from other sources.

Survey Background

The NAMCS is an annual probability survey of patient encounters with nonfederal, office-based physicians excluding the specialties of radiology, pathology, and anesthesiology. The content of the survey is based on characteristics of the patients, physicians, and encounters. Such items include patient age, sex, and race; physician specialty, size of office, and ownership status; and a host of visit characteristics such as physician diagnosis, medications prescribed, counseling and therapy administered, diagnostic tests ordered or provided, and so forth. The NAMCS has been conducted periodically since 1974 and annually since 1989. The NAMCS uses the masterfiles from the American Medical Association and American Osteopathic Association to define the sampling frame of physicians in 112 primary sampling units (PSUs). Physicians are stratified by specialty and systematically sampled and assigned randomly to 52 one-week reporting periods throughout the survey year. A new sample of 3,000 physicians is taken each year. Sampled physicians are asked to fill out a one-page encounter form for a sample of their office visits in their reporting week. The physician is asked to complete the target 30 patient record forms (PRFs) himself or herself, but often someone in the physician's office completes the information retrospectively using the doctor's medical record of the patient. Data on about 25,000 patient encounters are obtained from about 70% of the in-scope physicians (Woodwell, 1999). The NAMCS does not gather any identifying information on the sampled patients, and thus no contact is made with the patient to verify any information collected in the survey.

Validation Studies Performed

In order to assess how well respondents understood the NAMCS questions, we conducted a reinterview of 32 NAMCS sample physicians from the 1994 panel (Stussman & Schappert, 1995). Census Bureau field representatives (FRs) were trained in how to conduct the reinterview study. They spoke individually with each person in the office who helped to complete the PRFs. Information was also collected on who completed the forms (e.g., physician, nurse), what the original source of the data was (e.g., medical record, doctor), and what the nature of the data source was (e.g., written text, checkbox). From this study we discovered several things that affected the validity of our survey items. First, contrary to what we believed, the physician was not completing the PRFs most of the time. In fact, physicians were completing only 28% of the forms, and Census Bureau FRs had to abstract the form from medical records 13% of the time. The remainder of the forms were completed by someone on the doctor's staff such as an office manager, nurse, or clerk. This finding had implications for the meaning of items where the information might be known by the doctor but was not recorded in the medical record. For example, the PRF had an item on whether the patient smoked cigarettes (yes, no, unknown). The NAMCS data estimated that one in four doctors did not know whether their patients smoked cigarettes (Schappert, 1996). When we discovered from the reinterview study that more than half of the forms were completed by nurses or other office staff, this estimate became more understandable. The doctor may have known whether the patient smoked cigarettes but did not record that information in the visit notes for

The author is at the National Center for Health Statistics.

the office personnel to transcribe; thus the "unknown" box was checked.

Information on the nature and source of the original encounter characteristics revealed that half of the data came from the medical record and of that, 25% were from checkboxes and 75% were from written text. About 22% of the data came from direct observation of the patient or the encounter itself, with the doctor providing 12% of the information directly.

Information on how well the people completing the forms understood the survey questions revealed that the meaning of most items was completely understandable; that is, the respondents knew what we wanted to include or have excluded in each response. However, there were some exceptions. Approximately one-quarter of the respondents did not understand that they were to include medications that the patients were told to *continue* using, in addition to new prescriptions. One out of three respondents did not realize that we wanted them to mark all expected sources of payment when the patient had multiple sources. And finally, 40% of respondents did not limit their estimate of the visit duration to time spent with the physician.

To correct some of these problems for the 1997 NAMCS panel, we changed the name of the "duration" item to "time spent with physician" and added a "providers seen" item so we could edit out times provided where no physician was seen by the patient. After these changes, the percentage of visits with no direct physician/patient contact increased from 1.6% to 3.2% between 1994 and 1997, but the mean duration time did not change. We changed the "expected source of payment" item to "primary expected source of payment" and added a separate item on whether the patient was an HMO member. We added more instructions on what to include in the medications item, and we removed the cigarette smoking item from the form.

Some validation studies of the NAMCS have not necessarily been conducted by NCHS. Stange et al. (1998) conducted an in-depth study comparing medical record content with direct physician/patient encounter observations. Nearly 5,000 patient encounters with 138 family physicians were observed by trained nurses who completed a data observation checklist for each encounter. The checklist data were compared with the data in the corresponding medical record for each encounter. The results indicated that the medical record was fairly sensitive to the accuracy of information about specific physician exams such as rectal or heart, screening services such as PAP tests, lab tests such as EKG or urinalysis, immunizations, and some reasons for visit (e.g., prenatal, acute, or well care visit). The medical record was not very sensitive to picking up health habit counseling, such as smoking cessation or diet advice. For those encounters that included various counseling services (as measured by the observation checklist), the medical record contained mention of that service from 57% to 8% (the highest was for alcohol history and the lowest was for diet advice about calcium). This would imply that the NAMCS, which relies heavily on medical record abstraction, would be grossly underestimating the provision of counseling services unless our sample doctors were more diligent in their record keeping.

Last year, we tried a different method for providing information on validity of some of the NAMCS survey items. We used geographic contextual variables from the Area Resource File (ARF) (HRSA, 1998) to determine if the NAMCS item responses were distributed as expected. This validation technique involves building hypotheses to test theories associated with the item content. For example, we could hypothesize that a doctor visit was more likely to be made by a person 65 years of age or older in geographic areas that had high proportions of people who were in that same age category. In 1997 we added several new items to the NAMCS PRF to measure the impact of managed care on patient/physician encounters. We were curious as to how well these item responses mirror contextual characteristics. For example, we could hypothesize that a visit was more likely to be made by a patient who was a member of an HMO in geographic areas that had high concentrations of HMO enrollees. Similarly, we also examined some long-standing PRF items that measure expected source of payment and patient's race.

Geographic Contextual Methods

Patient visit records from the 1997 NAMCS were merged with the variables derived from the 1997 Area Resource File (ARF) (HRSA, 1998) by modified state and county Federal Information Processing Standards (FIPS) codes. The FIPS county codes were established by the National Bureau of Standards to provide a consistent coding scheme to identify counties. The FIPS codes for the visit data were taken from the sample frame rather than the patient, which means that the FIPS codes were for the physician's location (preferred mailing address—either home or office) rather than the patient's residential address. The variables derived from the ARF included *percentage of county population 65 years and over, percentage of population below the poverty line, rate of HMO penetration among the population, percentage of population that is white,* and *percentage of population that is black.* NAMCS variables used in this analysis included several managed care items: *Was authorization required for care? Is the patient a member of an HMO? Was the patient referred by another physician or by a health plan for this visit? Are you the patient's primary care physician? Is this a capitated visit?* Other NAMCS variables included the patient's race, age, and expected source of payment.

The numerator of the HMO penetration rate on the ARF was taken from the *Interstudy Competitive Edge HMO Directory*, prepared by InterStudy in 1992–1996. All HMO membership is reported in the county where the HMO's corporate address is; therefore, it should not be necessarily assumed that all those members listed in the county actually visit doctors who work in the county (HRSA, 1998). However, we decided to use the data anyway, because while there were HMOs headquartered in states such as Pennsylvania whose members were in Maryland and Delaware, there were other HMOs headquartered in Maryland whose members lived in Pennsylvania and Delaware. We hoped that the discrepancies between areas and plans would even out.

The denominator for the ARF rates was taken from the 1996 population estimates produced by the Census Bureau, as was the numerator for the percentage of the population aged 65 and over. The percentage of the population below the poverty line, percent black, and percent white were taken from the 1990 Census of Population and Housing. Because data from the ARF are not all related to one year, the rates derived from the variables included in the ARF should be considered approximations for the variables of interest. Fine differences in rates should not be considered reliable. For this reason, the contextual variables were recoded into terciles using SAS ranking procedures. Gross differences in the contextual variables (low, medium, and high) should be considered reliable, however.

The unit of analysis was a patient visit, which was weighted by the sampling weight. The sampling weight includes components that reflect the probability of selection at the PSU, physician, and encounter levels, and has been adjusted for survey nonresponse. The analysis compared the tercile contextual variables with the corresponding variables from the NAMCS (e.g., HMO penetration rate with percentage of NAMCS visits by patients who are HMO members). SUDAAN was used to run chi-square tests and logistic regression estimates so that the complex nature of the sampling design could be taken into account in determining significance (Shah, Barnwell, & Bieler, 1996).

Results

Initial linear logistic regressions indicated that managed-care variables did not associate well with the HMO penetration rate in the county. However, significant regression coefficients were found for the expected-source-of-payment items and patient's age and race (data not shown). For ease of display, only results from the analysis using the tercile contextual variables and the NAMCS item responses are presented. Table 1 presents the results along with the chi-square statistic and corresponding p-value for each pairing. The mean rate for each tercile is provided the first time the contextual variable is mentioned in the table.

The managed-care items showed no relationship with HMO penetration rate. For example, one would expect relatively more visits by patients who are members of HMOs in counties that have high HMO penetration levels. The results indicated that in counties with a high penetration of HMO enrollees (mean = 49.2%), approximately 45% of the physician office visits were by patients who were members of HMOs, but in counties with a low HMO penetration (mean = 4.7%), 39.5% of the office visits were by patients who were HMO members (p = .603). Similarly, one would hypothesize that in counties with high HMO penetration there would be relatively more visits to primary care physicians as opposed to specialists. However, the results indicated that the observed difference was not statistically significant (49.2% in high-HMO counties vs. 40.7% in low-HMO counties).

The associations among the demographic variables and expected source of payment with the corresponding county levels were as expected. For example, one would expect the percentage of visits paid by Medicare to be associated with the relative numbers of persons in the county aged 65 years and older. The results indicated that 24% of visits had Medicare as an expected source of payment in counties with a high percentage (16.1%) of people 65 years and over, compared with only 18% in counties with a low percentage (9.4%) of people 65 years and over (p = .007). Similar associations were found for percentage of visits by white patients and black patients. Figure 1 graphically displays the linear association between the Medicaid responses on the NAMCS with the percentage of the county's population below the poverty line. This is what would be expected. However, as Figure 2 indicates, the association between HMO penetration level and one of the managed-care items is very weak.

Discussion

In discussing the results of the geographic contextual validation analysis, the age-old problem of the value of the criterion variables that are included in the analysis arises. The HMO penetration variable from the ARF has long had some problems because the count of enrollees is from the county where the managed-care corporate office is located rather than where the doctor practices. The doctors for covered persons may practice in other counties and even other states. While we thought that county counts of enrollees may have evened out, this doesn't appear to be the case. In addition, there were missing InterStudy data for approximately two-thirds of the counties for this variable. This means that the analysis was based on a greatly reduced number of records, increasing the corresponding standard errors for the NAMCS percentages. Nonetheless, one would have expected to see more of a relationship with the NAMCS managed-care variables.

Recent research has shown some nonresponse bias in the percent HMO estimate from the NAMCS. We conducted a nonresponse study by sending a one-page mail-back survey instrument to the physicians who refused to participate in the 1998 NAMCS (Burt, 1999). The results indicated that solo physicians were more likely to respond to the NAMCS than nonsolo physicians (72% vs. 62%) and that solo physicians were less likely to see HMO patients compared to nonsolo physicians (24% vs. 36%). Therefore, since there is no current weight adjustment on the NAMCS for nonresponding nonsolo physicians, the total estimates are slight underestimates in the percentage of visits by HMO patients. But even this nominal nonresponse bias would not bring about a lack of association with the contextual variable.

The next step in a validation strategy would be to again conduct a reinterview study of physicians' offices completing the latest versions of the PRFs and ask them whether they understood the items attempting to measure managed-care characteristics. Feedback from the FRs indicates that many offices did not understand what a "capitated"visit was. While the"unknown" categories were removed from the above analysis, they were fairly high: 10% for HMO status, 8% for authorization, 5% for primary care physician, and 14% for capitated visit.

Table 1. Association of selected NAMCS item responses with geographic contextual variables

Tercile[1]	ARF Variable and Mean within Tercile[2]	NAMCS Variable	Chi-square	*P*-value
	HMO penetration ratio per 100 persons	Percent capitated visit	7.12	0.037
Low	4.7	21.8		
Medium	19.4	9.6		
High	49.2	22.0		
		Percent patients belonging to HMO	1.02	0.603
Low		39.5		
Medium		38.1		
High		45.3		
		Percentage of visits to primary care physician	1.35	0.515
Low		40.7		
Medium		44.1		
High		49.2		
		Percentage of patient visits referred	0.79	0.678
Low		22.8		
Medium		19.2		
High		22.0		
		Percentage of visits requiring authorization	0.75	0.690
Low		15.4		
Medium		13.8		
High		17.2		
	Percentage of population 65 years and over	Percentage of visits paid by Medicare	11.08	0.007
Low	9.4	18.0		
Medium	12.4	19.3		
High	16.1	24.9		
		Percentage of visits by patients 65 years and over	9.15	0.016
Low		21.6		
Medium		23.0		
High		28.3		
	Percentage of population below poverty line	Percentage of visits paid by Medicaid	24.70	0.000
Low	4.7	4.3		
Medium	9.4	7.7		
High	15.3	12.6		
	Percentage of population white	Percentage of visits by white patients	87.78	0.000
Low	60.0	75.4		
Medium	80.6	87.3		
High	94.5	94.9		
	Percentage of population black	Percentage of visits by black patients	90.54	0.000
Low	1.7	3.0		
Medium	9.8	8.6		
High	29.4	19.4		

[1]Terciles based on the continuous geographic contextual variables.
[2]Mean ARF variables shown only the first time for the terciles.

Figure 1. Item validation for Medicaid from the NAMCS

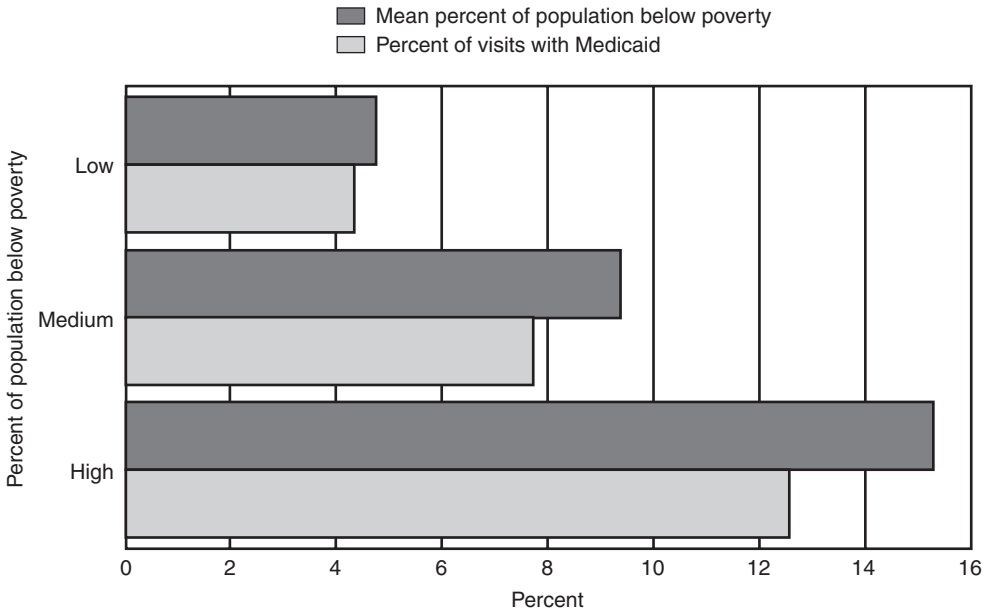
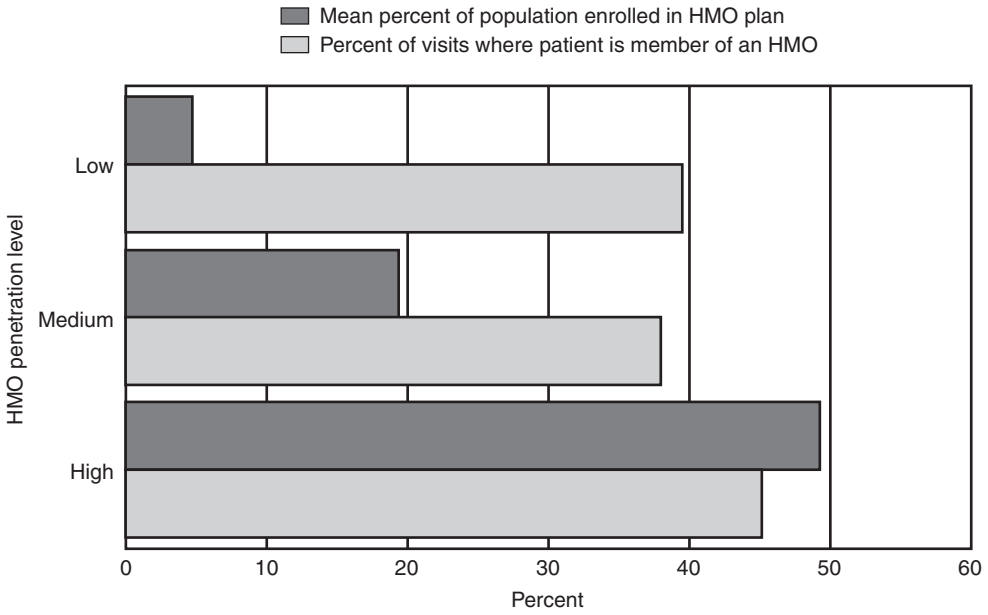


Figure 2. Item validation for HMO status from the NAMCS



Another step would be to expand the geographic contextual analysis to include data from sources other than the Area Resource File. For example, data from the National Health Interview Survey (NHIS) or the National Hospital Discharge Survey, which share many common PSUs with the NAMCS, could be used to provide other variables that could be categorized into areas high or low in a certain condition. These could be compared back to the NAMCS visit data to see if a similar pattern exists. The concern is to find an explanation when a relationship does *not* exist: Is it because there is lack of validity, or because there are confounding factors that affect health care utilization? These factors would have to be included in the analytical models.

This geographic contextual analysis study has several limitations in addition to the use of the InterStudy data to measure HMO penetration. The state and county of the patient would be a better indicator to match with geographical contextual data. While most patients go to doctors in their own county, there are some who will travel to another county to seek treatment. Combining NAMCS data with health service area data rather than state and county data should produce better results. Finally, the use of more sophisticated measurement tools would enhance the analysis.

Conclusion

This paper reviewed several techniques to measure the validity of records-based survey items: use of reinterviews to determine the extent to which survey instructions were followed and the meaning of survey questions understood; use of an independent measure of the physician/patient encounter through direct observation by a third party; and use of hypothesis-driven comparisons of survey estimates and geographic contextual variables. In addition to standard methods for assessing the validity of survey content, validity techniques that rely on comparisons with previously collected data may be the most cost-effective because there is no data collection effort to fund, but they could not be used exclusively for validity determination. Nonetheless, geographic contextual analysis is another tool that may be added to the survey methodologist's kit for better understanding survey results. Certainly, before such variables are used to analyze differences in health services utilization across counties, the expected relationships should exist.

References

Burt, C. (1999). Applying response propensity models to record-based health surveys. Submitted for publication.

HRSA. (1998). User documentation for the area resource file (ARF): February 1998 release. Health Resources and Services Administration.

Schappert, S. (1996). *National Ambulatory Medical Care Survey: 1994 summary.* Advance Data from Vital and Health Statistics, no. 273. Hyattsville, MD: National Center for Health Statistics.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1996). SUDAAN user's manual. Release 7.0. Research Triangle Park, NC: Research Triangle Institute.

Stange, K. C., Zyzanski, S. J., Fedirko Smith, T. et al. (1998). How valid are medical records and patient questionnaires for physician profiling and health services research? A comparison with direct observation of patient visits. *Medical Care, 36* (6), 851–867.

Stussman, B. & Schappert, S. (1995). National Ambulatory Medical Care Survey and National Hospital Ambulatory Medical Care Survey 1994 reinterview study. Unpublished paper, Division of Health Care Statistics, National Center for Health Statistics.

Woodwell, D. A. (1999). *National Ambulatory Medical Care Survey: 1997 summary.* Advance Data from Vital and Health Statistics, no. 305. Hyattsville, MD: National Center for Health Statistics.

Discussion Notes

Graham Kalton

The papers in this session employ a variety of ways to address the issue of evaluating the quality of survey responses. This issue is a critical one for survey research. At an aggregate level, there is serious concern when the estimates from one survey differ markedly from those of another survey or from those based on administrative records. Unfortunately, such differences are not uncommon. Examples from the papers presented here are the differences in the numbers of Medicaid enrollees estimated from the Current Population Survey (CPS) and from administrative records (Blumberg & Cynamon), and the differences in the number of uninsured estimated from the CPS and the Kaiser Survey of Family Health Experiences (K-SOFHE) (Harter, Kuby, & Moore). Other examples can be found in the areas of disability, poverty, crime victimizations, employment, and housing vacancies.

At an individual level, there is concern when the answers provided by a respondent are inaccurate. This concern exists even if the response errors of different respondents cancel out in aggregate, since such errors affect the relationships between variables. Thus the evaluation of survey responses should not be confined to aggregate analyses.

In his discussion, Sudman explores why response errors may occur in health classification items and makes some suggestions for reducing them. To complement Sudman's contribution, I have chosen to review the range of methods available for studying response errors, using the different methods employed in the papers as illustrations. My focus is on the strengths and limitations of the various methods. All the methods have significant limitations that need to be recognized. A welcome feature of all the papers is that they each clearly acknowledge these limitations.

A basic distinction between the methods that can be used to evaluate survey responses is that between macro-level and micro-level methods. Macro-level methods compare the estimates from a survey with estimates obtained from another source. The other source may be either another survey (as in Harter et al. and in Pascale) or administrative records (as in Blumberg & Cynamon). Micro-level methods compare individual survey responses with data obtained from another source. In this case, the other source may be records (as in Belli, Lepkowski, & Kabeto; Blumberg & Cynamon; and Fowler, Gallagher, & Homer) or reinterviews (Burt reports on reinterviews, but of a different type). For completeness, it may be noted that the other source could be another survey or census (as with a match of the CPS to the decennial census), but rarely will the samples of two surveys overlap sufficiently to make this approach feasible.

A further method of evaluation cuts across the macro/micro divide. This method involves checking on the consistency of the responses in relation to other information. At one extreme, it includes editing, in which a set of clearly inconsistent responses is modified to enforce consistency. In general, this method is a test of construct validity in which the validity of survey responses is assessed by examining how well the responses conform to theories relating them to other variables. Often the other variables used in tests of construct validity are responses to other questionnaire items, and the tests are conducted at the micro level. In Burt's application, the other variables are geographical contextual variables taken from an external source, and the tests are in a sense conducted at the macro level.

It is important to conduct macro-level evaluations since, as noted above, unexplained differences are problematic for data users. However, such comparisons have serious limitations when used to evaluate survey responses. The comparisons confound many different factors, including differences in definitions of the concept being measured (e.g., currently uninsured vs. uninsured in the past week), differences in population definitions, and differences in timing. When estimates from two surveys are being compared, the comparisons reflect noncoverage, nonresponse, processing errors, and sampling error as well as response error in both surveys. Sampling error can be taken into account in making the comparisons, but all the other sources are confounded with differences in survey responses. When an estimate from a survey is compared with that from administrative records, errors in the population of records (duplicate, dead, and missing records) and in the information on the records affect the comparisons. Estimates from record data are frequently mistakenly accepted as a gold standard, whereas they are often seriously flawed.

In making macro-level evaluations, it is sometimes possible to deal with some of the factors that may account for the differences observed as in, for example, the restriction by Harter et al. of the comparisons to the subpopulation common to the two surveys. It may also be possible to gain a fuller understanding of the differences between estimates by more detailed analyses, for example, examining whether the differences occur mainly for certain population subgroups. However, macro-level evaluations are usually unsatisfactory in that they are unable to provide convincing explanations for the differences.

A well-recognized macro-level method for examining the effects of certain components of the survey on survey responses is a randomized experiment. By experimental manipulation of

specified survey components with other components held constant, as was done in the Census Bureau's 1999 Questionnaire Design Experimental Research Survey described by Pascale, a better understanding of the effects of those components is obtained. Over many years, split-ballot and other randomized experiments have contributed greatly to our understanding of survey processes and thus to the development of survey methods. Nevertheless, such experiments have their limitations.

Generalization from an experimental setting may be uncertain because randomization is often obtained at the price of realism and representation (Kish, 1987). Thus, as Pascale notes, the generalizability of her findings is limited by the use of RDD methods, a low response rate, paper-and-pencil interviewing, and limited sample size. In addition, such experiments indicate only differences in the estimates obtained from the different experimental treatments; they cannot determine which treatment is to be preferred. Thus, reliance has to be placed on arguments such as "the more events reported, the better" (under the theory that events are underreported) or on other research methods (such as micro-level methods, cognitive research, and behavior coding) to choose the preferred treatment.

Burt's analyses of the associations between survey responses and geographical contextual-level variables face the general problems of macro-level analyses. Failure to observe the expected associations may occur for many different reasons, including measurement errors in the external data (e.g., HMO membership attributed to counties) and difference in time periods (the poverty estimates come from the 1990 Census). In addition, as is common to all tests of construct validity, the tests are as much tests of the underlying theory of the associations as of the validity of the survey responses. Furthermore, the theory generally predicts only the directions of associations but not their magnitudes. In this form construct validity tests are fairly weak.

Greater insight into the nature of response errors can be obtained from micro-level evaluation methods that compare individual survey responses with other data for the individual. Data from administrative records are often used for these evaluations. Three types of record check study can be distinguished: (a) reverse-record-check studies in which the survey data are collected from a sample selected from the records, thus covering only individuals in the record system; (b) forward-record-check studies in which the survey data are collected from a sample of the population, and record data are obtained for those reported to be in the record system; and (c) full-record-check studies in which survey data are collected from a sample of the population, and record data are collected for all sample members in the record system, whether or not they report in the survey that they are in the system. With a full-record-check study, the survey responses and the record data for an attribute such as being on Medicaid can be cross-tabulated as in Table 1.

Assuming for now that the record data are correct, B respondents incorrectly report the attribute and C respondents incorrectly report the absence of the attribute in the survey. The bias in the survey estimate is estimated by the difference between the survey estimate $(A + B)/N$ and the record estimate $(A + C)/N$, i.e., by the net difference rate $(B - C)/N$. To the extent that B and C cancel out, there may be little bias in the survey estimate, even if the gross error rate $(B + C)/N$ is large. In many cases, the proportion of the population with the attribute is small. In this situation, if the likelihood of misreporting is about the same for those with and those without the attribute, B will be larger than C, and the survey data will overestimate the prevalence of the attribute (see, for example, Blumberg & Cynamon's second study). It should be noted that bias in the survey estimate should not be the sole yardstick for evaluating the survey responses. A sizable gross difference rate can seriously affect the degree of association between the attribute and other characteristics even if the overreports (B) and underreports (C) cancel out.

Table 1. Comparison of survey responses and record data for the presence (yes) or absence (no) of some attribute

Survey Response	Record Data: Yes	Record Data: No	Total
Yes	A	B	$A+B$
No	C	D	$C+D$
Total	$A+C$	$B+D$	N

Reverse- and forward-record-check studies cannot provide the full data in Table 1. A reverse-record-check study gives only the first column of the table (i.e., A and C), and a forward-record-check study gives only the first row (i.e., A and B). Thus, neither study can produce an estimate of the bias in the survey estimate.

Record-check studies are a valuable tool for evaluating survey response but, as with other methods, they have their limitations. First, they are applicable only when the appropriate information is available on the records and the records are accessible. Second, differences in the definitions of the construct between the survey and the record system often have to be addressed (e.g., doctor visits in the Belli et al. study), as well as errors in the records. Third, there are usually significant problems with erroneous matches and failures to match that affect the analyses (see, for example, Blumberg & Cynamon).

The Fowler et al. study uses a record check but not for the purpose of evaluating survey responses. Fowler et al. aim to choose a set of items to identify children with special needs by seeing how various sets of items relate to MassHealth record data using enrollment in SSI or special programs as a rough surrogate for being in special need. With a positive response to any one item leading to a child being classified as having special needs, clearly the greater the number of items in the set, the fewer false negatives and the more false positives there will be. The limitation of the MassHealth record data is that they provide only a surrogate measure of special needs. In consequence, the findings from the study can be only suggestive.

Belli et al. use a record-check study to evaluate the effect of other aspects of the survey process on response errors.

They apply factor analysis to behavior codings to produce two factors that they label "cognitive difficulty" and "rapport." They then evaluate the effects of these factors on respondent accuracy in reporting the number of doctor visits, using (imperfect) HMO record data as the gold standard. This application of behavior coding is an interesting one, although caution is needed not to overinterpret the factors. Behavior coding has proved to be a useful technique for detecting questionnaire problems, and the extension to examine whether these problems affect survey responses is valuable. A useful related application would be to examine the relationships of individual behavior codes to response accuracy.

An alternative micro-level evaluation method obtains the comparison data not from records but from reinterviews. One form of reinterview study is simply to repeat the same interview using the same methods with a subsample of respondents to establish the reliability of their responses (assuming that the time interval between interviews is long enough so that respondents are not influenced by the first interview and short enough so that their survey characteristics have not changed). Another form is a validity study that employs improved methods with extended interviews for the reinterviews in an attempt to produce true values for the respondents. The reinterviews may be conducted by highly experienced, specially trained interviewers, who may employ many questions to obtain a full account of a respondent's characteristics, may employ detailed probing, may ask the respondent to consult personal records, and may well employ incentives to secure cooperation. This kind of approach has a long history (see, for example, Belson, 1963), but nowadays it appears to be out of fashion. It has advantages over a record-check study in that it can be applied to variables for which records are unavailable, it avoids problems of comparability and errors in record data, and it avoids matching and mismatching problems. Reinterview validity studies could usefully be more widely applied than is currently the case.

References

Belson, W. (1963). *Studies in readership*. London: Business Publications.

Kish, L. (1987). *Statistical design for research*. New York: Wiley.

Discussion Notes

Seymour Sudman

One of the most common purposes of a health survey is to determine whether or not an individual is a member of a specified class. This purpose might well be described by a song title from the forties: "Is you is or is you ain't my baby?" For example, the first four papers in this session all discuss the validity of reports that tell whether an individual is or is not covered by health insurance (specifically, Medicaid) or whether a child has or does not have special needs. Other very common examples include whether an individual does or does not have a specific disease, and more generally, whether an individual is of a specific race or has a telephone.

Compared to questions that ask about frequencies of events, or attitudes, classifying questions would appear to be relatively nonproblematic for respondents, but as seen from the papers in this session and from much earlier work on health and insurance variables, this is far from the case. When comparisons are made to validation information, one observes both Type 1 and Type 2 errors; that is, some people are erroneously omitted from classes to which they belong, while others are erroneously included into classes to which they do not belong.

Blumberg and Cynamon link survey responses on Medicaid coverage to enrollment lists in Minnesota and Texas. They found about 20% underreporting of Medicaid coverage in Minnesota and 15% underreporting in Texas. On the other hand, they found an overreporting rate of 8.5% in Texas. The other papers do not have comparisons to record data, but the substantial differences seen when different methods are used clearly indicate a lack of reliability.

I would like to draw upon the papers from this session, and from earlier work, to try to generalize as to why this is happening and to offer some suggestions for reducing classification errors. For the purposes of this discussion, I assume that these classification errors are real. We are all aware that the records used to validate survey responses are themselves subject to very significant errors that must be considered in any evaluation, but I will ignore these errors and concentrate on reasons for survey errors. Also for this discussion, I will assume that the questions being asked are not especially threatening. Questions that ask about drug use or other illegal behaviors create special problems of self-presentation that are not really addressed in the papers given here.

Let's start with some characteristics of the class that are likely to increase response errors. An obvious one is that membership in the class is not constant but changes over time. The three papers dealing with health insurance all make the point that conceptually, depending on how the question is asked, interval time periods yield greater estimates of those insured or uninsured at any point in the period as compared to point estimates. Cognitively, the task for the respondent is much more difficult for time interval estimates if there have been one or more changes in classification during the interval, and this difficulty increases with the length of the time period and the frequency with which the classification changes. For time interval estimates, the respondent must remember a series of dates in addition to information on classification. The reduced quality of reported data usually overbalances the increase in information obtained.

This issue is well discussed in the Pascale paper, which points out the very large differences between the Current Population Survey, which asks about a full calendar year, and the Survey of Income and Program Participation and the National Medical Care Expenditure Survey, which ask about the previous three or four months but can be aggregated to annual totals. The experimental data presented in the paper are complex. They seem to indicate an interaction between length of time period and method of asking about all household members. From my perspective, these results appear to confirm the cognitive complexity in asking about time interval classification.

A second characteristic of classes that increases cognitive complexity is when the classes are partial, ill defined, and not mutually exclusive. The obvious example is racial or ethnic background. A respondent may be one-sixteenth American Indian, one-eighth African-American, one-eighth Asian, and eleven-sixteenths Caucasian. Even with the new OMB regulations that permit multiple categories, such respondents will still have problems classifying themselves racially. In the health domain, complexity arises when a condition has been diagnosed and treated. If a child was diagnosed with attention deficiency, is being treated with medication, and is having no problems, does this child have special needs? If I take blood pressure medication and my blood pressure is normal, do I have high blood pressure? Such questions are often worded, either consciously or unconsciously, to increase or decrease estimates of occurrence, but they are often difficult for respondents to understand and remember.

The paper by Fowler, Gallagher, and Homer illustrates the problems that parents face when asked about the special medical needs of their children. Fowler and his colleagues started with a sample of children enrolled in SSI or other special programs. All of these children were identified as having special

The author is at the Survey Research Laboratory, University of Illinois.

needs. About two-thirds of these SSI children were identified as having a chronic condition using three basic questions. This proportion rose when additional questions were asked, but it is still clear that many parents and caregivers have difficulty with this concept. The more general and fuzzy the concept, the greater the likelihood that respondents will have cognitive difficulties. We should also note that this paper looked only at false negatives and did not study the problem of false positives.

I do not want to suggest that these class characteristic difficulties cannot be reduced by the researcher, and I'll give some obvious recommendations later, but there are some parts of a health survey that are even more directly under the control of the researcher. One critical requirement for a respondent is knowledge. Even in discussing their own insurance coverage, respondents may be unaware or confused about some of the details such as type of insurance and what is and is not covered. A fortiori, proxy reporters, even if they know of the existence of some kind of health insurance, will not likely know all the details of the insurance coverage of other household members. The greater the detail required and the greater the social distance between the proxy and the household member reported on, the greater the likelihood of misclassification.

Essentially, the same statements may be made about retrieving information from memory. We remember best information about ourselves as compared with proxies, although a child's caregiver will remember nearly as well about the child as about him- or herself. Memory quality declines as the social distance between the proxy and other household member increases. Memory quality also declines as the information required becomes more detailed.

What can survey researchers do to reduce the cognitive difficulty of health classification questions? First, they can make a judicious choice of the time period to cover. The evidence from the papers given in this session clearly suggests that asking about current classification such as current insurance coverage is cognitively easier and yields better data than asking about classification in some time interval.

For some disease classifications, it may be appropriate to ask an "ever" question: "Have you ever been told that you have . . . cancer, high blood pressure, diabetes, etc.?" Asking about a specific time period, be it last year, last quarter, or whatever, makes the task much tougher.

If maximum accuracy is desired, self- rather than proxy reports should be used, except for children, for whom the information should be obtained from caregivers. Proxy reporting is often used as a compromise to save resources when the individual household members are difficult to locate or do not wish to cooperate. It is clear that proxy reports about spouses and minor children are better than proxy reports about other adult household members, related or not. A reasonable compromise is to allow proxy reporting about spouses and minor children but to require self-reports from other household members.

Finally, it is clear from all the insurance papers, but especially the Blumberg-Cynamon one, that respondents are much more accurate in reporting gross categories, such as any insurance coverage, than they are in reporting specific detailed types of insurance. Of the 20% of parents in Texas who did not report Medicaid coverage, 80% of these reported some form of health insurance coverage. This suggests that the methodology in the NHIS that asks first about any insurance coverage and then asks about specific types of insurance is better than the methodology that asks first about specific types and may later ask about any kind of insurance. Here the desire to get details detracts from the validity of the overall estimate.

In my discussion to this point, I have considered only issues of question wording. It is also important to remember that sampling issues have an impact on the validity of classification information. One obvious example is the use of telephone methods for measuring the uninsured. Although only 5–6% of households are without phones, the majority of such households are also uninsured. Thus, phone surveys significantly underestimate the uninsured. The Harter, Kuby, and Moore paper makes the important point that how "family" is defined has a significant effect on estimates of uninsured families. While policy needs may require an estimate of uninsured families, methodological concerns suggest that estimates of individual insurance coverage are less sensitive to these definitional problems.

To sum up, the excellent papers in this session remind us—if we need reminding—that even so-called simple health classification questions can cause serious cognitive problems for respondents. Classifications will be most accurate when they do not change and only a single classification fits. The best example that comes to mind is gender, although one can point out some exceptions even with this. As an aside, date of birth is superior to age for the same reason.

Survey researchers can reduce the cognitive burden on respondents by asking about current classification rather than classification during some time interval. The respondent's task is easier for gross rather than detailed classification—any insurance coverage is easier than coverage by a specific type. Self-reports are superior to proxy reports for classification, and proxy reports about spouses and minor children are superior to proxy reports about other household members. The aim of obtaining perfect information is still far beyond our reach. Our aim should be to use methods that minimize error, to measure this error, and to report it to policymakers and other data users.

Discussion Notes, Session 4

Timothy Johnson and Willard Rodgers

Several themes dominated the discussion of this session, including sources of invalidity, methods for assessing validity, and potential ways to improve on existing methods.

Sources of Invalidity

One issue that was discussed concerned questionnaire context effects. It is important that we pay more attention to what precedes the questions of interest in the survey schedule. For example, if we change the frame of reference from past-year experiences to current health insurance status, respondents may be conditioned to answer using the previously defined frame of reference. Another issue considered was that of proxy versus self-reports. Self-reports are generally thought to be more valid, but this may not always be the case. For example, children and the elderly may not always provide the most accurate information about themselves. Also, the person most knowledgeable about the insurance status of an individual may be the person in the household who pays the bills rather than the target individual. The effects of behavior frequency and salience as well as interview mode were other sources of invalidity that were discussed.

Assessing Validity

Several of the papers presented in this session employed the assumption that increased reporting of the condition or behavior of interest represented increased accuracy. It was observed that "the more questions you ask . . . the more answers you get" and that you could thus produce higher reporting rates by increasing the number of questions concerned with the topic. This approach was challenged, and there was general agreement from the floor that it was just a device that should be complemented with other validation checks. Another approach is the use of administrative records as a gold standard for report validation. Yet this approach was also challenged, as several of the papers illustrated problems with the completeness and/or accuracy of such records. Clearly, appropriate gold standards are often elusive.

Potential Solutions

If no one source is perfect (that is, there is no pure gold standard), the suggestion was made to follow a strategy of evaluating validity using multiple indicators of the same phenomenon. The value of a general question that comes first or a "mop-up" question at the end (to get information that is missed in sets of detailed survey questions) was also discussed. Finally, the suggestion was made that more accurate information could be obtained if we tailored individual questions to respondents. For example, this technique could be used to trigger different types of memory retrieval mechanisms for individuals in different circumstances. Finally, it was noted that more research is needed regarding all three of these potential approaches.

Needs for State and Local Data of National Relevance

The devolution of responsibility to the states for the implementation of health and health-related programs has greatly increased the demand for state and local area data. Welfare reform and implementation of the State Children's Health Insurance Program are two examples of relatively new federally funded state programs that require strong surveys to evaluate program impact and effectiveness, and several other ongoing survey initiatives have also faced increased requirements to better address the localized nature of health care delivery and health policy and corresponding data needs. Meeting such demands has been an expensive endeavor, and localized efforts to mount important surveys have often been of uncertain quality or have varied so much in form, content, and methodology that comparisons across states (or even within states) has been problematic.

To address these escalating needs and the problems of meeting them, several federal agencies and private foundations have either established major new survey initiatives or honed and enhanced existing data systems to provide better data at the state level. Several federally sponsored data systems now provide state-level data on an annual basis. Some, like the Behavioral Risk Factors (BRFSS) and Youth Risk Behavior (YRBS) Surveillance Systems and the National Immunization Survey (NIS), have been in existence for many years. Others, like the National Household Survey on Drug Abuse (NHSDA), have been newly expanded and transformed from national to state-level surveys. Still others are completely new, such as the State and Local Area Integrated Telephone Survey (SLAITS), a spin-off from the NIS. Each of these is represented in the feature papers presented at this session.

Pooling State Telephone Survey Health Data for National Estimates: The CDC Behavioral Risk Factor Surveillance System, 1995

Ronaldo Iachan, Jane Schulman, Eve Powell-Griner, David E. Nelson, Peter Mariolis, and Carol Stanwyck

Introduction

As a state-based surveillance system, the Behavioral Risk Factor Surveillance System (BRFSS) was not designed to generate national estimates of health risks and health practices. Nevertheless, there is much interest among the research community in using the BRFSS for such estimates because all states use the same core instrument, sample size is relatively large, and the annual data are available within 6 months after collection. The BRFSS national estimates are more traditionally and widely used by state health policymakers as a benchmark against which they can compare the health practices of their citizens and for tracking progress on *Healthy People 2000* Objectives (Public Health Service, 1991).

Three main objectives are addressed in this study: (1) investigating the appropriateness of combining data across states in terms of variations in sample design as well as sampling and nonsampling errors; (2) comparing the usual method for computing BRFSS national estimates to a newer method that may take sample design more explicitly into account; and (3) assessing the correspondence between national estimates for a select set of health measures from the BRFSS and those from the National Health Interview Survey (NHIS).

This study is part of a project conducted by Battelle with CDC funding. The larger project used data from 1995, 1996, and 1997. In this presentation, however, we focus on 1995 because that was the most recent year for which NHIS data were available at the time of the Battelle analysis.

Background

The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project of the Centers for Disease Control and Prevention (CDC) and U.S. states and territories. The BRFSS, administered and supported by the Behavioral Surveillance Branch (BSB) of the CDC, is an ongoing data collection program designed to measure behavioral risk factors in the population 18 years of age or over living in households. The objective of the BRFSS is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases in the adult population. Factors assessed by the BRFSS include tobacco use, physical activity, dietary practices, safety-belt use, and use of cancer screening services, among others. Self-reported data are collected from a random sample of adults (one per household) through a telephone survey. Although 95% of U.S. households have telephones, coverage ranges from 87–98% across states and varies for subgroups as well (U.S. Bureau of the Census, 1994). No direct method of compensating for non-telephone coverage is employed by the BRFSS; however, post-stratification weights are used and may partially correct for any bias caused by non-telephone coverage. These weights adjust for differences in nonresponse and noncoverage, and must be used for deriving representative population-based estimates of risk behavior prevalences.

Field operations for the BRFSS are managed by the health departments under guidelines provided by the BSB via various mechanisms, including survey manuals, numbered memoranda, and other forms of training. The health departments collect and process the data either in-house or through use of contractors. The data are transmitted to the National Center for Chronic Disease Prevention and Health Promotion's Behavioral Surveillance Branch at the CDC for further editing, processing, weighting, and dissemination.

The questionnaire has three parts: the core component, optional modules, and state-added questions. Only the core component is considered in the current research, so discussion will be restricted to it. The core is a standard set of questions asked by all states. It includes queries about current health-related perceptions, conditions, and behaviors (e.g., health status, health insurance, diabetes, tobacco use, selected cancer screening procedures, and HIV/AIDS risks) and questions on demographic characteristics. Many questions in the core are taken from established national surveys, such as the National Health Interview Survey or the National Health and Nutrition Examination Survey. This practice allows the BRFSS to take advantage of questions that may have been tested and allows states to compare their data with those from other surveys. Any new questions proposed as additions to the core must go through cognitive testing prior to their inclusion on the survey. BRFSS protocol specifies

Ronaldo Iachan and Jane Schulman are at the Battelle Centers for Public Health Research and Evaluation, Baltimore.

Eve Powell-Griner, David E. Nelson, Peter Mariolis, and Carol Stanwyck are at the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Behavioral Surveillance Branch.

that all states ask the core component questions without modification. The CDC provides Ci3 CATI programming for the questionnaire. The questionnaire is available in both English and Spanish.

Interviews are conducted through computer-assisted telephone interviewing (CATI) by almost all survey areas, and interviewing specifications provided by the CDC are used by the state health personnel or contractors conducting interviews. The core portion of the questionnaire lasts an average of 10 minutes. The CDC provides materials for use in training interviewers, supervisors, and coordinators, as well as handbooks for the Ci3-CATI system used by BRFSS, BRFSS protocol specifications, and other materials related to data collection and management. States and contractors monitor interviewers. All states are required to do verification callbacks for a 5% sample of completed interviews as part of their quality control practices.

Telephone interviewing is conducted during a two-week period each month, and calls are made seven days per week, during both day and evening hours. Standard procedures in interviewing are followed for rotation of calls over days of the week and times of the day, and detailed specifications are provided for dispositioning calls.

After data are collected each month, the state transmits it to the CDC for editing, processing, and weighting. Following completion of the data collection processing, documentation materials assessing data quality issues as well as content are prepared and disseminated.

Combining Data Across States

We consider several issues related to the appropriateness of combining BRFSS data across states, including state sampling designs, sampling errors (precision), and nonsampling errors (bias). We first identify the type of design and sample size for each state. We then examine several indicators of data quality: the coefficient of variation of sample weights, the design effect averaged across 20 leading health risk indicators, CASRO (Council of American Survey Research Organizations) response rates, and the upper-bound response rate. If states use substantially different sampling designs—that is, if they are not using RDD designs or not covering the population specified by BRFSS protocol–or if the quality of data varies so substantially across states that state comparisons of estimates are invalid, it may be less appropriate to combine data across states for national estimates.

We look first at the effect of state survey sample designs. The 50 states participating in the BRFSS in 1995 can be classified according to use of two basic sample designs: Waksberg-type and list-assisted designs (Waksberg, 1978). Thirty-four states used a Waksberg-type sampling method, and 16 used some form of stratified list-assisted sample design (Table 1).

Waksberg-Type Designs

These telephone sample designs are variations of a three-stage cluster design. First, telephone numbers are grouped into primary sampling units (PSUs); each PSU is a cluster of 100 phone numbers, also designated as a 100-block. Each set of 100 phone numbers has the same area code, prefix, and first two digits of the suffix. Clusters are sampled randomly, and within each selected cluster, a single phone number is randomly selected to be dialed. If the selected phone number is a household, the entire PSU is selected for further sampling. In the third sampling stage, an adult 18 years of age or older is randomly selected from eligible households. Telephone numbers from accepted PSUs are dialed until three completed interviews are obtained.

List-Assisted Design

Telephone numbers are grouped into 100-blocks with the same area code, prefix, and first two digits of the suffix. The 100-blocks with one or more listed household numbers are put into a high-density stratum, which is expected to contain a large proportion of households. The 100-blocks with no listed household numbers are put into a low-density stratum, which is expected to contain a small proportion of households. Both strata are sampled to obtain a probability sample of all households with telephones, but the high-density stratum is sampled at a higher rate than the low-density stratum. Some states have modified this basic design so that they (1) consider more than two density strata; (2) consider different thresholds for classifying telephone numbers into high- and low-density strata (e.g., require at least two or three listed household numbers to classify a 100-block into the high-density stratum); or (3) truncate the frame rather than undersample low-density blocks. For example, Wisconsin assigns telephone numbers to high- and low-density strata based on previous experience with their prefixes. In addition, they identify three density strata instead of two. Michigan defines its high-density stratum as 2+ blocks and its low-density stratum as 1– blocks. Nebraska generates a simple random sample of telephone numbers from a sampling frame of all working telephone numbers in the state.

Irrespective of the design used, the BRFSS standard for participating areas is that sample records must be justifiable as a probability sample of all households with telephones in the state. Generally, for a given data year, almost all states meet this criterion. In 1995 Alaska, California, Hawaii, Nevada, and Texas did not adhere to the standard BRFSS design. Hawaii and Texas used only 1+ blocks, and California used 3+ blocks. Alaska excluded numbers with a low probability of being household numbers in two of their strata. Alaska's sample design excludes an estimated 9% of all residential numbers. The sample design used by California and Texas led to the exclusion of about 2–3% of all household numbers. Nevada restricted the sampling frame to 100- blocks that contained five or more listed household numbers.

Large sampling errors imply imprecise survey estimates. State sample designs that lead to variability in survey weights also lead to large standard errors. State-level sampling errors were examined using the coefficient of variation (CV) of the survey weights and the design effect (DEFF) averaged over 20 key health risk factors (Table 1). The coefficient of variation

Table 1. State sample designs and outcome measures, 1995 BRFSS

State	Sampling Design	Sample Size	CV	DEFF	UB Rate	CASRO Rate
Alabama	Waksberg cluster	1,792	0.543	1.3	79.9	68.30
Alaska	List-assisted	1,535	1.114	2.2	80.0	68.40
Arizona	Waksberg cluster	1,913	1.016	2.0	73.6	65.10
Arkansas	Waksberg cluster	1,800	0.450	1.3	76.0	65.70
California	List-assisted	4,046	1.149	3.1	70.3	52.30
Colorado	List-assisted	2,461	0.852	2.0	86.1	77.40
Connecticut	List-assisted	1,869	0.601	1.5	78.9	65.10
Delaware	Waksberg cluster	2,112	0.515	1.3	89.8	68.30
Florida	Waksberg cluster	3,335	0.524	1.4	77.1	54.60
Georgia	Waksberg cluster	2,388	0.536	1.3	83.4	77.60
Hawaii	Waksberg cluster	2,157	0.709	1.5	82.2	48.60
Idaho	Waksberg cluster	2,743	0.463	1.2	76.0	66.80
Illinois	Waksberg cluster	2,889	0.532	1.4	73.2	61.60
Indiana	Waksberg cluster	2,412	0.444	1.2	86.5	78.90
Iowa	Waksberg cluster	3,600	0.459	1.2	86.9	73.30
Kansas	List-assisted	2,009	0.404	1.3	89.9	73.60
Kentucky	Waksberg cluster	2,388	0.516	1.3	87.1	72.60
Louisiana	Waksberg cluster	1,657	0.503	1.3	77.4	67.30
Maine	Waksberg cluster	1,279	0.465	1.2	83.8	70.20
Maryland	Waksberg cluster	5,107	0.506	1.4	75.7	60.90
Massachusetts	List-assisted	1,768	0.520	1.5	69.1	60.40
Michigan	List-assisted	2,475	0.434	1.4	80.1	56.00
Minnesota	Waksberg cluster	3,943	0.446	1.2	92.9	78.10
Mississippi	Waksberg cluster	1,592	0.535	1.4	85.3	75.40
Missouri	Waksberg cluster	1,572	0.514	1.3	68.6	59.10
Montana	Waksberg cluster	1,193	0.446	1.2	88.7	77.50
Nebraska	List-assisted	1,819	0.444	1.3	80.4	67.50
Nevada	List-assisted	1,802	0.588	1.7	90.4	77.40
New Hampshire	Waksberg cluster	1,502	0.525	1.3	78.7	59.50
New Jersey	List-assisted	1,251	0.619	1.7	86.7	66.90
New Mexico	Waksberg cluster	1,298	0.583	1.5	60.5	52.50
New York	Waksberg cluster	2,477	0.589	1.4	72.4	60.20
North Carolina	Waksberg cluster	3,340	0.538	1.4	86.8	69.20
North Dakota	Waksberg cluster	1,860	0.441	1.2	95.0	84.50
Ohio	Waksberg cluster	1,346	0.596	1.4	81.4	69.50
Oklahoma	Waksberg cluster	1,779	0.589	1.5	79.2	76.20
Oregon	Waksberg cluster	2,845	0.443	1.3	67.9	56.90
Pennsylvania	List-assisted	3,601	0.765	1.8	69.0	64.10
Rhode Island	List-assisted	1,776	0.538	1.5	76.9	68.70
South Carolina	Waksberg cluster	2,038	0.585	1.4	83.9	74.50
South Dakota	Waksberg cluster	1,810	0.422	1.3	89.0	81.20
Tennessee	Waksberg cluster	2,040	0.484	1.3	79.7	68.70
Texas	List-assisted	1,703	0.628	1.3	75.6	60.20
Utah	Waksberg custer	2,891	0.851	1.7	87.4	78.50
Vermont	List-assisted	2,490	0.606	1.3	89.3	74.50
Virginia	Waksberg cluster	1,799	0.503	1.4	75.9	62.30
Washington	List-assisted	3,351	0.469	1.5	72.0	61.40
West Virginia	Waksberg cluster	2,434	0.508	1.3	85.0	77.50
Wisconsin	List-assisted	2,210	0.805	1.8	74.5	71.90
Wyoming	Waksberg cluster	2,437	0.450	1.2	79.5	69.20

CV refers to the coefficient of variation of the survey weights. DEFF refers to the design effect averaged over 20 key health risk factors. UB rate is the upper-bound rate and includes only refusals (02), terminations (09), and completed interviews (01). The resulting estimate reflects the cooperation of eligibles contacted and is not affected by differences in telephone sampling efficiency.

CASRO rate is the response rate developed by the Council of American Survey Research Organizations (CASRO) that apportions dispositions with unknown eligibility status (ring-no-answer and busy) to dispositions representing eligible respondents in the same proportion as exists among all calls of known status (all other BRFSS call dispositions). The resulting estimate reflects telephone sampling efficiency as well as the degree of cooperation among eligibles contacted.

was calculated as the standard deviation of the weights divided by their mean. We noted that the CV of the weights is largest for California (1.149) and Alaska (1.14). The average DEFF was computed over a subset of 20 key health risk items for 1995 and ranged from 1.2 to 3.1, with California (3.1), Alaska (2.2), Arizona (2.0), and Colorado (2.0) having the highest average DEFF. These large CV and DEFF values are related to the use of a sample design with large differences in the probability of selection of numbers. Although not true for California, the primary characteristic of designs with large differences in probabilities of selection of numbers is that they sample geographically defined strata at different rates. This results in an inefficient design, but the data are not *necessarily* of low quality.

Nonsampling errors imply bias in survey estimates. We examined two indicators of nonsampling errors: the CASRO and upper-bound response rates. The response rates measure the extent to which interviews were completed from among the telephone numbers selected for the sample. The higher the response rate, the lower the potential for bias in the data. The CASRO rate apportions dispositions with unknown eligibility status to dispositions representing eligible respondents in the same proportion as exists among all calls of known status. The resulting estimate reflects telephone sampling efficiency as well as the degree of cooperation among eligibles contacted. The CASRO rate ranged from 48.6 to 84.5, with a median of 68.4. CASRO rates were especially low for California (52.3), Florida (54.6), Hawaii (48.6), and New Mexico (52.5). The upper-bound rate calculation includes only refusals, terminations, and completed interviews. The resulting estimate reflects the cooperation of eligibles contacted and is not affected by differences in telephone sampling efficiency. The state surveys had an upper-bound response rate ranging from 60.5 to 95.0. New Mexico (60.5), Oregon (67.9), Pennsylvania (69.0), Massachusetts (69.1), and California (70.3) had lower upper-bound rates in 1995 compared with other areas (Table 1). These variations suggest that data collection procedures may be less than optimal or that there may be differences in the cooperativeness of respondents across states. Differences in outcome measures should not be overly emphasized because they do have limitations as measures of data quality. The CASRO is affected by the overall distribution of resolved numbers, which differs significantly between list-assisted and Waksberg designs, as well as between state telephone systems. For example, rates calculable from BRFSS data for Waksberg states include only the stage 2 numbers, that is, the numbers from accepted clusters that were called after the initial screening. Those numbers therefore include the calls to the number in each cluster that was specifically identified as a household and used to accept the cluster. The upper-bound rate does not incorporate nonresponse from households that do not explicitly refuse. Thus, these outcome measures are indicators of quality but cannot, in and of themselves, provide definitive answers about data quality and potential bias in the data. In this presentation we have included noncoverage due to households without telephones, although it is a source of nonsampling error.

Comparison of Pooled and Stratified National Estimates for Key Health Risk Factors

The second research objective was to compare the national estimates obtained using two different methods: pooled and restratified. We refer to the method that uses the weights assigned on the BRFSS data file[1] as the pooled method.

The restratified method refers to the method whereby the states are separated into two classes according to design used, an additional poststratification factor is applied to adjust individual weights to sum to the population totals for the two groups of states, and national estimates are derived as weighted sums of the estimates for the two groups of states. Specifically, design-group-level estimates are first computed for those using Waksberg-type designs and then computed for those using list-assisted designs. Each group-level estimate is a weighted sum of the individual state estimates, where the strata weights are proportional to the state's population. That is, the ith respondent in state h is assigned a weight $W(hi)$ so that state-level estimates, $y(h)$, are approximately unbiased. To combine the state-level estimates, the stratum weight $T(h) = N(h)/N$ is applied to the estimate $y(h)$, or equivalently to the individual weight $W(hi)$, where $N(h)$ is the state population and N is the total population at the design-group level. In generating the estimates for each of the two groups, a two-step post-stratification procedure to adjust the weights before and after combining the state estimates was used to mitigate the potential impact of state level noncoverage and nonresponse rates. We first used the final state-specific post-stratified weights available from the BRFSS. The second design- group level adjustments forced the individual weights to sum to the population totals for the two groups of states, using post-(re)stratification cells

[1]The data are weighted according to the following general formula. Where a factor does not apply, its value is set to 1. FINALWT = GEOWT * DENWT * 1 OVER NPH * NAD * CSA * POSTSTRAT FINALWT is the final weight assigned to each record. GEOWT is the inverse of the ratio of the estimated sampling fraction of each area code–prefix combination subset to the area code–prefix combination subset with the largest estimated sampling fraction. It weights for the unequal probability of selection by area code–prefix combinations intended to cover specified geographic regions. Almost always, the regions are discrete subsets of counties and the boundaries of the area code–prefix combinations do not correspond exactly to the boundaries of the specified geographic regions. DENWT is the inverse of the ratio of the sampling fraction of each subset of 100-blocks (sets of telephone numbers with identical first eight digits and all possible final two digits) sampled at a given rate to the 100-block subset with the largest sampling fraction. It weights for the unequal probability of selection by presumed household density of hundred block. This is generally used in a design in which telephone numbers from 100-blocks with one or more listed residential numbers (one-plus blocks) are sampled at a higher rate than telephone numbers from hundred blocks with no listed residential numbers (zero blocks). 1/NPH is the inverse of the number of residential telephone numbers in the respondent's household. NAD is the number of adults in the respondent's household. CSA is the ratio of the expected cluster size to the actual cluster size. POSTSTRAT is the number of people in an age-by-sex or age-by-race-by-sex category in the population of a region or a state divided by the sum of the preceding weights for the respondents in that same age-by-sex or age-by-race-by-sex category. It adjusts for noncoverage and nonresponse and forces the sum of the weighted frequencies to equal population estimates for the region or state.

Table 2. Comparison of pooled and stratified national estimates (1995 BRFSS)

Key Health Risk Factor	Estimate (Standard Error)		Absolute Difference	
	Pooled	Stratified	In Estimate	In Standard Error
Percent reporting ever having been told by a doctor that he/she had diabetes	5.54% (0.13)	5.80% (0.17)	0.26%	0.04
Current smoking prevalence	22.25% (0.21)	22.15% (0.22)	–0.10%	0.01
Percent reporting any kind of health coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare	86.92% (0.20)	86.36% (0.22)	–0.56%	0.02

defined by gender, race, and age. Finally, we computed national estimates as weighted sums of the estimates for the two groups of states, where the weights reflect the proportion of the national population in each design group of states. The final step draws on the stratified estimation methodology by considering the two groups of states as the new primary strata. This procedure allows for more explicit consideration of any differences in Waksberg-type versus list-assisted designs.

We chose a set of 20 key health risk factors from the BRFSS for analysis based on (1) the CDC's needs, (2) the consistency of the items across states, and (3) the consistency of the BRFSS items with the NHIS. We included questions in the following areas: medical history, tobacco use, HIV/AIDS, alcohol use, and health care coverage. The procedure was implemented with SUDAAN.

Table 2 illustrates the results of this analysis using 1995 survey data for a subset of 3 of the 20 key health risk factors we examined: dianosis of diabetes, current smoking prevalence, and percent reporting any kind of health coverage. The absolute difference in the point estimates is minimal. For example, 5.8% of respondents reported ever being told by a doctor that they had diabetes using the stratified estimate. The corresponding figure obtained using the original BRFSS weights was 5.5%. In fact, the absolute differences for the point estimates were less than 1% for all 20 key health risk factors we examined.

The differences between the point estimates using either of the two estimation methods are quite small. There are several factors that might account for the differences observed for the two methods. First, the population data for the pooled estimates came from 1995 Demo-Detail (1995) files, while the restratified population data were obtained from the Census (U.S. Bureau of the Census, 1998), so that the populations used were not identical. Second, the post-stratification process for the two methods varied in that the restratified method added another level beyond that used in the pooled method. The additional post-stratification factor was used to adjust individual weights to sum to the population totals for the *two design groups of states*, and national estimates then derived as weighted sums of the estimates for the *two design groups of states*. In contrast, the pooled method adjusted individual weights to sum to the population totals for the individual states, and national estimates were simply the sum of those states.

The major differences between the two methods are the relatively larger standard errors in the restratified method. These differences are coming primarily from the list-assisted states. Different SUDAAN statements were used in the restratified method compared with the pooled method. For example the restratified design statement specifies WOR while the pooled method specifies WR. Further, the restratified method uses a value of 100 for the PSU variable in Waksberg states, whereas the pooled method PSU variable contains a unique number for the completes from each PSU. The result of this latter procedure is that every record in a Waksberg state has the same PSU value (100) under the restratified method, but in the pooled method only three records share the same PSU value, with the range of values varying by state. There are also other differences in the specification of the SUDAAN nest statement that may be contributing to the differences in the standard errors.

Comparison of NHIS and BRFSS Stratified National Estimates

Our last research objective was to compare the national estimates obtained from the BRFSS with those obtained from the NHIS. Briefly, the NHIS is a nationwide survey of households that collects data on acute conditions, injuries, chronic conditions, health status, and medical care utilization. The Census Bureau conducts in-person interviews, and the data are edited and analyzed by NCHS. The NHIS uses a multistage probability sample. Primary sampling units consist of a county, a small group of contiguous counties, or a metropolitan statistical area. The 52 largest PSUs are forced into the sample. The remaining PSUs are organized into 73 strata, and 2 PSUs are chosen from each stratum with probability proportional to population size. Selected PSUs are organized into secondary sampling units (SSUs) of four to eight households. SSUs are sampled at different rates to meet design objectives such as oversampling blacks or Hispanics. Generally, all households within a selected SSU are targeted for interview.

We excluded 4 of the 20 key health risk factors from this analysis because we felt that the questions on the two surveys were too dissimilar to allow fair comparison. Table 3 compares a subset of three of these key national estimates from the 1995 BRFSS with national estimates from the 1995 NHIS. As illustrated here, we found that the two surveys generally gave consistent results. The difference noted for percent reporting general health is excellent or very good, that is, 4.6%, is the second largest difference we found among all 16 comparisons.

Table 3. Comparison of 1995 NHIS and BRFSS stratified national estimates

Key Health Risk Factors	Estimate (Standard Error) BRFSS	NHIS	Absolute Difference in Estimate
Percent reporting ever having had a pneumonia vaccination	13.95% (0.20)	11.75% (0.30)	2.20%
Percent reporting general health is excellent or very good	57.05% (0.27)	61.64% (0.31)	–4.56%
Average number of cigarettes smoked per day by everyday smokers	18.40 (0.13)	20.20 (0.22)	–1.80

Conclusions

In summary, the following major conclusions may be drawn from the various tasks described here:

- Combining data across states is feasible but depends on each state's ability to minimize both sampling and nonsampling errors. Valid and precise state-level estimates are important so that the data can be combined across states. It is important that documentation of data quality be carefully considered prior to aggregating or comparing data.
- Although there is very little difference between the national estimates obtained using the restratified and pooled methods, the former technique may yield more accurate standard errors when more than one type of design is used because it may more explicitly consider the varying state sample designs and state population sizes. In practical terms, the additional complexity introduced by restratification may be less warranted because BRFSS has moved away from Waksberg designs since the mid-1990s. For example, only two states use Waksberg-type designs in 1999; all others use list-assisted.
- The BRFSS and the NHIS give consistent results for most items we examined. Comparisons require careful attention to skip patterns and question wording. In addition, it is important to remember that the BRFSS and NHIS use different methodologies and cover somewhat different populations, and therefore cannot reasonably be expected to yield identical results.

Although the results of the current study suggest that researchers who wish to use the BRFSS for national estimates may do so with some caveats, additional work is needed to further define the uses and limitations of BRFSS for this purpose. For example, BRFSS data are often used to look at subgroups by race, gender, ethnicity, age, and so forth. The analysis here refers only to overall estimates for which there are very large sample sizes. There is a need to assess the utility of the BRFSS for estimating national estimates for small subgroups. Although the comparison of BRFSS and NHIS estimates yields reassuring results for these items, it is always advisable to consider the implementation of the BRFSS in individual states. The CDC is making progress in this area through the development of additional databases related to design implementation and data quality, and expects to continue expanding such activities as well as continuing methodological studies and adding to data documentation. Additional focus upon implementation of the survey rather than just upon stated survey design may reveal information about comparability among states as yet uncovered. Although such information is unlikely to cause researchers to decide to forgo state comparisons or national estimates, additional research will be beneficial in assessing overall quality of the data, and may also be useful in guiding research choices related to subgroups, health indicators, and methodology.

References

Demo-Detail. (1995). *1995 population estimates by county.* 2303 Apple Hill Road, Alexandria, Virginia 22308.

Public Health Service. (1991). *Healthy People 2000*: *National health promotion and disease prevention objectives.* DHHS Publication No. (PHS)91-50212. Washington, DC: U.S. Department of Health and Human Services.

U.S. Bureau of the Census. (1994). *Phoneless in America.* Statistical Brief 94-16. Washington, DC: U.S. Department of Commerce.

U.S. Bureau of the Census. (1998). 1990 to 1997 annual time series of state population estimates by age, sex, race, and Hispanic origin. July 1, 1995 files. Internet web site: http://www.census.gov/population/www/estimates/st_sasrh.html.

Waksberg, J. S. (1978). Methods for random digit dialing. *Journal of the American Statistical Association, 73,* 40–46.

State Estimates of Substance Abuse Prevalence: Redesign of the National Household Survey on Drug Abuse (NHSDA)

Joseph Gfroerer, Doug Wright, and Peggy Barker

Introduction

The National Household Survey on Drug Abuse (NHSDA) is the federal government's primary source of information on the magnitude of substance use and abuse in the general population of the United States. Conducted since 1971, the survey collects data by administering questionnaires to a representative sample of persons aged 12 and older at their place of residence. The survey is administered by the Substance Abuse and Mental Health Services Administration (SAMHSA), with data collection done under contract. Since 1988, Research Triangle Institute (RTI) has been contracted to conduct the survey. Data from the survey are used by policymakers and researchers to measure the prevalence and correlates of licit and illicit drug use, to identify and monitor trends in substance use, and to analyze differences by population subgroup. Starting with the 1999 survey, the sample was expanded and redesigned to provide the capability to estimate substance use prevalence for all 50 states. This paper describes this new NHSDA design and the data it will generate, as well as the problems and issues involved in its implementation. Also included are a summary of the old survey design and a discussion of the other major change implemented in the NHSDA in 1999—the conversion of the survey from paper-and-pencil interviewing (PAPI) to computer-assisted interviewing (CAI).

Data Collection Methodology

The basic methodology has remained unchanged since the survey began. In each sample dwelling unit, a knowledgeable household respondent is asked for information on the household composition, including the age, race, ethnicity, and gender of each household member. This screening takes about five minutes. Based on this information, either zero, one, or two household members are selected for the full NHSDA interview. Face-to-face interviews are then conducted with sample persons, with demographic questions interviewer-administered and the more sensitive questions about drug use self-administered. No proxy interviews are allowed, no respondent names are collected, and both anonymity and confidentiality are promised. Response rates have generally been around 93% for the household screening and around 80% for the interview.

Old (Pre-1999) NHSDA Methodology

Prior to 1999, screening and interviews were both done using PAPI. The self-administered questions (SAQ) were given to the respondent on a series of answer sheets, which he or she read and filled out without assistance from the interviewer, unless assistance was requested. The answer sheet methodology was used to encourage honest reporting of sensitive information by allowing respondents to report these behaviors in privacy. To enhance respondents' privacy and keep their task simple, the answer sheets that covered basic use patterns of major substances did not include any skip patterns. After its completion, each answer sheet was placed in an envelope, and at the end of the entire interview the envelope was sealed and mailed to the contractor's data processing site. Thus, there was no field editing of the respondent-completed answer sheets, and respondents were not recontacted to resolve inconsistencies in their responses. The methods used in the NHSDA to enhance privacy, anonymity, and confidentiality have been shown in several studies to improve reporting of drug use behaviors, relative to other methods such as telephone surveys or interviewer-administered face-to-face surveys (Turner, Lessler, & Gfroerer, 1992).

Development of CAI for the NHSDA

While it is clear that privacy is essential and the self-administered method is preferred, it is difficult to design written questionnaires with a level of literacy appropriate for children as young as 12 years old and older persons with varying reading abilities. Audio computer-assisted self-interviewing (ACASI) is a methodology that may balance the need for privacy, the need for reducing response burden, and improved data quality through edit checks during the interview. The ACASI methodology allows the respondent to listen to questions through a headset and/or read the questions on the computer screen. Respondents also key their own answers into the computer. Under ACASI, the self-administered format found to increase reporting of sensitive behaviors in the NHSDA can be maintained, and greater privacy can be assured for the

The authors are with the Division of Population Surveys, Office of Applied Studies, Substance Abuse and Mental Health Services Administration, Rockville, Maryland.

respondent even in interview settings that might not otherwise be considered sufficiently private. Programming the questionnaire also allows for more complex skip logic in a format where the routing is less visible to the respondent. Thus, it may be less obvious to a respondent how answering a question in a particular way will influence the number and type of additional questions asked.

In 1995, SAMHSA decided to initiate the development and testing of computer-assisted interviewing in the NHSDA. The development was accomplished primarily under a contract awarded to RTI in early 1996. Research showing the feasibility of ACASI and its potential for improving the reporting of sensitive behaviors was pivotal in SAMHSA's decision to actually develop and test CAI for use in the NHSDA (Turner, Ku, Sonenstein, & Pleck, 1996; Duffer, Lessler, Weeks, & Mosher, 1996). Recently, Turner et al. (1998) obtained greatly increased estimates of injection drug use and other HIV-risk behaviors using ACASI compared to a paper SAQ.

In addition to incorporating the use of CAI instruments for collecting data from respondents, the use of electronic screening of households was implemented in 1999. Prior to 1999, interviewers had to manage a difficult paper-and-pencil interaction while conducting a mini-interview on the housing unit composition, and then execute the respondent selection algorithm, which makes use of complex paper forms. These paper forms were difficult to manage, prone to error, expensive to process, and limiting in terms of the sample selection algorithms that could be implemented. The 1999 NHSDA uses a pen-based, handheld computer (Apple Newton) for conducting the NHSDA dwelling unit screening process.

SAMHSA carefully considered the shift to computer-assisted screening and interviewing, requiring extensive testing and proof of the feasibility of the new technology for the NHSDA. The testing protocol included a small (n = 400) initial field test in the fourth quarter of 1996; small-scale cognitive laboratory testing; a second, larger (n = 1,982) field test in the fourth quarter of 1997; and a final pretest conducted in August 1998. A comprehensive report on the development and testing of the NHSDA CAI instrument will be published by SAMHSA in 2000.

1999 Methodology

The CAI NHSDA ultimately fielded in 1999 incorporated the features tested throughout the prior two years of development that were determined to decrease burden and increase data quality. These include electronic screening, range edits and consistency checks throughout the questionnaire, and inconsistency resolution in ACASI. The CAI questionnaire was programmed in Blaise 4.0 for Windows.

Once selected dwelling units have been located, the screening process, including case management, is accomplished entirely without paper. Each field interviewer's (FI) list of assigned cases, responses to screening questions and resulting selection of sample person(s), record of calls, and status of each case are stored on the handheld Newton computer. Each day, FIs transmit all screening work to RTI headquarters, and data are available through the project web site to RTI and SAMHSA staff for monitoring field progress.

After a sample person is selected, a unique questionnaire ID number is generated by the Newton, which is then entered into the laptop computer by the FI in order to start the interview. The FI conducts the initial CAPI portion of the interview and then turns the laptop over to the respondent, pointing out the keys he or she will use and the headset volume control. Before the actual ACASI portion of the interview begins, the respondent is presented with a short, interactive ACASI tutorial that provides basic instructions and practice in entering responses to different types of questions, changing responses, and having questions repeated. The FI makes every effort to ensure that no person other than the respondent can see or hear the questions. After the respondent completes the ACASI portion, the FI administers the remaining CAPI portion of the interview. Interview data are also electronically transmitted daily by FIs.

In order to provide the capability to adjust prior estimates to maintain comparability for long-term trend analyses, a supplemental PAPI sample of 15,000 was included in the 1999 NHSDA. This sample size is large enough to measure differences in drug use prevalence estimates caused by the change in instrumentation. For this sample, the questionnaire used is essentially identical to the 1998 PAPI questionnaire.

Questionnaire Content

The NHSDA instrument is divided into "core" and "supplemental" components. Questions designated as "core" are included in the first half of the interview and compose the critical data items in the NHSDA. "Core" data are collected on certain basic demographic characteristics (age, gender, employment, education, marital status) and on the use of 12 substances: tobacco, alcohol, marijuana, cocaine, "crack" cocaine, heroin, hallucinogens, inhalants, and nonmedical use of analgesics, tranquilizers, stimulants, and sedatives. Data are collected on the recency, frequency, and initiation of use of each substance. Questions designated as "supplemental" appear in the second half of the interview. These items have been revised, reordered, or even replaced with different items from year to year. Supplemental questions include opinions about drugs, problems associated with drug use, and drug abuse treatment experience. Also collected are data on income, health status, health insurance, utilization of various health services, mental health, workplace issues, and responses about various risk and protective factors associated with youth drug use. In developing special topic modules, the NHSDA has made extensive use of cognitive laboratory testing and small-scale field tests. The content of the 1999 CAI questionnaire is similar to the 1998 PAPI questionnaire, although there were a few changes, the most significant being the addition of questions on usual brand for each tobacco product. The median time to complete the 1999 NHSDA interview is 55 minutes for youths age 12–17 and 49 minutes for adults age 18 and older. Interviews rarely take more than 90 or less than 30 minutes. The CAI inter-

view is considerably shorter than the NHSDA PAPI, which took an average of one hour to complete. With CAI and PAPI, the survey employed both English and Spanish versions of the questionnaire.

Data Limitations

Many of the important prevalence measures from the NHSDA are based on self-reports of sensitive and often illegal behaviors. Although the methods used in the survey have been shown to be effective in reducing reporting bias, there is still probably some unknown level of underreporting that occurs. Furthermore, the underreporting is believed to be more severe for the most deviant behaviors, such as heroin use or heavy use of other drugs. Second, the sample coverage of the populations with the most severe drug problems is probably worse than the coverage of less heavy drug users or nonusers. Some populations with large numbers of heavy drug users (e.g., the prison population) are excluded from the sample by design. One method that SAMHSA has used to account for this underestimation in national estimates is a ratio adjustment procedure that links NHSDA self-reports of arrest and drug treatment with separate counts of arrest and treatment from external data sources believed to be more accurate for these measures (Wright, Gfroerer, & Epstein, 1997). Finally, the underreporting and undercoverage problems are compounded by the fact that many of the behaviors of interest are rare in the population, resulting in only a small number of sample persons with the behavior even in a survey as large as the NHSDA.

Sample Design

1971–1998 NHSDAs

Between 1971 and 1990, the NHSDA sample covered the household population in the 48 contiguous states. In 1991, the sample was redesigned so that study results could be used to make inferences about the entire civilian, noninstitutionalized population aged 12 and older. In order to do this, Alaska and Hawaii were added to the sample frame, as were civilians living on military bases and individuals living in noninstitutional group quarters, such as homeless shelters, college dormitories, and boarding houses. Persons excluded from the sample are the homeless who never use shelters, active military personnel, and residents of institutional group quarters, such as jails and hospitals.

The sample has always been a stratified, multistage, area probability sample of the target population. The first stage of sampling was a set of primary sampling units (PSUs) (usually around a hundred) that consisted of counties or groups of counties such as metropolitan areas. Within a PSU a sample of segments, formed from Census blocks, was selected, and within a segment a sample of dwelling units was selected. At each dwelling unit, zero, one, or two persons were selected for interview. Although some sample segments were included in the sample for consecutive years, a new sample of dwelling units was used each year.

From 1971 to 1990, the survey was conducted every 2–3 years and each was done over a several-month period (different for each survey). Since 1990, the survey has been administered every year, and since 1992 the NHSDA has been conducted continuously (January–December each year) using quarterly samples, eliminating seasonal bias. Since 1991 the sample size has varied between approximately 18,000 and 31,000 persons per year, based on a national sample of 18,000 and supplemental samples in six metropolitan areas during 1991–1993 and in California and Arizona in 1997 and 1998. Oversampling of young age groups (since 1971) and of blacks and Hispanics (since 1985) have been incorporated into the sample design. To achieve the desired sample allocations, it has been necessary to screen approximately 3–4 times more households than the number of completed interviews.

Beginning in 1997, in order to better measure household effects on substance abuse, the sample design was modified to produce representative samples of pairs of household residents. In previous NHSDAs, selection probabilities were assigned to individual household members, and within-household pairs occurred ad hoc, depending on the particular within-household sampling algorithm used. Starting with the 1997 NHSDA, the selection of these pairs is random, with known selection probabilities, and with analytic pair weights calculated for use in special analyses. Giving all pairs a probability of selection has a very small effect on the precision of other estimates.

Why the NHSDA Sample Was Expanded in 1999

With the passage in 1996 of voter initiatives that legalized marijuana use for medical purposes in two states (California and Arizona) and the substantial role of federal block grant funds given to states for substance abuse prevention and treatment, Congress and the Clinton administration concluded that it would be useful to have comparable state-level estimates of substance abuse prevalence. The House Appropriations Committee Report accompanying the DHHS FY 1997 appropriation bill called for SAMHSA to expand the survey and use the new state-level data to improve the provision of treatment and prevention services in states with high levels of substance abuse. Based on the level of funding allocated by the legislation, key to the sample expansion was the ability to make estimates for most states using only modest sample sizes in conjunction with small-area estimation modeling techniques. SAMHSA determined that this was feasible based on its prior experience with modeling for selected states using the 1991–1993 NHSDA (SAMHSA, 1996, 1997), as well as a sampling plan for 1999 that would facilitate state-level estimation.

Description of New Sample Design

The sample design in 1999 will ensure that each of the eight states with the largest populations have a sample designed to

yield 3,600–4,630 completed interviews per year, allowing direct estimates to be produced annually. These eight states are California, New York, Texas, Florida, Pennsylvania, Illinois, Michigan, and Ohio. The remaining 42 states plus the District of Columbia will each have a sample designed to yield 900–1,030 completed interviews each year, for a total sample of about 70,000. There is no longer any oversampling of racial or ethnic groups. Approximately one-third of the sample will be allocated to each of three age groups (12–17, 18–25, 26+). To achieve this age allocation, approximately 200,000 dwelling units will be screened.

Each state was first stratified into a number of geographic field interviewer (FI) regions of roughly equal population size. The eight large states have 48 FI regions, while the other states have 12 FI regions, for a total of 900 regions. The FI regions are typically composed of counties, groups of counties, or subareas of counties depending on the density of the population. The first stage of sampling is at the segment level, with about 7,200 sample segments selected (eight sample segments per FI region) from a total of about a half million segments in the entire United States. Segments are defined by combining adjacent blocks to create nonoverlapping areas that contain at least 150 occupied dwellings. There is no initial stage of county-based PSU selection, as is done in many national household surveys, including prior NHSDAs. In each of the 7,200 selected segments, a listing of all dwelling units is made and a sample of these dwelling units is selected for screening.

With data collection being continuous throughout the year in order to avoid any quarterly seasonality effects, the sample of (eight) segments in each FI region is randomly allocated across the four quarters. Half of the 1999 sample segments (one from each quarter-segment combination) will be retained for the 2000 NHSDA sample, and a new sample of 3,600 segments will be selected for the year 2000. This 50% overlap will continue in successive surveys, with all segments retained for two years (with a new sample of dwelling units in the second year).

Difficulties in Implementing the Expanded Sample

The size and distribution of the 1999 sample units across the 50 states posed a challenge for data collection operations. In 1998 and earlier years, the sample was distributed across 100 to 135 PSUs with segments selected within each PSU. Each PSU consisted of a county or multiple counties, ranging upward to 6–7 counties in some areas of the country. This sample design had been fairly easy to staff for the necessary data collection because recruiting efforts could be focused in the areas in and around the PSUs, given that the segments were clustered and contained within the counties. With the increased sample size in 1999 and the distribution of the sample across 7,200 sample segments distributed across 900 FI regions located in all 50 states and the District of Columbia, the task of recruiting experienced and suitably qualified interviewing personnel was a problem for project managers. An interviewing staff of approximately 1,350 personnel is needed to handle the sample allocation and distribution. A large number of the interviewers hired for the 1999 survey were inexperienced. Although all attended a seven-day training program before going into the field for the first time, many still lacked the skills required to convincingly explain the requirements of the survey and to obtain the cooperation of sample respondents. As a result, there was a decline in response rates relative to prior NHSDAs.

Estimates That the New Design Will Provide

The sample was designed to produce both model-based and sample-based state-level estimates of a variety of substance use measures. Generally, model-based estimates will be used for annual estimates for states with small sample sizes, and direct sample-based estimates will be used for the eight states with large samples. The larger sample will also provide much greater capability for national and regional estimates, and these will continue to be a major focus of NHSDA analyses.

Sample-Based Estimates

The precision of sample-based state estimates will depend on the state sample sizes. The eight largest states will have sample-based estimates that are twice as precise as those for the other states. In other words, standard errors for comparable measures will be about half as large for the large states as for the small states. Sample-based estimates for the smaller states may require the accumulation of a number of years of data to achieve sufficient precision. Over three years, the sample for large states will be 10,800, and for small states 2,700 persons. These sample sizes will provide reliable state-level and substate-level estimates for the multi-year span that would be useful for a number of purposes, assuming that year-to-year changes are not significant or critical to the analysis. Since the sample is designed to be approximately self-weighting within each state by age group, and the effect of clustering is minimal, the effective sample size will be only slightly smaller than the actual sample size for these subgroups.

At the national and regional levels, sample-based estimates of substance use prevalence will be produced just as in prior years, but with improved precision. Taking into account all of the design changes, the approximate fourfold increase in the sample size from 18,000 in 1996 to 70,000 in 1999 is expected to increase the effective sample size for national estimates of illicit drug use prevalence by a factor of about 4 overall, and by a factor of about 6 or 7 for the age group 12–17. Prevalence estimates for rare behaviors and for small population groups will be possible to a much greater extent than with previous NHSDAs.

Model-Based Estimates

Small-area estimation (SAE) modeling techniques will be used to develop annual prevalence measures for the 42 smaller

states and the District of Columbia. The approach used for these models will be a survey-weighted empirical Bayes approach that estimates the parameters in a mixed model. The modeling combines the direct survey data from each state in a regression model that employs a variety of local indicators with model-based results from the rest of the nation. The methodology will be an improved version of the methodology used previously by SAMHSA for 1991–1993 state estimates (SAMHSA, 1996, 1997). Improvements made to the methods used in the previous (1991–1993) SAE analysis, including better evaluation methods, have been developed and tested using the 1994–1996 NHSDA data. The method of estimation to be used is the full hierarchical Bayes method, which is more robust for low prevalence items and small samples than was the method (penalized quasi-likelihood) used for the 1991–1993 data. In addition, more accurate prediction intervals are possible because of the use of Gibbs sampling methods.

The estimates will be validated in a number of ways. One method will be to take subsamples of 900 in each of the eight large states, run the models, and compare the resulting model-based estimates to the direct estimates based on the full samples of 3,600 in those states. This method was tested with the 1994–1996 data from California, which had a large and representative sample. Results show that the SAE statistics have prediction intervals that are approximately half as wide as the comparable design-based confidence intervals.

The SAE produces state-level estimates by first estimating prevalences for each block group and then aggregating block group–level estimates to the state level. Since there are no random effects at the block group level, it is impossible to provide comparable prediction intervals for block group prevalence estimates. Also, since the samples at the block group level are extremely small, estimates at that level would mainly reflect the national model. Estimates for other substate geographic levels, such as the county or FI region, may be possible given sufficient sample size because random effects are included at these levels, and staff are exploring this possibility.

The list of dependent variables for the 1999 SAEs currently includes ones that focus on youth prevention and substance abuse treatment need: past-month binge alcohol use, past-month cigarette use, past-month marijuana use, past-month illicit drug use, past-month illicit drug except marijuana use, past-year dependence on illicit drugs, past-year dependence on alcohol or illicit drugs, past-month alcohol use, past-year incidence of marijuana, perceived great risk of smoking marijuana once a month, perceived great risk of smoking one or more packs of cigarettes every day, perceived great risk of having five or more alcoholic drinks once or twice a week, past-year receipt of treatment for illicit drugs, past-year receipt of treatment for illicit drugs or alcohol, past-year needed treatment for illicit drugs or dependency on alcohol, past-year needed treatment for illicit drugs, past-year cocaine use, and past-month tobacco use. The state SAEs for these variables will be more precise than the comparable sample-based estimates; however, it is important to note that the SAE is limited to these variables, while sample-based estimates can be made for any variable in the survey instrument.

Based on results from both the 1991–1993 and 1994–1996 data, state estimates tend to bunch around the center of the distribution and spread out some in the tails. For this reason, it will not be practical to provide a unique ranking for each of the 50 states. However, it is expected that the estimation will identify those states with relatively high or low estimates of substance use.

Estimation and Analysis Issues

Comparability of NHSDA State Estimates

A major analytic issue is the issue of balancing the need for good national estimates with the need for good state estimates. Since the sample sizes are much larger and the design effects somewhat smaller for the 1999 survey than for prior years, it will not be difficult to meet the national precision goals. However, since one of the main purposes of the increased sample is to make state estimates in order to compare states, the focus must be on maintaining the comparability among states.

One important aspect of state comparability is trying to maintain similar levels of response in all states. It would not be desirable to have a higher national response rate if it meant unacceptably low response rates in certain states. For the first half of 1999, there was significant variation in response rates across states. It will be important to report the response rates for individual states so that analysts can have a sound basis for judging the comparability of the estimates. In addition, it will be important to appropriately weight the data across the four quarters of data collection (i.e., give nearly equal weight to each quarter's data within each state) so that variations in sample size in the quarters do not bias state comparisons.

In the same vein, the desire is to conduct all aspects of estimation within each state rather than across states. These aspects include editing, imputation, nonresponse adjustment, weight trimming, and post-stratification. This kind of estimation preserves the "unbiasedness" of state estimates at the expense of increasing the national mean squared error. However, this could also create difficulties because of the relatively small sample sizes in each state, which may be inadequate for implementing methods such as hot deck imputation for drug use behaviors with low prevalences.

It should be noted that because of the new CAI data collection, the need for editing of inconsistent responses is not as significant as it was in the NHSDA PAPI. The use of skip patterns in the core drug use items and the numerous consistency checks that are built into the ACASI portion of the questionnaire result in much more consistent data than was seen with the SAQ without skips, where respondents were asked multiple times for the same information. The editing and imputation methods employed can potentially significantly impact some of the rare hard-core drug use measures. The goal will be to minimize this and to report the amount of incomplete data for these kinds of drugs. At the state level, variable rates of missing data are likely, so that documentation at the state level becomes imperative.

Comparisons of NHSDA State-Level Estimates to Other Surveys in States

There are a variety of surveys of substance use conducted by states and others that result in state-level data. These other surveys are, for the most part, school surveys or household telephone (random digit dialing) surveys. A number of them are directed at youth. The sample sizes of many of these surveys are significantly larger than the NHSDA state sample sizes, and questions concerning comparability will arise when the NHSDA state estimates become available. No doubt the NHSDA results for some states will be at odds with other survey results for those states, creating confusion among policymakers and researchers. Whether the NHSDA estimate for a given state is comparable to a state estimate from another source will depend on the similarity of the questions and their context, the mode of administration, the size of the sample, and the quality of the design implementation. This question will have to be analyzed on a state-by-state basis. Earlier comparisons of the NHSDA to a nationally representative school-based substance use survey (Monitoring the Future) have shown that there are significant differences in results due to the unit of data collection, the mode of collection, questionnaire differences, and potential biases introduced by implementation (Gfroerer, Wright, & Kopstein, 1997). States can probably make the best use of the NHSDA by using the same wording as the NHSDA questions, even if the mode or other aspects of data collection are different. Comparisons with various administrative data that a particular state may have to NHSDA estimates would also help in interpreting the estimates.

Assessing Trends within States

There is interest in assessing trends within states, for research, surveillance, and evaluation purposes. Patterns of substance use often emerge in small geographic areas or larger regions, and then may spread to other areas. The new sample design makes the NHSDA a potentially powerful tool in measuring these phenomena. Policymakers will also look to within-state trends to draw conclusions about the effectiveness of programs and policies that will be implemented at different points in time and place. These kinds of analysis will need to be done cautiously, taking into account the small sample sizes in small geographic areas and the limitations of the model-based estimates. For the model-based estimates, SAMHSA plans to utilize the same national model from one year to the next to maximize comparability over time. By the "same national model," what is meant is that it will be desirable to use the same variables in the model from one year to the next, so that change within a state can be attributed to true changes at the state level as opposed to changes induced by a change in the model. This approach would still anticipate refitting the national model each year with new data. This in turn will typically result in somewhat different coefficients for the same variable from one year to the next.

Data Release and Disclosure Limitation

The results of the 1999 survey will be disseminated in a series of reports that include various measures of the quality and the limitations of the data. The confidentiality of individual respondents must be maintained. This is one of the issues involved in determining what to include with respect to geographic identifiers below the state level in the data files. For researchers, it is our desire to make available public-use files and to include state-level identifiers on the files in order to make them more useful to states. All public-use files will be subjected to disclosure analysis procedures. One possibility for meeting different research needs is to develop two public-use files: one would include the state-level identifiers with more limited substate geography or with more confidentiality disclosure methods applied to the substate geography, and the other would be a national file identifying all levels of sampling except the state level.

References

Duffer, A., Lessler, J., Weeks, M., & Mosher, W. (1996). Impact of incentives and interviewing modes: Results from the National Survey of Family Growth Cycle V Pretest. In *Proceedings of the Health Survey Research Methods Conference*, pp. 147–152.

Gfroerer, J., Wright, D., & Kopstein, A. (1997). Prevalence of youth substance use: The impact of methodological differences between two national surveys. *Drug and Alcohol Dependence*, *47,* 19–30.

Substance Abuse and Mental Health Services Administration (SAMHSA). (1996). *Substance abuse in states and metropolitan areas: Model based estimates from the 1991–1993 National Household Survey on Drug Abuse, Summary Report*. DHHS Pub. No. (SMA) 96-3095.

——— . (1997). *Substance abuse in states and metropolitan areas: Model based estimates from the 1991–1993 NHSDA—Methodology report*. OAS Metholological Series M-1. DHHS Pub. No. (SMA) 97-3140.

Turner, C., Ku, L., Sonenstein, F., & Pleck, J. (1996). Impact of ACASI on reporting of male-male sexual contacts: Preliminary results from the 1995 National Survey of Adolescent Males. In *Proceedings of the Health Survey Research Methods Conference*, 171–176.

Turner, C., Lessler, J., & Gfroerer, J. (Eds.). (1992). *Survey measurement of drug use: Methodological studies.* DHHS Pub. No. (ADM) 92-1929.

Turner, C., Ku, L., Rogers, S., Lindberg, L., Pleck, J., & Sonenstein, F. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science, 280,* 867–873.

Wright, D., Gfroerer, J., & Epstein, J. (1997). Ratio estimation of hardcore drug use. *Journal of Official Statistics, 13,* (4) 401–416.

The National Immunization Survey: A Surveillance System for State and Local Estimates of Childhood Vaccination Levels

Philip J. Smith, Michael P. Battaglia, Danni Daniels, Victor G. Coronado, and J. N .K. Rao

Introduction

The CDC's National Immunization Survey (NIS) was developed to provide ongoing national, state, and local estimates of vaccination coverage levels among children aged 19–35 months. The NIS was implemented in April 1994 to monitor vaccination coverage levels as part of the Childhood Immunization Initiative, a national strategy to ensure high vaccination coverage of children during the first two years of life. The NIS is also used to monitor progress toward achieving the Year 2000 national vaccination coverage goals.

Sample Design

The initial development phase of the NIS examined five immunization surveillance designs:

1. birth certificate follow-back surveys
2. retrospective school-entrant surveys
3. area probability sample surveys
4. provider surveys
5. random-digit-dialing surveys

Evaluation criteria included:

1. need for quick design, testing, and implementation
2. ability to provide vaccination estimates in a timely manner
3. reasonable cost of data collection
4. feasibility of large-scale implementation to provide reliable four-quarter vaccination estimates for 78 separate Immunization Action Plan (IAP) areas covering the entire United States
5. validity of the vaccination estimates with respect to non-coverage bias, nonresponse bias, and response bias

Philip J. Smith, Danni Daniels, and Victor G. Coronado are with the National Immunization Program, Centers for Disease Control and Prevention, Atlanta, GA.

Michael P. Battaglia is with Abt Associates Inc., Cambridge, MA.

J.N.K. Rao is with Carleton University, Ottawa, Canada.

The random-digit-dialing (RDD) approach was chosen after consideration of the strengths and weaknesses of alternative designs. The main concern associated with the RDD approach is the exclusion of children in non-telephone households.

The NIS RDD design uses independent samples of telephone numbers in each IAP area for each quarter. The quarterly samples make it possible to combine four consecutive quarters to form four-quarter estimates of vaccination coverage. The RDD sample in each IAP area is used to screen for households with one or more age-eligible children. The target number of completed interviews per quarter has been 110 per IAP area. With a sample size of 440 interviews for a four-quarter period, the total number of interviews is 34,320 per four-quarter period. The NIS seeks to achieve a 5% coefficient of variation for the four-quarter IAP area estimates of vaccination coverage levels.

The main challenges of the sampling process are to ensure that a valid probability sample of telephone numbers is selected for each geographic area included, to ensure that the desired level of precision for the vaccination estimates is met in each geographic area, to minimize in a cost-effective manner the number of age-eligible children excluded from the survey, and to maintain an up-to-date sampling frame of telephone numbers.

The NIS uses a list-assisted method of RDD (Lepkowski, 1998). This method is used to select an equal-probability-of-selection sample of telephone numbers from the banks of 100 consecutive telephone numbers in an IAP area that contain one or more directory-listed residential telephone numbers. Brick, Waksberg, Kulp, and Starer (1995) indicate that the list-assisted method is subject to a small coverage bias, but that this bias is offset by gains in survey efficiency and lower cost. One approach to eliminate any bias is to also sample from the zero-banks. Using information from the Current Population Survey, we estimate that only 3.4% of telephone children age 1–3 years are in the zero-banks. Sampling from the zero-banks is not cost-effective, because the bias in the overall estimates of vaccination coverage levels caused by the elimination of the zero-banks is very small, and the very low residential-working-number rate in the zero-banks would increase the cost of the survey.

When the IAP area is a city, a county, or a combination of counties, some prefix areas may cover part of the IAP area and part of an adjacent IAP area. In such situations the NIS

applies a plurality rule: If at least 50% of the directory-listed households in a prefix area fall inside the IAP area, the prefix area is assigned to that IAP area. The survey obtains residence location information, which is used to fine-tune the assignment of prefix areas to IAP areas.

The drawing of a quarterly probability sample of telephone numbers to meet the target number of completed interviews with age-eligible children in an IAP area faces some challenges, because the 78 IAP areas differ considerably along several dimensions. They differ on (1) the percentage of telephone numbers that are working residential numbers, (2) the likelihood of contacting a person among those numbers that are residential, (3) the percentage of contacted households willing to complete the screener interview, (4) the percentage of households with an age-eligible child, and (5) the willingness of parents to complete the immunization interview. Although we attempt to minimize differences in the screener and interview response rates among the IAP areas, the other factors lead to considerable variation in the total sample size of RDD telephone numbers needed to achieve a target sample size of completed interviews. We have dealt with this problem by implementing three design tools. First, we use an automated procedure to eliminate a portion of the nonworking and business telephone numbers in the sample before it is dialed by the interviewers (Battaglia, Starer, Oberkofler, & Zell, 1995). Second, we use a statistical model we have developed to predict the number of sample telephone numbers needed in each IAP area for a given quarter of interviewing. Third, after drawing the required sample size of telephone numbers for an IAP area, we divide that sample into random subsamples called replicates. By administering the sample release on a replicate-by-replicate basis, we are able to control the total number of interviews obtained and to spread the interviews for each IAP area evenly across the entire calendar quarter. This ensures that we do not complete all of the interviews in an IAP area with a high birth rate in, say, the first month of the quarter, while the interviews in an IAP area with a low birth rate are spread out over all three months of the quarter.

Household Survey Data Collection

The screening for eligible households and the vaccination history interview are conducted by CATI. Because the sample of telephone numbers covers six time zones, the telephone center is in operation from early morning to past midnight seven days per week. Calls to a specific time zone are not conducted beyond 9 P.M. The initial screening questions determine whether the telephone number reached is a residence and whether the household has any children 12 months to 3 years old. The CATI system determines whether any children are eligible for the survey using birth dates provided by the household respondent. The final set of questions in the screening section identifies the respondent most knowledgeable about the child's vaccination history and asks that person to locate the child's written vaccination record.

The immunization questions used in the NIS were adapted from the in-person immunization questionnaire used in the National Health Interview Survey (NHIS). If a written record is available, the interviewer asks the respondent to report the number and dates of vaccinations. If a written record is not available, the interviewer asks the respondent to recall from memory the number of shots for each vaccine, but not the dates. Upon completion of the immunization questions, the interviewer proceeds to obtain demographic and socioeconomic information. Finally, the interviewer asks the respondent to give the names and addresses of all health care providers who vaccinated the child. Verbal consent is obtained from the parent or guardian to contact the named providers.

The NIS incorporates several special procedures aimed at maximizing household survey response. First, an advance letter is mailed to households with directory-listed telephone numbers (Camburn, Lavrakas, Battaglia, Massey, & Wright, 1995). Second, after reaching an answering machine for the third time, interviewers leave a brief message with a toll-free number for the respondent to call. Third, an extensive ongoing interviewer training and monitoring program is used. Fourth, refusal conversion procedures are used to deal with various types of refusals. Fifth, up to 24 call attempts are made to sample telephone numbers. Sixth, the NIS uses two special approaches to accommodate households whose primary language is not English; the questionnaire has been translated into Spanish, and the NIS also uses a real-time language translation service that provides the ability to conduct the interview in more than 140 languages. Finally, the CATI system uses "hot keys" to allow an interviewer to quickly complete the screener interview for an elderly household without any children.

One of the primary measures of the success of the NIS RDD survey is the overall CASRO response rate (Frankel, 1983). Three assumptions underlie the overall response rate:

1. The proportion of households among unresolved numbers is equal to the proportion of households found among resolved numbers.
2. The proportion of eligible households in the unresolved group is the same as the proportion of eligible households found among households in which screening was completed.
3. The proportion of eligible households among the known but unscreened households is equal to the proportion of eligible households found among screened households.

Under these assumptions, the overall response rate in 1997 (82.9%) is equivalent to the product of the resolution rate (90.2%), the screening completion rate (97.9%), and the interview completion rate (93.8%). The overall response rate varied from a low of 72.0% in New Jersey–City of Newark (followed by 75.2% in New Jersey–Rest of State and 75.5% in the District of Columbia) to a high of 89.1% in Arkansas, with a median value of 83.3%.

Provider Survey

The information that a respondent reports on a child's vaccinations is subject to response bias. If the respondent does not have a vaccination record for the child, some vaccinations may be forgotten and hence not reported. The interview asks about several individual vaccines, and a child may often receive more than one of these in a single visit to a provider. Even when the respondent has a vaccination record, that record may be incomplete (Battaglia, Shapiro, & Zell, 1996). Each child interviewed in the NIS is therefore eligible for inclusion in the provider (record check) survey.

The information from the household respondents is used to contact the providers. Written requests for vaccination histories are mailed to the providers in order to obtain reports of vaccinations from medical records. Providers have the option of responding via mail or facsimile. Postcard reminders and telephone follow-up are used to encourage nonresponding providers to participate in the study. For the 1997 NIS as a whole, the parents of 86% of the 32,742 children in the household survey gave verbal consent. The consent rate ranged from 92% in Texas–El Paso County to 82% in Connecticut, Louisiana–Orleans Parish, New Jersey–Rest of State, and Ohio–Franklin County, with a median of 86%. The percentage of children in 1997 with adequate provider data for use in estimation was 70.1%. Variation was observed in the IAP area rates: 56% in New Jersey–City of Newark to 80% in Maine, Vermont, and Wyoming.

Weights

Estimation procedures in the NIS have been oriented mainly toward calculation of population-based estimates of vaccination coverage in each IAP area, in entire states, and in the nation for a set of four consecutive quarters. Before the actual calculation of the estimates, however, a number of further steps are required:

- Impute missing data for certain data elements.
- Adjust weights of households with multiple telephone numbers.
- Compensate for unit nonresponse in the interviewing process.
- Post-stratify to compensate for noncoverage of households without telephones (and to ensure agreement with population totals on race/ethnicity, mother's education, and age of child).
- Use the results of the provider survey to compensate for biases in the information that respondents give on children's vaccinations.

Each child with data in the NIS receives a base sampling weight, equal to the reciprocal of the probability of selecting the household's telephone number into the sample. Specifically, this weight is the ratio of two totals for that IAP area: the number of telephone numbers in the 1+ working banks, and the number of telephone numbers drawn from those banks and actually released for use. In a few instances the interview reveals that the child actually resides in an IAP area other than the one from which the telephone number was sampled. Because a much larger weight can substantially increase the variance of estimates, each child's base weight is not allowed to exceed three times the base weight for the IAP area in which the child resides, as calculated above.

A household with two or more residential telephone numbers has a proportionally higher probability of being selected into the RDD sample. To preserve the relationship between the base sampling weight and this probability, an adjustment divides the base weight for such a household by its number of nonbusiness voice-use telephone numbers, up to a maximum of 3.

Nonresponse can occur at several points in the NIS interviewing process. At each point a different amount of information is available about the nonresponding telephone number. The NIS applies a separate weighting-class adjustment for each of three amounts of information:

1. The interviewer has identified an eligible household, but the interview has not been completed.
2. The survey has reached a household, but nothing more is known.
3. It is unknown whether the telephone number is residential.

Within each of a set of cells or classes the adjustment increases each respondent's base sampling weight to account for the nonrespondents. For example, where each nonrespondent is known to be an eligible household, each respondent's base weight is multiplied by the ratio of the number of respondents and nonrespondents to the number of respondents. The cells are defined by IAP area, the residential directory-listed status of the sample telephone number, and telephone-exchange-level demographic and socioeconomic characteristics. The result of applying the three adjustment factors in turn is the "nonresponse-adjusted base sampling weight."

Random-digit dialing yields a sample of children in households that have telephones, but the NIS aims to measure vaccination coverage levels for all children 19 to 35 months of age. Data from the NHIS indicate that vaccination levels are generally lower among children from non-telephone households than among children from telephone households. In some IAP areas a substantial proportion of age-eligible children reside in non-telephone households. In attempting to compensate for such potential noncoverage bias, the NIS employs strategies based on post-stratification. Battaglia, Malec, Spencer, Hoaglin, and Sedransk (1995) discuss these and other approaches.

Post-stratification separates the actual sample into cells defined by characteristics that are related to noncoverage and vaccination status. Then the weighted distribution of completed interviews over the cells is brought into agreement

with a corresponding set of population totals. The NIS uses 12 post-stratification cells, based on race/ethnicity (three categories: Hispanic, black non-Hispanic, and white or other race non-Hispanic), mother's education (two categories: less than or equal to 12 years, greater than 12 years), and age of child (two categories: 19 to 25 months, 26 to 35 months). Within each IAP area these cells are combined, according to an overall set of rules, as needed to ensure that the resulting cells contain specified minimum numbers of children. For each of those cells the National Center for Health Statistics' natality files provide a universe of live births. Adjustments for infant mortality, immigration, and migration then yield the population control totals.

Post-stratification assumes that the vaccination rate in each cell is the same for non-telephone children as for telephone children. This strategy, however, does not adequately account for lower vaccination coverage among non-telephone households. Thus the weights currently used in calculating estimates in the NIS are obtained from modified post-stratification, which splits each cell into two subcells: children whose vaccinations are up to date and children whose vaccinations are not up to date. For this purpose, "up to date" is defined in terms of the 4:3:1:3 series of shots. To develop the control totals for these two subcells, one begins with the control total N_g for post-stratification cell g. Applying P_g, the proportion of children in cell g who (according to the 1996–97 CPS) reside in telephone households, yields the number of children in telephone households, $N_{g1} = N_g P_g$, and the number of children in non-telephone households, $N_{g0} = N_g(1 - P_g)$. To obtain the number of children who are up to date and the number of children who are not up to date within cell g, one uses cell-specific up-to-date rates for telephone and non-telephone children. For telephone children the NIS directly estimates the IAP-area-specific 4:3:1:3 up-to-date rate, r_{21g}. In this notation the first subscript indicates the source of the estimate (1 = NHIS, 2 = NIS), and the second subscript distinguishes non-telephone children (0) from telephone children (1); the third subscript denotes the post-stratification cell. The number of telephone children who are up to date is $N_{g1} r_{21g}$. For non-telephone children, data from the NHIS give a national estimate of the 4:3:1:3 up-to-date rate, r_{10g}. It is reasonable to assume that the "up-to-date ratio" r_{10g}/r_{11g}, the ratio of r_{10g} to the corresponding rate for telephone children, applies also at the time of the NIS in 1997. Thus it can be used in place of r_{20g}/r_{21g} to estimate the number of non-telephone children who are up to date:

$$N_{g0} r_{21g} (r_{10g} / r_{11g})$$

Together with the number of telephone children who are up to date (derived above), this yields the control total of children who are up to date in post-stratification cell g:

$$\hat{N}_g \quad N_{g1}$$

The difference,

$$N_g - \hat{N}_g$$

is the control total of children who are not up to date. Thus, splitting each post-stratification cell into two subcells allows the RDD children to receive a post-stratified weight that is a function of whether they are 4:3:1:3 up to date (Hoaglin & Battaglia, 1996).

The modified post-stratification weight is currently used in forming estimates of vaccination coverage levels. We are, however, examining the use of respondent-reported interruptions in telephone service in the previous 12 months as an alternative method for compensating for the exclusion of non-telephone children (Frankel, Ezzati-Rice, Wright, & Srinath, 1998). The weights of telephone children in households that experienced an interruption in telephone service of one week or longer in the past year can be adjusted to represent the population of telephone children with an interruption plus non-telephone children with an interruption plus non-telephone children with no telephone service for the entire year.

Estimation

The primary goal of the NIS is to provide annual estimates of vaccination coverage levels for the 78 IAP areas and the 50 states. The vaccination coverage estimates include up-to-date status on the individual vaccines and completion of vaccination series. In the early years of the NIS, the vaccination reports obtained from providers and the household report of the up-to-date status of the child and use of written vaccination records were used to form stratified two-phase estimates of vaccination coverage (Zell, Ezzati-Rice, Hoaglin, & Massey, 1995). The first-phase sample consists of all children with completed household interviews, and the second-phase sample comprises only children with provider vaccination information.

The statistical estimation methodology that is currently in use for NIS has been designed specifically to adjust vaccination coverage estimates for "vaccination history nonresponse" bias (Smith, 1999). Within each IAP area, the methods achieve this by grouping sampled children into adjustment cells according to both the similarity of their response propensities to have a provider-reported vaccination history, and the similarity of their predictive means of being up to date with respect to the 4:3:1:3 vaccination series. A group of children who have similar response propensities and predictive means will also be similar with respect to the background variables that are predictive of these factors. In this important respect, children within each adjustment cell are ostensibly comparable. Because of this, all of the sampled children in the cell may be represented fairly by the sampled children within the cell who have provider-reported immunization histories by dividing these children's first-phase sampling weights by the cell's weighted response rate. In this way, the bias in immunization coverage rates attributable to differences between sampled children who have and do not have provider-reported immunization histories is reduced to an extent that depends upon the similarity of the background variables that are associated with the response propensities and predictive means used to construct the adjustment cells.

As a first step in forming adjustment cells, a response propensity model was developed using logistic regression. The response propensity is the probability that a sampled child has a provider-reported vaccination history. As candidates for predictors to the response propensity logistic model, we used demographic, socioeconomic, and household vaccination report variables that have been found to be associated with immunization status in other research conducted by CDC. Forward stepwise logistic regression was then used to select predictors among these candidates.

To develop the predictive mean model, within each IAP area the estimated response propensities were used to form preliminary adjustment cells according to the quintiles of the distribution of the estimated response propensities. Within each of these preliminary adjustment cells, the first-phase sampling weights of children with adequate provider vaccination information were divided by the cell-specific weighted response probabilities. The resulting weights were used as prior weights in a logistic regression of the indicator of whether the child's provider indicated that he or she was up to date on 4:3:1:3 on the demographic, socioeconomic, and household vaccination report variables. The forward stepwise model selection method was used to develop a predictive mean model for whether a child is up to date for the 4:3:1:3 series.

The final adjustment cells were formed in a manner so that unweighted sample counts were equal in each adjustment cell. To adjust for vaccination history nonresponse bias, within each adjustment cell children with vaccination histories are assigned a revised set of weights that are obtained by dividing their first-phase sampling weights by the cell-specific weighted response rate. By dividing the first-phase sampling weights of children who have provider vaccination information by their adjustment cell-specific weighted response rate, these children more fairly represent all of the children in the cell as a whole.

However, the revised weights may not match the post-stratification totals used to construct first-phase sampling weights. Also, the revised weights may not match the first-phase sample weighted totals of variables that are known to be important predictors of being up to date. To reduce bias attributable to these differences and to maintain the nonresponse bias adjustment, we rake the revised weights to match post-stratification totals, outcome predictor totals, and the adjustment cell-specific first-phase sampling weight totals.

These methods were applied separately to 1996, 1997, and 1998 NIS data sets. After review of the empirical results obtained by applying thse new methods to the 1998 data, it was found that the reduction in bias obtained using an optimal strategy employing both response propensities and predictive means did not differ significantly from using five adjustment cells formed using the quintiles of the distribution of response propensities in each IAP area. In view of these results, five adjustment cells was designated as the number of cells to be used for adjusting for vaccination provider nonresponse in 1998.

The current NIS estimation methodology also offers the advantage of allowing immunization analysts at CDC to calculate weighted estimates of vaccination coverage in a straightforward fashion.

Dissemination of NIS Vaccination Coverage Estimates

Vaccination coverage estimates for the nation, each of the 50 states, and 28 selected urban areas is disseminated in a timely manner six months after the end of data collection via the CDC's *Morbidity and Mortality Weekly Report*.

References

Battaglia, M., Starer, A., Oberkofler, J., & Zell, E. (1995). Pre-identification of nonworking and business telephone numbers in list-assisted random-digit-dialing samples. In *Proceedings of the Section on Survey Research Methods* (pp. 957–962). Alexandria, VA: American Statistical Association.

Battaglia, M., Malec, D., Spencer, B., Hoaglin, D., & Sedransk, J. (1995). Adjusting for noncoverage of nontelephone households in the national immunization survey. In *Proceedings of the Section on Survey Research Methods* (pp. 678–683). Alexandria, VA: American Statistical Association.

Battaglia, M., Shapiro, G., & Zell, E. (1996). Substantial response bias may remain even when records are used in a telephone survey. In *Proceedings of the Section on Survey Research Methods* (pp. 452–455). Alexandria, VA: American Statistical Association.

Brick, M., Waksberg, J., Kulp, D., & Starer, A.(1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly, 59,* 218–235.

Camburn, D., Lavrakas, P., Battaglia, M., Massey, J., & Wright, R. (1995). Using advance letters in random-digit-dialing telephone surveys. In *Proceedings of the Section on Survey Research Methods* (pp. 969–974). Alexandria, VA: American Statistical Association.

Frankel, L. (1983). The report of the CASRO task force on response rates. In F. Weisman (Ed.), *Improving data quality in sample surveys*. Cambridge, MA: Marketing Science Institute.

Frankel, M., Ezzati-Rice, T., Wright, R., & Srinath, K. (1998). Use of data on interruption in telephone service for noncoverage adjustment. In *Proceedings of the Section on Survey Research Methods* (pp. 290–295). Alexandria, VA: American Statistical Association.

Hoaglin, D., & Battaglia, M. (1996). A comparison of two methods of adjusting for noncoverage of nontelephone households in a telephone survey. In *Proceedings of the Section on Survey Research Methods* (pp. 497–501). Alexandria, VA: American Statistical Association.

Lepkowski, J. (1988). Telephone sampling methods in the United States. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, & J. Waksberg (Eds.). *Telephone survey methodology* (pp. 73–98). New York: John Wiley & Sons.

Smith, P. (1999). Compensating for nonsampling errors in the national immunization survey. Paper presented at the International Conference on Analysis of Survey Data, Southampton, England.

Zell, E., Ezzati-Rice, T., Hoaglin, D., & Massey, M. (1995). Adjusting for respondent bias on vaccination status in a telephone survey. In *Proceedings of the Section on Survey Research Methods* (pp. 684–689). Alexandria, VA: American Statistical Association.

Targeting Approaches to State-Level Estimates

Jennifer H. Madans, Trena M. Ezzati-Rice, Marcie Cynamon, and Stephen J. Blumberg

Existing population-based surveys, such as the National Health Interview Survey (NHIS) and the Medical Expenditure Panel Survey (MEPS), provide much relevant information on many important health-related issues at the national level. Historically, however, none of the major national surveys can provide data on the performance and impact of various programs at the state or local level—data that are necessary to inform public health policy. The National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), has developed several approaches to providing needed data at the state level. Using existing NHIS data, NCHS has created enhanced NHIS data files containing state identifiers for public use. To improve accessibility to more detailed geographic data that has been previously unavailable, NCHS has also established the Research Data Center. In addition, the NHIS sampling frame has been reconfigured to improve state coverage. Further sample design changes under consideration for future use include the augmentation of the in-person interview sample with telephone interviews. Finally, the State and Local Area Integrated Telephone Survey (SLAITS), a population-based, multipurpose, flexible survey that expands the existing National Immunization Survey (NIS), was created to conduct surveys on targeted populations. This paper provides an overview of these efforts with an emphasis on SLAITS.

Background

State-level data are recognized as increasingly important to the public health and health policy communities, especially with the recent changes in the health care market and the increasing responsibilities gained by states for administering various service delivery programs. The development of performance partnership indicators and state-based initiatives in welfare reform and health care coverage make the need for state level data all the more critical. While considerable data are available at the national level to track and monitor these issues, data at the state level are much more limited.

Jennifer Madans is Associate Director for Science at the Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Hyattsville, Maryland. Trena Ezzati-Rice is with the Office of Research and Methodology (ORM), CDC/NCHS. Marcie Cynamon and Stephen Blumberg are with the Division of Health Interview Statistics, CDC/NCHS. We would like to thank Linda Tompkins and Pradip Muhuri, NCHS/ORM, for their assistance in preparing the weighted estimates and estimates of standard error in this paper.

NCHS operates a broad-based program of vital and health statistics data collection systems designed to obtain information on a wide range of health and health-related topics at the *national* level. This includes the collection of information on vital events such as births and deaths for each state, and the collection of a broad range of national health and health-related data either from population-based household interview surveys such as the NHIS or through direct physical examinations such as the National Health and Nutrition Examination Survey. In addition, NCHS conducts a family of record-based surveys designed to supply information about the use of health services provided by medical care providers such as office-based physicians, hospitals, and long-term care facilities. However, few of these surveys could produce statistically reliable data for more than the largest states, and confidentiality concerns often limited the release of data that were available.

Approaches to Obtaining State-Level Estimates

Efforts to obtain state-level estimates have focused primarily on the NHIS. The NHIS has been a major ongoing source of information on the health of the U.S. civilian noninstitutionalized population since the 1950s. The NHIS is primarily designed to provide estimates of various aspects of health and health care for the nation and for subpopulations defined by demographic characteristics. Because it is also the largest in-person survey conducted by NCHS and its sample is used for other surveys, this survey was an ideal starting point for enhancements that would lead to state estimates.

State Files

Initial efforts to permit these state-based estimates led to the creation of the NHIS State Data Files. These files differ from the NHIS public-use data in that statistical noise at both the variable level and record level have been added to allow for the protection of respondent confidentiality and, at the same time, allow for release of the files with state identifiers. Some variables were excluded from these files, and others were recoded to prevent linkage of this file to the annual NHIS data files. Finally, states where the sample size was very small were combined into two groups of three to five states each. Despite these changes, these publicly released data files allow for the production of direct state estimates.

These files are currently available for data years 1990 through 1995 (at this writing, 1995 is still in the review process).

Research Data Center

Even with the state files, requests to NCHS to provide ever more detailed geographic information have increased. A greater number of researchers wish to link NCHS data with exogenous data files. And sometimes the statistical noise and variable suppression in the state files have hindered certain research efforts. These requests have been granted on a limited basis under tightly controlled circumstances; however, with the increased use of the NHIS sampling frame for other surveys such as the National Survey of Family Growth and the Medical Expenditure Panel Survey, granting these requests has become more complicated. In particular, protecting against disclosure of identifiable information—that is, information that could be used by itself due to its unusual characteristics or combined with other information to potentially identify survey participants—has become more difficult. The need to respond to these requests for access to more and detailed information about respondents while still protecting the confidentiality of respondents led NCHS to create the Research Data Center (RDC).

The RDC is located at NCHS headquarters in Hyattsville, Maryland. Researchers can use one of several methods to acquire access to restricted data files—that is, files that contain detailed information such as state, county, or census tract identifiers that are not generally available to the public. Through the RDC, researchers may link files exogenous to NCHS to enhance the amount or type of information available (such as detailed demographic and economic data for census tracts). Research proposals and data output are reviewed by an advisory committee and RDC staff, respectively, to ensure that the utmost care is given to disclosure avoidance. At no time may researchers access direct identifiers such as names, addresses, or Social Security numbers of survey participants. Individuals may work on-site or by remote electronic transmission or by having the RDC staff run programs. The costs associated with each of these activities are borne by the researcher.

The state files and the Research Data Center were created with the goal of increasing the availability of existing data. Efforts have also been made to improve the collection of new information so that state-level estimates can be made more accurately.

Sample Design

Approximately every 10 years, the NHIS sample is redesigned following the decennial census to update the information upon which the sample is based and to incorporate new sample objectives for the coming decade. The most recent redesign is in place for the period 1995–2004. One of the sample objectives for this redesign was to improve the estimates for specific racial, ethnic, and economic subdomains and to enhance the geographic distribution of the sample. The cost of designing the NHIS so that reliable statistics can be produced for all 50 states, as well as for subgroups defined by race and ethnicity and for the nation, is prohibitive. However, some efforts were made with the 1995 redesign of the NHIS to facilitate state-level estimates. The most important was making the primary sampling units (PSUs) respect state boundaries. In addition, at least two PSUs (one metropolitan and one nonmetropolitan) were selected from each state that contained both metropolitan and nonmetropolitan areas. The number of PSUs was also increased from 198 to 358 (National Center for Health Statistics, 1999).

The net result is that all states now have the sample size necessary for state-level estimates when the prevalence rates are relatively large ($p = 0.15$ and $p = 0.20$) and the design effects are relatively small (1.0–1.5) (Westat Inc., 1999). However, for rarer prevalence rates and larger design effects, the number of states with sufficient sample size decreases significantly. In addition, when the population of interest is a subdomain (e.g., children under 18), the number of states with the necessary sample size is smaller still.

Dual Frame Estimation

To boost the sample size, NCHS is investigating the ability to produce direct state-based estimates from the NHIS in combination with a random-digit-dial (RDD) telephone survey. Multiple sampling frames are sometimes used in tandem for population surveys when there is concern about potential bias due to undercoverage by a single frame. In addition, supplementing an area frame sample with an RDD sample has cost savings relative to simply increasing the area frame sample size in the smaller states in the NHIS. Research is ongoing to examine the utility of supplementing the states with an insufficient NHIS sample with telephone interviews to meet desired levels of accuracy using dual-frame estimators. The investigation will include the costs associated with the RDD supplementation and the effect of using dual-frame estimation, response rates, coverage, and bias.

In addition to supplementing the existing NHIS area sample, an RDD survey can also provide the opportunity to collect data on additional topics that might not have been included in the NHIS. Often, the content of ongoing NCHS surveys does not fully address the needs of individual researchers or agencies, especially at the state level. Therefore, NCHS created the State and Local Area Integrated Telephone Survey (SLAITS), a population-based survey with flexible content and sample coverage that can be used to supplement the NHIS, to target specific populations, and to provide a platform for fielding questions in new content areas.

SLAITS

Specifically, SLAITS is a standardized survey mechanism that allows for the collection of population-based data at the state and local levels to address emerging health and health-related issues. It also allows for the comparison of data across states and with national data. To accomplish this, SLAITS

was integrated with two major national surveys—the National Immunization Survey (NIS) and the NHIS. The NIS is an ongoing RDD survey designed to monitor vaccination coverage levels among children 19–35 months old in all 50 states and 28 large urban areas. (The design of the NIS has been previously published; see Ezzati-Rice, Zell, Battaglia, Ching, & Wright, 1995.) Because the proportion of households with age-eligible children in the NIS is small (about 4%), about 25 randomly selected households must be screened to identify a single household with an age-eligible child. In fact, approximately 1 million households are contacted each year in the NIS. By sharing the NIS sampling frame and using this large number of telephone numbers for its base, SLAITS has been able to economize on the cost of selecting and screening households. All households contacted for the NIS are potentially eligible for SLAITS, regardless of whether or not they are eligible for NIS.

In addition, SLAITS makes use of questions in the NHIS and its data for telephone and non-telephone households to account for the effect of not covering households without telephones in SLAITS. Many of the questions included in the SLAITS questionnaire modules are taken from the NHIS so that SLAITS data can be compared with national data and because many of the questions have undergone extensive testing.

SLAITS is designed to allow flexible sample selection and questionnaire modules to address geographical and/or topical requirements. It was created in conjunction with the Department of Health and Human Services' plan for the integration of existing surveys (Ezzati-Rice, Cohen, Khare, & Moriarity, 1998). By using the NIS sampling frame and questions from multiple existing surveys—including the NHIS, the Survey of Income and Program Participation (SIPP), the Current Population Survey (CPS), and other surveys—SLAITS has increased analytic capabilities at decreased costs. Since SLAITS is a telephone survey, it has additional cost efficiencies, with all interviews conducted from one central location using Computer Assisted Telephone Interviewing (CATI). This allows for rapid implementation, rapid data collection, and timely data release—desirable characteristics of any survey.

Due to the fact that certain sociodemographic groups (for example, those with low family income) are more likely not to have telephones, potential non-telephone coverage bias is an issue in this survey, just as it is for all telephone surveys. Therefore, direct adjustments are made for noncoverage of non-telephone households. Two different methods have been investigated, including the use of information from telephone and non-telephone households in the NHIS for selected key variables included in both surveys. More recently, information on interruption in telephone service has been examined as an enhanced method to adjust for non-telephone coverage bias (Frankel, Ezzati-Rice, Wright, & Srinath, 1998).

Key Features of SLAITS

SLAITS has been designed as a population-based survey mechanism with topical questionnaire modules, oversampling options, quick data turnaround, and flexibility. In brief, selected key features of SLAITS include:

- A centrally administered state-based telephone survey linked to the NHIS, effectively creating state-level health interview surveys
- Use of standardized questions, survey methodology, and mode of administration to provide data that are comparable across states and with national data
- Use of questions from the NHIS and other existing national household surveys that are administered in person to allow for statistical adjustments for households without telephone coverage and to allow for eventual dual-frame surveys using both area frame and RDD surveys
- Efficient use of the sampling frame for the existing NIS, an ongoing telephone survey designed to produce state and selected urban area estimates of vaccination coverage levels among young children
- Flexible sampling frame to allow states to sponsor the collection of data at the substate level
- Flexibility to target policy-relevant subgroups of the population and to customize questionnaires to meet state-specific needs for data (e.g., states can sponsor the inclusion of questions to ascertain respondents' knowledge of program availability and requirements or to assess utilization of specific state-based services or programs)
- Rapid implementation and quick turnaround of data
- Demonstrated high response rates

SLAITS Pilot Studies and Selected Results

Prior to large-scale implementation of SLAITS, pilot studies of two different questionnaire modules and study designs were undertaken. The primary objectives of these pilot studies were to evaluate the integration of the NIS and SLAITS sample design, to test a general health module that would provide NHIS data at the state level for an abbreviated set of variables, to test a mechanism for collecting data for all members of the household, to test a child well-being and welfare module including special oversampling design features, to test an alternative statistical adjustment procedure to the one in use by the NIS to adjust for non-telephone coverage bias, and finally to evaluate survey participation rates. At the present time, two questionnaire modules have been tested using the SLAITS mechanism: a Health Module and a Child Well-Being and Welfare (CWBW) Module. Other modules are under development for implementation in 2000.

Health Module

Study Design

The SLAITS Health Module closely approximates a state-based NHIS. Questions from the NHIS have been supplemented

by items from the SIPP. This module focuses on access and barriers to care, health insurance coverage, health status and limitations of activity, health care utilization, demographic characteristics, family income, and family structure. This module is designed to collect information on all members of the household.

A pilot test of this module was conducted in two states, Iowa and Washington, in summer 1997. The sample of households selected for SLAITS was essentially a subsample of households screened for the NIS in the states of Iowa and Washington. In selecting the sample of households for SLAITS, the NIS sampling frame (a representative sample of telephone households) was used as a second-phase sampling frame in the two pilot study states. The goal for the pilot study was 1,000 completed household interviews in each state. For those households that were eligible for NIS (approximately 50 of the 1,000), the SLAITS questions were simply appended to the regular NIS immunization interview. For those households that were eligible only for SLAITS (approximately 950 of the 1,000), the Health Module questions followed the NIS screening questions. As in the NIS, an advance letter was sent to those households with a directory-listed telephone number in order to increase survey participation. The interview was completed with a household member age 18 or older. The interviews averaged 20.2 minutes in length among NIS-ineligible households and 34.0 minutes (including the NIS immunization interview) for NIS-eligible households.

Response Rates and Selected Results

The combined total number of households interviewed in Iowa and Washington was 2,089. These interviews provided data for 5,541 persons of all ages (2,675 in Iowa, 2,866 in Washington) and 1,543 children under age 18 (738 in Iowa, 805 in Washington). The interview completion rates among households were 76% in Iowa and 75% in Washington.

Taking into account the resolution rate and enumeration rate as well as the interview completion rate, the final CASRO (Council of American Survey Research Organizations, 1982) response rates were calculated as 68% in Iowa and 66% in Washington. Unfortunately, due to resource limitations and schedule considerations, the data collection period for some replicates was ended before all cases could be completely worked. If the resources had been available to extend the field data collection period, higher response rates could have been achieved.

For producing population-based estimates of totals and percentages, a survey weight was attached to each sample person. This weight combined the base sampling weight (reflecting the probability of selection of an individual in the sample) with an adjustment for households that have multiple telephone numbers, an adjustment to compensate for unit nonresponse, a post-stratification adjustment to a set of known population totals, and finally an adjustment to account for noncoverage of non-telephone households using information from the NHIS. Weighted estimates and estimates of standard errors were obtained using SUDAAN (Shah, Barnwell, & Bieler, 1997).

To assess the reliability of the data collected, data from the SLAITS module were compared with data collected using the NHIS in Iowa and Washington. For example, the uninsured rate in Iowa (6.5%) was lower than in Washington (9.9%). In both states, the uninsured rate for children under 18 (4.2% in Iowa, 5.7% in Washington) was lower than the uninsured rate for adults aged 18–44 years (10.9% in Iowa, 16.7% in Washington). As expected, uninsured and insured respondents reported different health status and health care experiences. Nearly twice as many uninsured persons reported fair or poor health status compared with the insured. In addition, the uninsured were less likely to have visited a doctor in the past year and were six to seven times more likely than the insured to report a problem in getting medical care. These findings closely mirror those found by the NHIS for these states.

Child Well-Being and Welfare Module

Study Design (Texas)

The SLAITS Child Well-Being and Welfare (CWBW) Module is designed to provide estimates of measures related to the transition from welfare to work. The impact of this transition on children in poor families is of particular interest. Therefore, the questionnaire is targeted to families below 200% of the federal poverty level and covers aspects of child well-being such as family structure, stability and turbulence, psychosocial characteristics of parents, neighborhood characteristics, academic and school behavior, and child care arrangements. This module also includes economic indicators of well-being such as welfare program participation, income and earnings, health insurance coverage, and education and employment of adults in the household. The measures used were drawn from numerous national surveys including the NHIS, SIPP, the National Household Education Survey, the Survey of Program Dynamics, and the National Survey of America's Families (conducted by the Urban Institute). This module was recently tested with a random sample of households with children in Texas.

The SLAITS CWBW Module in Texas collected information on the health and well-being of children under 19 years along with information on welfare program participation of households with the targeted sample children. This pilot study targeted a sample of 660 households with children whose screened household income was below 200% of the federal poverty level and a sample of 588 households with children under 19 years and income above 200% of poverty. To accomplish this, a general income screening question was added early in the interview. A household respondent 18 years of age or older was asked about the number of household members, the number of children under age 18, and the standard NIS eligibility questions. Respondents in NIS-ineligible households were then asked whether the household income in the last calendar year was above or below a dollar amount determined to be 200% of the federal poverty level based on the household's size. Respondents in NIS-eligible households first completed the NIS interview and then were asked this income question. All households with children that

reported income below 200% of poverty were included in the sample. Initially, households that reported income above 200% of poverty were subsampled at the rate of 1 in 1.8. This subsampling rate was later adjusted because the actual number of households reporting income below 200% of poverty was greater than anticipated from 1997 CPS data. Using 1997 CPS data, the target sample size, an assumed response rate of 85% to the screening interview, and an assumed 85% interview response rate, the screening sample size was determined to be 5,742 households.

For this study, rather than collecting data for the child well-being questions for every child in the household (which could make for a very long interview), a maximum of two children were randomly selected from each sample household. Children in each household were stratified into two age groups: All children from 0 to 5 years old were in the first stratum, and children between 6 and 17 years old were in the second stratum. If children were present in both age groups, then one child was selected at random from each age group. If there were children in only one age group, then no more than two children were selected at random from that household. In single-child households, the child was included in the sample.

The CWBW interviews (including the SLAITS portion of the screener) averaged 30.3 minutes in length among NIS-ineligible households and 45.1 minutes (including the NIS portion) for NIS-eligible households. Among the NIS-ineligible households, the interview ranged from a mean of 26.9 minutes for a single-child household to 32.5 minutes for a household with two children. Spanish-speaking respondents were interviewed by a bilingual interviewer using a Spanish translation of the questionnaire.

Response Rates and Selected Results (Texas)

The total number of households interviewed in Texas was 1,265. These interviewed households yielded completed interviews for 686 children aged 0–5 tears and 1,323 children aged 6–17 years. The interview completion rate among households was 88%. The American Association for Public Opinion Research recommends that final response rates for surveys that involve screening take into account the resolution rate, the screener completion rate, and the interview completion rate. The final response rate was calculated as 70% in Texas—slightly higher than the rates for the Health Module.

For producing population-based estimates of totals and percentages, a survey weight was attached to each sample person using the same weighting process as for the Health Module. Likewise, weighted estimates and estimates of standard errors were obtained using SUDAAN (Shah et al., 1997). Because this module did not include many questions from the NHIS, the reliability of the data was assessed by comparing data for children living in lower-income families with data for children living in higher-income families (see Table 1). For example, compared with children in higher-income households, children in lower-income families were more likely to be living with only one parent (16.6% vs. 37.6%). Income levels were also related to important measures of child well-being. For example, nearly twice as many children in lower-income families lived with aggravated parents (11.5%), compared with children in higher-income families (6.0%). This difference is particularly important because high stress and aggravation among parents are associated with poor cognitive, social, and emotional development in children. Finally, as might be expected, the proportion of children without any type of health insurance coverage was substantially higher for those children living in lower-income households (36.4%) compared with those residing in higher-income households (5.0%).

Study Design (Minnesota)

Implementation of the CWBW Module was recently completed in Minnesota, where it focused strictly on families with children insured by means-tested state-based health coverage (e.g., Medicaid and MinnesotaCare). In addition to providing well-being data for publicly insured children, this survey also provides information on how accurately respondents report their children's Medicaid coverage. A similar analysis of the accuracy of Medicaid reporting was undertaken in Texas by linking the pilot-test data to state-maintained administrative databases. Initial results from these two special SLAITS Medicaid evaluation studies are presented elsewhere in these proceedings (Blumberg & Cynamon, 1999). Though additional analyses are necessary, the information will be used to further clarify the issues related to erroneous reporting of health care coverage and other program participation. These results will likely be of interest to many health care researchers and could be adapted for other studies undertaking surveys of the low-income population.

Future SLAITS Activities

SLAITS is in the process of developing additional new modules. For example, a Children's Health Insurance and Health Care Module has been developed for use by states to plan programs and monitor progress toward increasing health insurance coverage and improving access to care, as required in the recently enacted Children's Health Insurance Initiative (Title XXI of the Social Security Act). The current draft of this module draws on questions from the NHIS, the Consumer Assessment of Health Plans, and the Medical Expenditure Panel Survey conducted by the Agency for Health Care Policy Research. Questions focus on health insurance coverage, access to care, use of preventive health services (e.g., well-child care, dental screening), children's health status, and unmet needs. Additional indicators of health care quality and satisfaction are also included. Pending funding, this module will be tested in one or more states.

Additional SLAITS questionnaire modules are in the development phase. The national survey of Families with Young Children is being planned for early next year in collaboration with the American Academy of Pediatrics and the UCLA Center for Healthier Children, Families, and Communities as part of a larger project sponsored by the Gerber Foundation. The overall design is to have companion surveys

Table 1. Selected measures of well-being and health insurance coverage by poverty status in Texas, 1999

	200% FPL,[1] % (SE)	> 200% FPL,[1] % (SE)	All Income Levels, % (SE)
Family structure			
Living with no parents[2]	3.7 (0.76)	2.2 (0.58)	2.9 (0.47)
Living with one parent[3]	37.6 (2.14)	16.6 (1.74)	27.7 (1.39)
Parents never married to each other	24.6 (1.93)	5.6 (0.81)	16.0 (1.12)
Child well-being			
Read Stories 3+ times per week (age 1–5)	73.8 (2.93)	90.9 (2.75)	80.7 (2.00)
Participated in extracurricular activities in last year (age 6–17)	54.4 (2.82)	76.9 (2.38)	64.9 (1.82)
Live with aggravated parents	11.5 (1.50)	6.0 (1.04)	9.6 (0.95)
Health insurance[4]			
Private	35.4 (2.22)	88.1 (1.46)	59.5 (1.54)
Public	28.9 (1.93)	8.9 (1.26)	19.6 (1.18)
Other	3.4 (0.66)	5.1 (0.92)	4.1 (0.53)
Uninsured	36.4 (2.15)	5.0 (0.99)	22.0 (1.31)

[1]FPL = federal poverty level
[2]Includes biological, foster, step, and adoptive parents, but not other guardians.
[3]One-parent families may include unmarried partners.
[4]Columns may not sum to 100% because respondents could report more than one type of insurance coverage.

of pediatricians and parents of children under 3 years of age to provide data on critical contemporary questions surrounding the health and development of children. Some of the topics to be addressed include how pediatricians deal with developmental issues, parents' concerns about developmental issues, the primary stresses and concerns of parenting, how pediatricians address parenting issues, access to and satisfaction with the health system, and how health care services are coordinated.

Finally, a module is being developed to assess health care and health care access for children with special health care needs, to be conducted in collaboration with the Maternal and Child Health Bureau of the Health Resources and Services Administration, DHHS. This survey will be conducted in 50 states and the District of Columbia beginning in July 2000. Its focus will be to provide baseline estimates for federal and state performance measures and the year 2010 national prevention objectives, and data for each state's five-year needs assessments for Title V of the Social Security Act. In each state, 5,000 households with children will be screened for special health care needs. The estimated 750 children per state that screen in as having special health care needs will be administered a questionnaire that covers functional status, health insurance, adequacy of health care coverage, access to care, care coordination and satisfaction, and family impact.

In conclusion, SLAITS has the capacity to grow into a broad-based ongoing surveillance system at the state and local levels to track and monitor the health and well-being of both adults and children. The SLAITS survey mechanism and questionnaire modules have been designed to complement the content of existing national and state surveys and systems. The SLAITS survey mechanism also may be used in the future to supplement the NHIS by targeting specific populations at the state level. Its unique ability to collect comprehensive data on specific health and welfare-related topics and for specific at-risk subdomains of the population make it extremely useful for informing public policy. This flexibility provides health policymakers with many analytic possibilities to track and monitor specific state health and welfare programs and to evaluate emerging public policy issues at the state and local levels.

References

Blumberg, S. J., & Cynamon, M. L. (1999). Misreporting Medicaid enrollment: Results of three studies linking telephone surveys to state administrative records. In *Proceedings of the 7th Conference on Health Survey Research Methods.*

Council of American Survey Research Organizations. (1982). Special report: On the definition of response rates (report of the CASRO Completion Rates Task Force). Unpublished manuscript, Audits and Surveys Co., Inc., New York.

Ezzati-Rice, T. M., Cohen S. B., Khare, M., & Moriarity, C. L. (1998). Using the National Health Interview Survey as a sampling frame for other health-related surveys. In *1998 Proceedings of the Section on Survey Research Methods* (pp. 121–129). Alexandria, VA: American Statistical Association.

Ezzati-Rice, T. M., Zell, E. R., Battaglia, M. P., Ching, P., & Wright, R. A. (1995). The design of the National Immunization Survey. In *1995 Proceedings of the Section on Survey Research Methods* (pp. 668–672). Alexandria, VA: American Statistical Association.

Frankel, M. R., Ezzati-Rice, T. M., Wright, R. A., & Srinath, K. P. (1998). Use of data on interruptions in telephone service for noncoverage adjustment. In *1998 Proceedings of the Section on Survey Research Methods* (pp. 290–295). Alexandria, VA: American Statistical Association.

National Center for Health Statistics. (1999). National Health Interview Survey: Research for the 1995–2004 redesign (DHHS Publication No. [PHS] 99-1326). *Vital and Health Statistics, 2,* 126.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1997). *SUDAAN user's manual: Software for the statistical analysis of correlated data, release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

Westat Inc. (1999). Evaluation of the NHIS Capacity to serve as the sampling frame for the DHHS consolidated surveys. Task 9 report: Assessing the NHIS capacity to provide state-level estimates. Unpublished manuscript, National Center for Health Statistics, Hyattsville, MD.

Needs for State and Local Data of National Relevance

James M. Lepkowski

State and local area needs for health data often include many of the same topics as those developed for national purposes. The relative importance of the topics at the state and local levels are usually quite different from those at the national level. State and local data, as for national purposes, are needed to aid health policy formulation, to allocate resources across health programs and between health and other state funding needs, to evaluate existing programs, and to observe trends in health and health care. And, just as for national data, state and local health data are needed for domains of interest within the state or local area.

The need for survey systems that provide state-level estimates has been recognized and addressed outside health statistics for several decades. Agriculture and economics, for example, have long-standing data collection series conducted through federal-state cooperative agreements that provide state-level estimates. The development of a number of survey systems that provide state-level estimates for health status, health care coverage and utilization, immunization, and health behavior in the last decade is long overdue and widely welcomed by those formulating state and local area health policy.

The four health surveys described in the papers in this session have different strategies and methodologies, as well as common design elements, for obtaining health data for state and local areas. The Behavioral Risk Factor Surveillance System (BRFSS) is a set of 50 separate surveys coordinated by the Centers for Disease Control and Prevention (CDC) but administered by each state. The National Household Survey on Drug Abuse (NHSDA), the National Immunization Survey (NIS), and the State and Local Area Integrated Telephone Survey (SLAITS) are administered by a single central agency and have separate samples for each state as well as for many local (in this case, metropolitan) areas. The BRFSS, the NIS, and the SLAITS use telephone sampling methods to select telephone households, and face issues of covering non-telephone households. The NHSDA uses an area frame to cover the entire household population and supplemental frames to cover relatively small populations of particular interest in the field of substance use, the homeless, students in college dormitories, and civilians living in military installations. All four surveys use computer-assisted data collection, and one, the NHSDA, uses audio computer-assisted self-administered interviewing. Standard statistical estimation procedures are used to compensate for nonresponse by weighting class adjustments and noncoverage through post-stratification in the four surveys, although one employs hierarchical Bayes methods to obtain state-level estimates for 42 states with smaller populations.

These surveys all share important design goals: producing valid and reliable estimates for the entire country and for states (the same goals may be stated for the kind of local area estimates that are important in the NIS and NHSDA, and eventually SLAITS). A corollary to the latter state-level estimation goal is to produce reliable estimates of differences among states, as in state-to-state comparisons. As is evident from the different survey design features across surveys, these goals are in conflict. A useful device for thinking about the conflict between national and state estimation from the same survey is the notion of stratified sampling. Consider the states as separate strata, or in this case domains, for which separate estimates are to be produced. A standard approach in sampling methods is to use a proportional allocation of the sample across states for the goal to produce reliable national estimates. That is, allocate more of the sample to the most populous states, proportional to the size of the state. Thus, California would have the largest sample (approximately 10% of the total). A proportionate allocation of sample across states has been used in other health surveys, such as the National Health Interview Survey (NHIS), but such designs are very inefficient at yielding reliable state-level estimates for all but the largest states.

None of these designs employ proportionate allocation across the states. All have chosen to emphasize (although not exclusively) the goal of producing valid and reliable state-level estimates, and obtaining reliable state-to-state comparisons. Under this state-level goal, the best allocation of sample across states is, under equal variance and cost for each state, an equal allocation: each state should have the same sample size, regardless of the size of the state. Thus, California should have the same sample size as Wyoming. Equal variance and equal cost across states for many of the kinds of measurements made in these surveys is probably correct. For telephone interviewing, the cost per unit (i.e., completed household or person interview) should be very similar across states. For face-to-face interviewing, some states will be more costly per unit, such as Alaska or Wyoming, because of travel costs. Sampling theory indicates that they ought to be allocated smaller samples than the other states. However, sampling theory results are benign, penalizing only minimal departures from the optimal smaller allocation to such states.

The author is at the University of Michigan.

For the sake of simplicity, and out of a general sense of political fairness, sample sizes remain approximately equal despite anticipated higher costs in some states.

Two of these four surveys employ approximately equal allocation across the states (and local areas): the NIS and the linked SLAITS. Their primary goal thus appears to be to provide equally reliable estimates for each state and many local areas. The other two surveys depart from an equal allocation for a variety of reasons, even though they share the goal of comparing state-level estimates. The BRFSS departs from the equal allocation because strategically each state has been enlisted to collect the data. States choose to increase sample size in order to improve the reliability of their own estimates, and to provide substate estimates that could not be obtained otherwise from the minimum sample size required by the Behavioral Surveillance Branch at the CDC. Still, the underlying principle of approximately equal sample sizes across states is fundamental to the BRFSS design. The NHSDA departs from approximately equal allocation in order to produce more reliable national estimates. It began as a national survey and adapted state-level estimates as a secondary goal. The BRFSS, NIS, and SLAITS were designed initially for state and local area estimation. National estimates are a secondary goal.

Thus, differences in goals partly explain the reason for the different strategies these surveys employ to obtain state-level estimates. The BRFSS is a state-level survey in administration and design, and begins with a recommendation for approximately equal sample sizes across states. State resources and interest increase sample sizes for many of the states. The NIS and SLAITS are also state-level surveys in design but not administration. Central design has allocated approximately equal sample size across states and local areas. Because these three surveys are fundamentally state-level surveys, estimates are obtained for each state directly from sample results. The NHSDA is primarily a national survey, with departures from the best allocation for national estimates made in order to improve the precision of state-level estimates. The departure, though, is not substantial enough to allow the same kind of direct estimation of state-level rates and other statistics. Thus, the NHSDA will produce estimates directly from sample results for several states, while Bayesian hierarchial estimation methods will be used for other states and local areas.

While these general principles of sample allocation are sound, they are also limited to considerations of sampling variance. The standard results ignore important sources of nonsampling error. All four surveys face substantial problems with nonresponse, and three have concerns about covering the population adequately. Nonresponse can lead to bias in results, reducing the validity of survey findings, when respondents and nonrespondents differ systematically with respect to the measurements of interest. Similarly, noncoverage of some population groups may bias measurements of interest if the covered and noncovered populations differ.

The BRFSS, NIS, and SLAITS all have noncoverage concerns because of the use of telephone sampling methods for the household portion of the population. Residents of non-telephone households are known to differ from those in telephone households on many health characteristics, such as health behaviors (e.g., seatbelt use, cigarette smoking, and frequency of exercise) and immunization. Smaller noncoverage concerns arise in the use of list-assisted telephone sampling methods, which leave out residents of telephone households that appear to have more recently received telephone numbers. Leaving out non-telephone households yields biased estimates of rates of health behavior or immunization levels not only for national but also for state and local area estimates. It is probably the case, however, that this noncoverage error is not as large in comparing national or state-level estimates from one time period to the next, but empirical study of the stability of noncoverage error is needed.

More importantly, noncoverage rates vary across states and local areas. Only 2% of the households in North Dakota, for example, do not have telephones, but 13% of the households in Mississippi are without telephones. National estimates will be biased through the "overrepresentation" of persons from states with low noncoverage rates. State-level comparisons will be confounded with differences in the telephone household coverage as well.

These three surveys employ a standard statistical estimation procedure (post-stratification) in order to partly compensate for noncoverage errors. By adjusting sample data to agree with national or state and local area population distributions on characteristics related to health behavior, health status, health care utilization, or immunization, survey designers attempt to reduce the difference between sample estimates (on average) and the population rate. However, post-stratification does not guarantee that sample estimates are free of bias due to noncoverage. There is no doubt that residual noncoverage bias remains across states and local areas.

While nonresponse is present in all four surveys, the potential nonresponse bias is greatest for the three surveys employing telephone sampling methods. Nonresponse rates in many states are extremely high (approaching 50%) in the BRFSS, and are likely to be nearly as high in the NIS and the SLAITS for some states (although state-level nonresponse rates are not reported). The NHSDA also has varying response rates across states. The variation in rates and associated potential bias across states has at least two detrimental effects on national, state, and local area goals. As for noncoverage, national data are difficult to interpret when states are disproportionately represented due to nonresponse, and state-to-state comparisons are confounded by differences in nonresponse bias. All four surveys employ statistical adjustment procedures to compensate for nonresponse. As for noncoverage, though, these methods are not expected to eliminate bias due to nonresponse.

The combination of noncoverage and nonresponse error in the three telephone surveys may be particularly problematic for state-level estimation. Although not universally true, states with high noncoverage rates generally have average or slightly below average nonresponse rates. On the other hand, many states with low noncoverage rates have high nonresponse rates. A few states have both. The quality of the data produced across these states will vary, with more valid esti-

mates expected for North Dakota (which has low noncoverage and nonresponse rates) than for California (which has very high nonresponse rates), despite the use of nonresponse and post-stratification weights.

Estimates from data combined across states, individual state-level estimates, and contrasts of state-level estimates are all threatened by nonresponse and noncoverage sources of error. National estimates are a goal of the BRFSS, NIS, and NHSDA. Such estimates will have smaller sampling error than state-level estimates because of the larger sample size, and test statistics as well as confidence intervals for national-level data will be based on those sampling errors. These nonsampling errors are difficult to incorporate into test statistics, but the error due to these sources is larger than the sampling error that is included in inferential statistics. Thus, conclusions drawn from national estimates are anticonservative, overstating confidence in the results.

Further, the BRFSS faces particular difficulties for national estimates. The nonsampling error properties of the individual state surveys are not well understood. There is substantial variation in coverage, response rates, and measurement across states, which raises serious concerns about combining data across BRFSS surveys. Some critics say that the data from the separate states should never be combined. Others recognize the need to account for different nonsampling error levels across states, and to develop methodologies for combining state estimates that have better error properties than the current estimates. It is also important to know the effective sample size of the BRFSS national sample. While the 1995 BRFSS state surveys have over 110,000 completed interviews, the effective sample size is bound to be smaller, especially given the already sizable design effects in several states.

On the other hand, state and local area estimates will, because of the smaller samples sizes allocated at the state or local area level, have larger sampling errors, although the nonresponse and noncoverage error will, for many state and local areas, be little different from that at the national level. In principle, at least, inferences based on state and local area estimates of sampling error will ignore less error than national estimates. Further, when states with similar levels of noncoverage and nonresponse error (and not just rates) are compared, one could speculate that these errors will tend to cancel one another. However, when the noncoverage and nonresponse errors are different between two states and fail to cancel one another in comparisons, it may become more difficult to find differences among the states.

These concerns emphasize the importance of understanding these sources of error and their contributions to errors in inferences about states and local areas. There are several tactics to address these concerns, only one of which is consistently described across these four surveys: compensatory weights for noncoverage and nonresponse. Statistical adjustment to survey results is relatively easy and readily implemented, and these surveys would be deficient without such adjustments. But the focus of the adjustments should be at the state or local level, in order to improve the quality of such estimates.

A second tactic is to seek ways to reduce the extent of such problems in survey design rather than in estimation. Noncoverage can be addressed through the use of supplemental frames. The NHSDA uses such frames to reduce noncoverage errors for important, although small, groups of the population left out (e.g., the homeless) or poorly covered (e.g., college students living in dormitories) in household sampling frames. The BRFSS, the NIS, and the SLAITS all have the potential to use the full household frame of the NHIS as a supplement to cover non-telephone households. The SLAITS explicitly mentions the use of dual-frame estimation procedures to integrate telephone household coverage of the SLAITS with non-telephone household coverage of the NHIS. Dual-frame methods have been developed for national survey estimation, including the important issue of altering allocations of sample size to address nonsampling errors such as nonresponse bias. Adaptation of dual-frame methods to state and local area estimates in the context of national surveys is needed. Here, synthetic estimates based on models will be needed for small states and for local areas where NHIS data on non-telephone households is limited. Methodological development in this area is incomplete at the present time.

A third tactic to address nonsampling errors is to devise methods that will reduce the extent of the problem. The evidence from the state-level data in the BRFSS, NIS, and SLAITS indicates that methods that reduce nonresponse rates in telephone surveys are urgently needed. Although telephone survey methods are desirable because of cost, nonresponse rates of nearly 50% for some states and 30% for national surveys pose serious threats to confidence in the findings of these surveys. While one can call for more resources to be directed to reducing nonresponse rates, through increased callbacks and incentives for nonresponding households, the problem is more fundamental than finding clever ideas or tricks to win household cooperation. Survey research does not fully address the cultural, psychological, social, and economic factors that drive nonresponse. The discussion on cross-cultural issues at this conference on Saturday illustrates how little survey researchers understand the cultural differences that contribute to nonresponse. Theoretical models of survey nonresponse have been developed, although those models need elaboration and further testing before they can be used to generate methodologies that will reduce nonresponse. State and local area data needs will not be adequately addressed in these current surveys unless such research is productive. States and local areas that need health survey data for policy development should be as concerned as national survey researchers about addressing these problems. State and local area surveys will be less useful in the future unless greater attention is paid to the problems of noncoverage and nonresponse reduction.

There are two additional issues of great importance to state and local area health surveys that are, by intent, not addressed in these papers. One issue is the content of health survey data needed for states and local areas. These four surveys address a number of important topics in health status, health care coverage and utilization, and health behavior. They point to national health goals in these areas. States and local areas are

also concerned with these national goals, especially when mandated to implement federal programs based on national goals. It is also the case that the states or local areas have other content areas that are higher in priority. For example, states and local areas may have greater concern about the health consequences of violence, food safety, the quality of care offered to different subgroups of the population, and consumer satisfaction with health care. In addition, just as national surveys are concerned with subnational estimates for important geographic areas or key subgroups, states have substate areas and subclasses for which estimates are needed.

There is an interesting contrast among these four surveys in how different content and substate estimates can be and are being handled. The BRFSS and the SLAITS allow and encourage variation in content across states. The SLAITS has the potential to give states the opportunity to mount telephone surveys on unique topics, as in the child well-being and welfare survey in Texas. The BRFSS partnership with state health departments provides greater opportunity for examination of content that is state specific. The BRFSS and the NIS specifically provide for substate estimates, either through state control of the sample allocation across state regions in the BRFSS or the targeting of local areas in the NIS. The SLAITS has the potential to assist states and local areas in targeting geographic, cultural, or demographic subgroups through the large sample sizes available in each state and in some local areas. It would be helpful to see, however, a discussion of the wider range of content and substate estimates that could be addressed through these surveys, and the development of methods to make it easier for states to participate in content specification and substate estimation.

Finally, these four surveys are nationally coordinated or administered. Many states and local areas are conducting their own surveys, no doubt with varying levels of sophistication and success. A compilation of such surveys may exist, although if it does it is not widely known. Either such a compilation needs to be more widely disseminated or needs to be developed. Simple description of the surveys and survey designs would be a useful reference for understanding how well states and local areas are handling their existing data needs. Individual state surveys can potentially be integrated with one or more of these nationally administered surveys to provide states and local areas with even better data for policy development. Further, and perhaps most important, federal expertise in survey implementation, and particularly knowledge about nonsampling errors and their control and reduction, needs to be disseminated to the states. Joint research to examine how nonsampling errors differ across states will be important to the successful long-run implementation of state and local area surveys. Methodological development aimed at improving state and local area surveys, whether by reducing nonresponse rates or making the content and operations culturally sensitive, can be conducted in concert with ongoing national survey methodology research.

Discussion Notes, Session 5

Donald Camburn and Arthur Hughes, Rapporteurs

The need to collect state and local area data was recognized as a legitimate and emerging requirement for surveillance and monitoring systems. The presentations in the session focused on a number of federal surveillance systems that currently provide, or will soon provide, state and local area estimates. These range from BRFSS, a state-level data system that is looking to provide national level estimates, to the National Immunization Survey, which since 1994 has provided vaccination coverage rates for all 50 states and 28 local immunization action plan areas.

The discussion, however, revealed a number of issues surrounding the systematic collection of state and local level data. These issues covered the following areas:

- Variation in quality of response rates, nonsampling error, interviewer effects, and differences in error structure
- Timeliness and extent of data releases
- Balancing national, state, and local needs and determining the appropriate locus of control
- How federal surveillance systems can provide within-state estimates for various demographic and geographic subgroups
- The most appropriate method for analyzing and reporting the data, in light of the large number of estimates that are being generated and the different estimation methods (e.g., direct versus model-based)

These areas are summarized separately in the following sections.

Variation in Data Quality

There were some concerns raised about the variation in response rates among states with surveys. One comment focused on differential nonresponse in large metropolitan areas, CMAs, rural areas, and so forth. Even when additional resources are allocated to combat the problem, it still remains a pervasive issue. There does not appear to be a good explanation for why it is happening, but it remains a problem and it is growing. For telephone studies such as the BFRSS, there are concerns about the impact of the degree of telephone infrastructure and differences in telephone use on data quality. California, with 10% of the U.S. population and only a 50% response rate, is different from many states in the large number of phone numbers per household. Also, with the high proportion of Hispanics and other minority populations, there may be cultural differences in the level of respondent cooperation. These differences have not been factored in and may present challenges in interpreting findings, particularly across states. From the NHSDA, response rates have varied somewhat across states. In 1999, North Dakota has had high response rates while Connecticut is on the other end of the spectrum. Even with additional resources used in states with high nonresponse, there has not been any considerable improvement in this problem. Another comment centered on the fact that nonsampling errors in general may not be equivalent across states. One suggestion was that, in telephone studies, there may be interviewer effects by state across time, which may lead to differential error structure across states. Estimating interviewer effects could help inform the analysis of house effects. In response to the concerns raised about the BRFSS, one comment was that steps have been taken to improve data quality through identification of problem states, to understand in some detail the nature of the problem, and to document them.

Timeliness of Data Release: What to Release and When

Issues were raised about the timeliness of data release for all of the studies presented in this session. For the NHSDA, where direct and model-based (empirical Bayes) estimation methods will be employed, the challenge is to release both micro and tabular data in a timely fashion. Issues such as producing a public-use file with state identifiers that may be limited in other ways (to minimize disclosure risk) may be a common occurrence. Also, while model-based estimates will be used in the 42 smaller states, over time users will have the ability to pool data across years and use direct estimation techniques in all states. In order to reduce possible misuse or misinterpretation of the data under these cirucumstances, an effort will be made to produce consistent data collection protocols across states and survey years; also, users and policymakers will be informed about the limitations of the data and the dangers of overinterpretation, particularly when multiple estimation methods are possible. While there may be some release issues during the first year of the expanded NHSDA sample, the plan is to publish national and state estimates approximately 8 months after the close-out of data collection. Sometimes choices have to be made between data quality and timelinesss of release. In order to get data out to meet

scheduled deadlines for publication of current state health profiles, BRFSS data were used instead of preferred data from the NHIS. In an effort to release state-level NHIS estimates in a more timely fashion, the survey changed to CAPI in 1997, which was one of serveral reasons why CAPI is now used. Currently, the plan is to release the state NHIS data approximately 6 to 9 months after the end of the data collection year. Another effort is to produce a select set of estimates for HP2010 and other important health monitoring systems much sooner (e.g., 4 months after data collection). The feeling is that once the 1995 state NHIS files are completed, future releases will be available a lot quicker. The 1995 file is the first state NHIS file being converted to public-use. The 1998 NIS data were released within 6 months of close-out. Future plans are to release data on a quarterly basis; also, release of a public-use file is planned at the end of 2000. Overall, the feeling among the presenters was that there is or will be great interest in their respective state estimates, and they are planning to do all they can to release the appropriate data in a timely manner, to document limitations, and to minimize micro-data disclosure risk.

Balancing National, State, and Local Needs

The federal surveillance systems covered in this session vary in their sponsorship. On one end of the spectrum, BRFSS is a decentralized collection of state-sponsored and state-run data collection systems that are implemented using CDC-produced guidelines and technical assistance. The state-level BRFSS data are submitted to CDC, and the results are tabulated and released in CDC publications. At the other end of the spectrum of sponsorship are the other surveillance systems covered in this session (NIS, SLAITS, and NHSDA), which use highly centralized data collection and analysis systems that are wholly sponsored by the federal government. Federal agencies also determine the methodology and content, and report the results.

One view expressed during discussion was that the control over content exercised by states in the decentralized model is an important feature that heightens the utility of the data to the sponsoring state. Because states use the BRFSS data, for example, to conduct internal evaluations, the need for centralization is less critical from the states' perspectives. To the extent that efforts to centralize control and standardize the methodologies reduce the autonomy of states, the utility of the data is diminished, since comparability across states is not a significant issue.

In contrast, the view was expressed that states are, indeed, very interested in comparing data with other states, especially adjoining states. While collecting high-quality data is an important goal, recognition needs to be given to the need to vary procedures across states to obtain high-quality data. Thus, as long as BRFSS produces good data within individual states, it provides data that can be used for within-state analysis, cross-state comparisons, as well as aggregated national estimates. It was noted that one advantage of the decentralized approach is that using local data collection contractors (e.g., local colleges or universities) may decrease nonresponse. Standardized, well-controlled methodologies may be an advantage in some circumstances but not others. CDC is taking steps to improve overall quality of BRFSS data and is enlisting researchers to identify those states where they are not comparable or are not at the same quality level as other states.

Within-State Estimates

One of the limitations of current state-level surveillance systems is that resource and time constraints restrict the amount of data that can be collected within individual states for smaller geographic domains or for demographic subgroups. This discussion echoed comments made during the discussion for session 2, Racial and Ethnic Populations: Cross-Cultural Considerations. For example, a question was raised about adding sexual orientation to the NHSDA, which would allow broad examination of risk factors for the gay population. Because of the multiple demands on the NHSDA, concerns about the effects of new questions on other data in the survey make this unlikely, given that the survey's primary purpose is to measure illicit drug use. While it is possible that such questions may be added to the National Health and Nutrition Examination Survey (NHANES), the sample size will be small and estimates will be available only at the national level. The National Health Interview Survey (NHIS) is not likely to add such questions. The State and Local Area Integrated Telephone Survey (SLAITS) is working with the HIV center at CDC to test using appropriate telephone interviewing technology to carry out a small-scale pilot study of 400 cases in order to examine the feasibility of collecting this kind of information at the state and local-area levels.

Analyzing and Reporting

An important issue for surveillance systems collecting data and providing estimates that cover a large number of areas is determining appropriate analysis methods and identifying appropriate methods for reporting the data that include area-specific data quality indicators. For example, in the National Household Survey on Drug Abuse (NHSDA), direct estimates are provided for the eight largest states while model-based estimates are calculated for the remaining states. In future years it will be possible to pool state-level data from direct-estimate states and compare them with pooled data for model-based states. Eventually, NHSDA plans to provide direct estimates for all 50 states. It will be important (and difficult), however, to provide policymakers with an appropriate understanding of these complex statistical issues and to caution them against overinterpreting the data. The discussion emphasized that it is important for state-level surveillance estimates to be used in conjunction with other data. The federal data are meant to complement data at the state and local levels, not replace them.

In addition, reporting data across a large number of geographic domains complicates the presentation of indicators of quality, such as response rates, precision estimates, and standard

errors. One observer noted that this has led to a reluctance to explain issues such as variability to the audience of data users.

It was noted during the discussion that while the presentations and discussion covered wide-ranging design issues, these data collection systems are ultimately designed to provide a substantial number of estimates and to provide estimates of change across time. The risk is that if careful analysis and interpretation of results are not carried out, spurious results may be reported. The CDC has appointed a task force that is investigating statistical methods and tools designed to identify aberrations in surveillance systems and changes in secular trends.

Although the session summaries in this volume cover specific research issues raised in the featured papers, discussant papers, and floor discussion, a few themes emerged from the proceedings that cut across several sessions. Three, in particular, seem especially worthy of mention as "conference" issues or themes that were raised periodically throughout these three days in Williamsburg:

- Protection of human subjects versus high response rates and data quality
- Letting subgroups "speak for themselves"
- Standardization versus tailoring of survey methods and measures

Protection of Human Subjects

One persistent issue or theme is the growing tension between the legitimate and important needs to protect human subjects—especially vulnerable population subgroups—and achieving high participation rates and acceptable levels of data quality. This is hardly a new issue in health survey research, but with increased demands to gather more and more data—and more sensitive data—from children, adolescents, and other subgroups of the population regarded as vulnerable to potential harm, this tension has mounted significantly in recent years. Within the federal statistics system, this is often played out as a balance between meeting requirements from the Office of Management and Budget (OMB) for acceptable participation and response rates versus those of the NIH Office for Protection from Research Risks (OPRR)[1] and local Institutional Review Boards (IRBs) for fully informed consent and other privacy and human subject concerns.

The papers in Session 1 on active parental consent in school-based research and a national study of child welfare both highlighted the nature of this dilemma in contemporary health survey research. Discussants and other researchers attending the conference noted recent studies that required review and approval by as many as 250 separate IRBs, while others noted instances where significant design changes required by one or more IRBs raised concerns among OMB reviewers that such studies could achieve participation rates high enough to justify the respondent burden and costs of conducting the study. Moreover, OMB has recently taken a more active, direct interest in privacy, consent, and human subject issues, thereby further raising the potential for significant conflict between an increasing number of parties with oversight and review responsibilities for important health surveys.

Protecting the confidentiality of data and the privacy of research subjects is without question essential to the conduct of successful and ethical research. Nevertheless, imposing inflexible (and at times inconsistent) methods to achieve these important protections and applying the most conservative interpretation of regulations often serves to undermine survey procedures and the resulting data. This dilemma was explicitly addressed in the first recommendation of the NAS Committee on National Statistics' Panel on Confidentiality and Data Access that "federal statistical agencies should follow a flexible, multilayered approach to informing data providers of the conditions under which they are being asked to provide information," but that all of the elements of informed consent should be provided.[2] The recommendations of the Panel uphold the spirit and letter of the law (and the ethical treatment of subjects) while directly highlighting the need to adapt our methods to the formidable, evolving challenges of conducting useful, relevant, high-quality research, an issue that is clearly fundamental (though not unique) to the health survey research enterprise.

Survey Participants "Speak for Themselves"

A second, related theme is a significant and growing need to reach children, adolescents, and various racial, ethnic, and minority groups and let them "speak for themselves" as survey participants and respondents. For example, several conference participants emphasized that such groups have a "right" to represent themselves in surveys, and current NIH guidelines require inclusion of ethnic minorities in all human subjects research. Mounting evidence indicates that this can in fact be done, is not harmful, and often provides quite different results from those observed when proxies provide data for these important subpopulations.

Several significant ethical considerations were raised in Session 2, and the examples and issues described there appeared and reappeared throughout the conference. The complex interplay of cultural, racial, ethnic, and psychological dynamics that permeate and influence the research enterprise among various ethnic and minority populations (in particular) must be recognized and attended to if effective and valid research is to be done.

[1]Recently elevated to the HHS department level as the Office for Human Research Protections.

[2]Duncan, George T., Jabine, Thomas B., and De Wolf, Virginia A., eds. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Panel on Confidentiality and Data Access [of the] Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council, and the Social Science Research Council. National Academy Press, 1993.

Standardization versus Tailoring

An additional complexity that is germane to allowing various population subgroups to "speak for themselves" reflects a debate in survey research on using standardized methods and measures for all survey respondents versus adapting or tailoring methods and measures to the characteristics of specific population subgroups. Those who promote standardized measures believe that it is of utmost importance that surveys use established, uniform measures to allow comparisons between surveys. Those who lean toward adapting techniques to the populations of interest give higher regard to the impact that diversity has on survey findings. Although differences in language and meaning are obvious examples, and the use of translation and appropriate translation techniques is now well established in health survey research, the more general issue is the extent to which traditional survey methods "work" with minority populations. Indeed, the appropriateness of some traditional methods, such as asking the same questions in exactly the same way of all respondents, has recently been questioned as a general strategy for all survey respondents.

With generally declining response rates and an increasingly diverse population, a one-size-fits-all approach to gaining the cooperation of subjects must be examined. More than ever before, interviewers must become proficient at identifying issues likely to affect decisions of specific potential participants and respond to their concerns. For example, anecdotal evidence from the ongoing National Health and Nutrition Examination Survey (NHANES) indicated a reluctance on the part of better-educated African Americans to participate in the survey, with specific mention of the government-sponsored Tuskegee project. NHANES staff responded by preparing a pamphlet that interviewers provide to survey respondents should Tuskegee be mentioned. The pamphlet underwent extensive testing with the target audience to achieve appropriate factual content, tone, and language.

Achieving a balance between consistency and change is critical to data quality. The need to make and justify choices has always been a significant methodological issue in health surveys, but the changing distribution and increasing diversity of our population is quickly rendering this issue one of the most fundamental methodological challenges that face us as we enter a new millennium.

Lu Ann Aday
University of Texas at Houston
School of Public Health
P.O. Box 20186
Houston, TX 77225
Phone: 713-500-9177
Fax: 713-500-9171
Email: laday@sph.uth.tmc.edu

Barbara Bailar
NORC
1155 E. 60th Street
Chicago, IL 60637
Phone: 773-256-6070
Fax: 773-753-7540
Email: bailar-b@norcmail.uchicago.edu

Peggy Barker
5600 Fishers Lane
Rockville, MD 20857
Phone: 301-443-4404
Fax: 301-443-9847
Email: pbarker@samhsa.gov

Mike Battaglia
Abt Associates, Inc.
55 Wheeler Street
Cambridge, MA 02138
Phone: 617-349-2425
Fax: 617-349-2605
Email: mike_battaglia@abtassoc.com

Bob Belli
Survey Research Center-ISR
University of Michigan
426 Thompson Street
Ann Arbor, MI 48104
Phone: 734-763-6020
Fax: 734-764-8263
Email: bbelli@umich.edu

Sandra Berry
RAND
1700 Main Street
Santa Monica, CA 90401
Phone: 310-451-7051
Fax: 310-451-6921
Email: sandra_berry@rand.org

Paul Biemer
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: 919-541-6056
Fax: 919-541-1261
Email: ppb@rti.org

Stephen Blumberg
National Center for Health Statistics
6525 Belcrest Road, Room 850
Hyattsville, MD 20782
Phone: 301-458-4107
Fax: 301-458-4035
Email: swb5@cdc.gov

Norman Bradburn
NORC
1155 E. 60th Street
Chicago, IL 60637-2799
Phone: 773-702-1066
Fax: 773-753-7886
Email: bradburn@norcmail.uchicago.edu

Mike Brick
Westat
1650 Research Blvd.
Rockville, MD 20850-3129
Phone: 301-294-2004
Fax: 301-294-2034
Email: brickm1@westat.com

E. Richard Brown
Center for Health Policy Research
UCLA School of Public Health
Box 951772
Los Angeles, CA 90095-1772
Phone: 310-825-5491
Fax: 310-825-5960
Email: erbrown@ucla.edu

Diane Burkom
Battelle Centers for Public Health Research and Evaluation
6115 Falls Road, Second Floor
Baltimore, MD 21209
Phone: 410-372-2702
Fax: 410-377-6802
Email: burkom@battelle.org

Cathy Burt
National Center for Health Care Statistics
6525 Belcrest Road, Room 952
Hyattsville, MD 20782
Phone: 301-458-4126
Fax: 601-458-4032
Email: cwb@cdc.gov

Don Camburn
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: 919-541-6696
Fax: 919-541-7198
Email: camburn@rti.org

Dick Campbell
Health Research and Policy Centers
University of Illinois at Chicago
850 W. Jackson, Suite 400
Chicago, IL 60607
Phone: 312-413-0480
Fax: 312-996-2703
Email: dcamp@uic.edu

Charles Cannell
13 Heatheridge
Ann Arbor, MI 48104
Phone: 734-769-0111
Fax: 734-764-8263

Joseph Catania
Dept. of Medicine
University of California, San Francisco
74 New Montgomery, Suite 600
San Francisco, CA 94105
Phone: 415-597-9161
Fax: 415-597-9395
Email: jcatania@psg.ucsf.edu

Grace Chang
Harvard Medical School
Brigham and Women's Hospital
75 Francis Street
Boston, MA 02115
Phone: 617-732-6775
Fax: 617-738-8703
Email: gchang@partners.org

Anne Ciemnecki
Mathematica Policy Research
P.O. Box 2393
Princeton, NJ 08543-2393
Phone: 609-799-3535
Fax: 609-799-0005
Email: aciemnecki@mathematica-mpr.com

Steve Cohen
Agency for Healthcare Research and Quality
2101 E. Jefferson St.
Rockville, MD 20852-4908
Phone: 301-594-6171
Fax: 301-594-2166
Email: scohen@ahrq.gov

Marcie Cynamon
NCHS
6525 Belcrest Road
Hyattsville, MD 20782
Phone: 301-458-4174
Fax: 301-458-4025
Email: MCynamon@cdc.gov

Chuck Darby
AHCPR
2101 E. Jefferson St., #502
Rockville, MD 20852
Phone: 301-594-2049
Fax: 301-594-2155
Email: cdarby@ahrq.gov

Terry DeMaio
Statistical Research Division
U.S. Bureau of the Census
Room 3000, FOB4
Washington, DC 20233
Phone: 301-457-4894
Fax: 301-457-4931
Email: Theresa_J_DeMaio@ccmail.census.gov

Kevin Dodd
National Institutes of Health
National Cancer Institute
6130 Executive Blvd. Suite 4103, MSC 7359
Bethesda, MD 20892-7359
Phone: 301-435-1834
Fax: 301-435-2046
Email: doddk@mail.nih.gov

Karen Donelan
Harvard Opinion Research Program
Harvard School of Public Health
677 Huntington Avenue
Boston, MA 02115
Phone: 617-432-3829
Fax: 617-432-0092
Email: kdonelan@hsph.harvard.edu

Brenda Edwards
NCI
6130 Executive Blvd., EPN 343
Bethesda, MD 20892-7350
Phone: 301-496-8506
Fax: 301-496-9949
Email: be2w@nih.gov

Michael Fendrich
Dept. of Psychiatry
University of Illinois at Chicago
423 IJR, MC-747
Phone: 312-413-1084
Fax: 312-413-1036
Email: fendrich@uic.edu

Jorge Ferrer
Division of Beneficiary Analysis
HCFA
7500 Security Blvd.
Baltimore, MD 21244
Phone: 410-786-9357
Fax: 410-786-8004
Email: dmiranda@hcfa.gov

Floyd J. Fowler, Jr.
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125-3393
Phone: 617-287-7200
Fax: 617-287-7210
Email: csr@umb.edu

Marty Frankel
Baruch College, CUNY
14 Patricia Lane
Cos Cob, CT 06807
Phone: 203-869-1324
Fax: 203-661-7456
Email: martin_frankel@abtassoc.com

Patricia M. Gallagher
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
Phone: 617-287-7200
Fax: 617-287-7210
Email: Patricia.Gallagher@umb.edu

Joe Gfroerer
SAMHSA
5600 Fishers Lane, Room 16-105
Rockville, MD 20857
Phone: 301-443-7977
Fax: 301-443-9847
Email: jgfroere@samhsa.gov

Ingrid Graf
Survey Research Laboratory
University of Illinois at Chicago
412 S. Peoria, 6th floor, M/C 336
Chicago, IL 60607
Phone: 312-413-0492
Fax: 312-996-3358
Email: ingridg@srl.uic.edu

Bob Groves
Survey Research Center
University of Michigan
426 Thompson Street
Ann Arbor, MI 48109
Phone: 734-763-2359
Fax: 301-314-7912
Email: bgroves@survey.umd.edu

Ann Hardy
6525 Belcrest Road, Room 850
Hyattsville, MD 20782
Phone: 301-458-4257
Fax: 301-458-4035
Email: amh1@cdc.gov

Rachel Harter
NORC
55 E. Monroe, Suite 4800
Chicago, IL 60606
Phone: 312-759-4058
Fax: 312-759-4090
Email: harter@norcmail.uchicago.edu

Jennifer Hawes-Dawson
RAND
1700 Main Street
Santa Monica, CA 90401
Phone: 310-393-0411x7238
Fax: 310-451-6921
Email: hawes@rand.org

Kris Hertenstein
Survey Research Laboratory
University of Illinois
909 W. Oregon, Suite 300
Urbana, IL 61801
Phone: 217-333-4273
Fax: 217-244-4408
Email: krish@srl.uic.edu

Mike Hilton
NIAAA
6000 Executive Blvd., Suite 505
Bethesda, MD 20892-7003
Phone: 301-443-8753
Fax: 301-443-8774
Email: mhilton@willco.niaaa.nih.gov

Rebecca Hines
HRSA
5600 Fishers Lane, 10A-30
Rockville, MD 20857
Phone: 301-443-6439
Fax: 301-443-4414
Email: rhines@hrsa.gov

Art Hughes
National Institute on Drug Abuse
Division of Epidemiology and Prevention Research
6001 Executive Blvd., Room 5153
Bethesda, MD 20892-9589
Phone: 301-402-1817
Fax: 301-443-2636
Email: ah62b@nih.gov

Tim Johnson
Survey Research Laboratory
412 S. Peoria, Sixth Floor
Chicago, IL 60607
Phone: 312-996-5308
Fax: 312-996-3358
Email: timj@srl.uic.edu

Graham Kalton
Westat
1650 Research Blvd.
Rockville, MD 20850-3129
Phone: 301-251-8253
Fax: 301-294-2034
Email: kaltong1@westat.com

Jon Klein
University of Rochester Medical Center
Adolescent Medicine Research Group
601 Elmwood Drive, Box 690
Rochester, NY 14642
Phone: 716-275-7760
Fax: 716-242-9733
Email: jonathan_klein@urmc.rochester.edu

Beth Kosiak
Health Care Financing Administration
7500 Security Blvd., MS S3-24-13
Baltimore, MD 21244
Phone: 410-786-1035
Fax: 410-786-8004
Email: bkosiak@hcfa.gov

Mary Grace Kovar
NORC
1350 Connecticut Ave., NW, Suite 500
Washington, DC 20036
Phone: 202-223-6040
Fax: 202-223-6104
Email: kovar@norcmail.uchicago.edu

Jon Krosnick
Dept. of Psychology
Ohio State University
1885 Neil Avenue
Columbus, OH 43210
Phone: 614-292-3496
Fax: 614-292-5601
Email: krosnick@osu.edu

Dick Kulka
Research Triangle Institute
3040 Cornwallis Road
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: 919-541-7008
Fax: 919-541-7004
Email: rak@rti.org

Jim Lepkowski
ISR-University of Michigan
426 Thompson Street
P.O. Box 1248
Ann Arbor, MI 48106-1248
Phone: 734-763-2359
Fax: 734-764-8263
Email: jimlep@isr.umich.edu

Judith Lessler
Research Triangle Institute
3040 Cornwallis Road
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: 919-541-6631
Fax: 919-541-7004
Email: lessler@rti.org

Jennifer Madans
National Center for Health Statistics
6525 Belcrest Road, Room 1140
Hyattsville, MD 20782
Phone: 301-458-4500
Fax: 301-458-4020
Email: jhm4@cdc.gov

David Maglott
HRSA
Maternal and Child Health Bureau
5600 Fishers Lane, 11A-22
Rockville, MD 20857
Phone: 301-443-0889
Fax: 301-480-2694
Email: dmaglott@hrsa.gov

Diane Makuc
Div. of Health and Utilization Analysis
National Center for Health Statistics
6525 Belcrest Road, Room 790
Hyattsville, MD 20782
Phone: 301-458-4360
Fax: 301-458-4037
Email: dmakuc@cdc.gov

Katherine Marconi
HRSA
HIV/AIDS Bureau
5600 Fishers Lane, 7A-07
Rockville, MD 20857
Phone: 301-443-6560
Fax: 301-594-2511
Email: kmarconi@hrsa.gov

Vickie Mays
Institute for Social Science Research
University of California at Los Angeles
405 Hilgard Ave, 1285 Franz Hall, Box 951563
Los Angeles, CA 90095-1563
Phone: 310-206-5159
Fax: 310-206-5895
Email: mays@ucla.edu

Lorraine Midanik
School of Social Welfare
University of California, Berkeley
120 Haviland Hall
Berkeley, CA 94720-7400
Phone: 510-642-7974
Fax: 510-643-6126
Email: lmidanik@uclink4.berkeley.edu

Lizza Miller
8835 Paisley Place NE, Second Floor
Seattle, WA 98115
Phone: 206-526-9985
Fax: 206-526-8920
Email: lizza@datstat.com

Danna Moore
Washington State University
Social and Economic Sciences Research Center
Wilson Hall 133
Pullman, WA 99164-4014
Phone: 509-335-1511
Fax: 509-335-0116
Email: moored@wsu.edu

Leo Morales
UCLA and RAND
1700 Main Street
Santa Monica, CA 90401
Phone: 310-794-2296
Fax: 310-794-0722
Email: morales@rand.org

Colm O'Muircheartaigh
Harris School
University of Chicago
1155 E. 60th St.
Chicago, IL 60637
Phone: 312-759-4017 (NORC),
773-702-8404 (UC)
Fax: 773-702-0926
Email: colm@uchicago.edu, colm@norcmail.uchicago.edu

Diane O'Rourke
Survey Research Laboratory
University of Illinois
909 W. Oregon, Suite 300
Urbana, IL 61801
Phone: 217-333-7170
Fax: 217-244-4408
Email: dianeo@srl.uic.edu

Linda Owens
Survey Research Laboratory
909 W. Oregon, Suite 300
Urbana, IL 61801
Phone: 217-333-4422
Fax: 217-244-4408
Email: lindao@srl.uic.edu

Elsie Pamuk
Office of Analysis, Epidemiology, and Health Promotion
National Center for Health Statistics
6525 Belcrest Road
Hyattsville, MD 20782
Phone: 301-458-4414
Fax: 301-458-4038
Email: epamuk@cdc.gov

Joanne Pascale
U.S. Bureau of the Census
Statistical Research Division
Federal Building 4, Room 3134
Washington, DC 20233
Phone: 301-457-4920
Fax: 301-457-4931
Email: Joanne.Pascale@ccmail.census.gov

Deb Potter
Agency for Healthcare Research and Quality
2101 E. Jefferson St.
Rockville, MD 20852-4908
Phone: 301-594-1061
Fax: 301-594-2166
Email: dpotter@ahrq.gov

Eve Powell-Griner
Centers for Disease Control
Behavioral Surveillance Branch
Mailstop K47
Atlanta, GA 30341-3724
Phone: 770-488-2524
Fax: 770-488-8150
Email: eep1@cdc.gov

Kenneth A. Rasinski
NORC
1155 E. 60th Street
Chicago, IL 60637-2799
Phone: 773-256-6278
Fax: 773-753-7886
Email: rasinski@norcmail.uchicago.edu

Anne Riley
Johns Hopkins School of Public Health
624 N. Broadway
Baltimore, MD 21205
Phone: 410-955-1058
Fax: 410-614-7189
Email: ariley@jhsph.edu

Todd Rockwood
Div. of Health Services Research and Policy
University of Minnesota
Box 729, 420 Delaware St., SE
Minneapolis, MN 55455-0381
Phone: 612-625-3993
Fax: 612-624-2196
Email: rockw001@tc.umn.edu

Bill Rodgers
Institute for Social Research
University of Michigan
426 Thompson Street, P.O. Box 1248
Ann Arbor, MI 48106-1248
Phone: 734-763-6623
Fax: 734-647-1186
Email: wrodgers@umich.edu

Rob Santos
The Urban Institute
2100 M Street, N.W.
Washington, DC 20037
Phone: 202-261-5291
Fax: 202-429-0687
Email: rsantos@ui.urban.org

Cathy Schoen
The Commonwealth Fund
One East 75th Street
New York, NY 10021-2692
Phone: 212-606-3800
Fax: 212-606-3500
Email: cs@cmwf.org

Kate Scott
New Zealand Ministry of Health
133 Molesworth Street
P.O. Box 5013
Wellington, New Zealand
Phone: 04-496-2000
Fax: 04-496-2340
Email: kate_scott@moh.govt.nz

Monroe Sirken
National Center for Health Statistics
6525 Belcrest Road, Suite 700
Hyattsville, MD 20782
Phone: 301-458-4505
Fax: 301-458-4039
Email: mgs2@cdc.gov

Phil Smith
Centers for Disease Control and Prevention
National Immunization Program
MS E-62, 1600 Clifton Rd. NE
Atlanta, GA 30333
Phone: 404-639-8729
Fax: 404-639-8613
Email: pzs6@cdc.gov

Christina Smith Ritter
Division of Beneficiary Analysis
HCFA
7500 Security Blvd.
Baltimore, MD 21244
Email: critter@hcfa.gov

Richard Strouse
Mathematica Policy Research, Inc.
600 Alexander Park, P.O. Box 2393
Princeton, NJ 08543-2393
Phone: 609-275-2332
Fax: 609-799-0005
Email: rstrouse@mathematica-mpr.com

Seymour Sudman
Survey Research Laboratory
University of Illinois at Urbana-Champaign
909 W. Oregon, Suite 300
Urbana, IL 61801
(Deceased May 2000)

Michelle van Ryn
Dept. of Health Policy, Management and Behavior
University of Albany–SUNY
One University Place
Rensselaer, NY 12144-3456
Phone: 518-402-0293
Fax: 518-402-0414
Email: mv479@csc.albany.edu

Dick Warnecke
Health Policy Center
University of Illinois at Chicago
850 W. Jackson, MC-275
Chicago, IL 60607
Phone: 312-355-1167
Fax: 312-413-9835
Email: warnecke@uic.edu

Robert Weech-Maldonado
Dept. of Health Policy and Administration
Pennsylvania State University
116 Henderson Building
University Park, PA 16801
Phone: 814-865-1926
Fax: 814-863-2905
Email: rxw25@psu.edu

Nicholas Zill
Westat
1650 Research Blvd.
Rockville, MD 20850-3129
Phone: 301-294-4448
Fax: 301-294-3992
Email: zilln1@westat.com

DEPARTMENT OF
HEALTH & HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics
6525 Belcrest Road
Hyattsville, Maryland 20782-2003

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

STANDARD MAIL (B)
POSTAGE AND FEES PAID
CDC/NCHS
PERMIT No. G-284

DHHS Publication No. (PHS) 01-1013
1-0123 (3/01)