Eighth Conference on

HEALTH SURVEY RESEARCH METHODS

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

Suggested citation:

Eighth Conference on Health Survey Research Methods.
Cohen SB, Lepkowski JM, eds. Hyattsville, MD: National
Center for Health Statistics. 2004.

Eighth Conference on

HEALTH SURVEY RESEARCH METHODS

Edited by
Steven B. Cohen and James M. Lepkowski

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

Hyattsville, Maryland
July 2004

DHHS Publication No. (PHS) 04-1013

DEDICATION

Charles F. Cannell (1913–2001) was a charter member of the group of researchers and research administrators who instituted the health survey research methods conference series. He attended each of the first six conferences and would have attended the seventh if illness had not prevented him from doing so. Charlie was a tireless supporter of research on health survey methodology, particularly the role of interviewers in the survey process. Charlie believed that valid data on the state of a society is essential to wise and humane policies. His gentle and kind nature, good humor, and thoughtful counsel will be missed in this series. This volume is dedicated to the memory of Charlie Cannell and his many contributions to the health survey research methodology conference.

CONTENTS

SESSION 5: SECURITY AND PRIVACY

FOREWORD Setting the Agenda for the *Eighth Conference on Health Survey Research Methods*

HEALTH SURVEY RESEARCH METHODS CONFERENCES

An instrumental component in the systematic study of health survey methodologies has been a series of occasional symposia and conferences. In the early 1970s, the National Center for Health Statistics (NCHS) and the National Center for Health Services Research (NCHSR, the predecessor to the Agency for Healthcare Research and Quality [AHRQ]) identified through a series of meetings and seminars a set of important health survey methodology issues that needed to be addressed through methodological research. Each had the potential to affect the quality of health survey data. In 1975, the first of a sequence of conferences examined critical methodological issues in health survey research. A small group of 40 researchers and federal health statistics system officials gathered to discuss methodological problems, set priorities for funding, and disseminate recommendations of the conference to a broad community of health researchers.

A second conference of 40 researchers was held in 1977. The conference participants concluded that a series of biennial conferences on health survey methods would add substantial value to the broader health research and policy committee. Additional conferences were held in 1979, 1982, 1988, 1995, and 1999. This volume is a report on the proceedings of the Eighth Conference held in 2004.

These conferences have shared features that are collectively distinctive. A steering committee of government and academic researchers develops conference topics and seeks prominent investigators to lead and stimulate discussion. Conferences employ formal papers and presentations but allow ample opportunity for discussion, both prepared and spontaneous. Participation is by invitation only to a limited number of participants. Invitations are extended to government researchers, survey data users, and academic methodological researchers actively engaged with health survey methods and data. Invited presentations are presented in plenary session, and rapporteurs develop summaries of floor discussion to accompany formal papers. Findings are disseminated through published proceedings volumes.

Conferences are supported by a variety of agencies and foundations. Early conferences were sponsored by the NCHSR (AHRQ today) and NCHS. Both agencies remain steady contributors to conference support, with more recent support provided by other agencies in the federal health and statistics system.

Themes across the eight conferences represent timely and enduring research problems faced by those conducting health surveys. For example, the first conference addressed questionnaire design, interviewers and interviewing techniques, validity of survey data, and total survey design. Subsequent conferences addressed a full range of survey methodology problems, examining such topics as nonresponse, respondent burden, standardized measurement, ethics, sample design and estimation, cross-cultural considerations, data collection mode, and state and local data needs from surveys.

EIGHTH CONFERENCE ON HEALTH SURVEY RESEARCH METHODS

The Eighth Conference was held February 20–23, 2004, in Peachtree City, Georgia. The steering committee met in person and by telephone conference several times before the conference to discuss potential themes. The steering committee

discussion led to a decision to organize sessions in a format that was a modest departure from more recent conferences. Each conference session was organized by a single member of the steering committee who had research interest in a conference theme. The steering committee assisted in identifying chairs, invited participants to review the current status of and future directions for the topical areas under consideration, and discussants. Contributed paper abstracts were subsequently solicited through a public announcement, and the committee selected one or more of the most relevant contributed papers for inclusion in the respective sessions.

The success of the Eighth Conference depended on the timeliness and enduring quality of the themes selected for the sessions. Steering committee discussions began with a review of the complex and challenging problems being faced by those conducting health surveys in the beginning of the 21st century. Discussions covered issues ranging from the increasingly diverse and dynamic nature of the U.S. population to changes in population response to survey requests for participation. Five themes emerged from these discussions:

- Capturing more effectively diversity and change in dynamic populations through longitudinal health surveys;
- Community participation in surveys and other types of research and the identification of community benefit from health surveys;
- The challenges of conducting health survey research in culturally diverse populations;
- Responding to rapidly decreasing cooperation levels in health surveys, whether in person, by telephone, or through self-administration, and the emergence of Internet technology and survey data collection; and
- The nature and impact of privacy concerns and survey confidentiality provisions on survey participation and the administrative procedures for conducting health surveys today.

The committee believed that these themes represented several of the most important challenges facing health survey research methods today and in the foreseeable future. Committee members and session organizers were particularly interested in providing a forum where the implications of current trends for the future of health survey research methodology could be discussed in the plenary sessions. These themes also complemented each other, providing an opportunity for researchers familiar with one theme to contribute to discussion on one or more of the other themes.

There have been important changes in major health policy issues that surveys have been used to address. The essential survey conditions faced by health survey methodologists have evolved as the population has changed and become increasingly diverse. Yet the methodological challenges continue to revolve around a common set of enduring themes that have been discussed at this and past conferences. Each conference provides the opportunity to select the most important methodological problems of the day and anticipate future health policy and survey methodology developments.

The first session, *Capturing Diversity and Change in a Dynamic Population*, provided a forum to address the utility of longitudinal studies as important vehicles for providing a better understanding of the factors associated with transitions in health care status, utilization and expenditure patterns, and health insurance coverage over time. In an increasingly diverse U.S. population, methods continue to be needed to improve sample selection in order to identify and follow populations with critical needs, from minority groups to those with conditions that are of particular interest to the health

care community. Attention also was given to supporting critical reflections of the value of longitudinal studies and providing examples of how to best execute them to obtain the data content needed by policy makers and health care researchers. The presentation of methods most appropriate for the analysis of time dependent data was another session objective. Consequently, the session was formalized to include a set of invited papers that focused on the following topics: longitudinal estimation in AHRQ's Medical Expenditure Panel Survey and its related analytical capacity; health findings from the British National Birth Cohort Study; design issues associated with the forthcoming National Children's Survey; and analytical strategies for longitudinal data from complex health surveys.

The second session on *Community Participation and Community Benefit* was designed to identify paths to help balance the tension between the needs of the community, the principles of scientific research, and prior practices of researchers and the institutions they represent. It was viewed as an essential thematic area for conference inclusion, particularly in light of the downward trend in response rates in national and local surveys. In surveys such as the National Health and Nutrition Examination Survey, methods have been emerging that recognize the importance of enhancing and conveying the community benefits of the research in influencing survey participation, within both geographically defined communities and broader communities of identity. Consequently, the goal for this session was to explore methods and means for balancing and enhancing both community participation and community benefit in the design and conduct of national, state, and local health surveys.

The third session, *Cross-Cultural Challenges in Health Survey Research*, was planned at the outset to include a mix of invited and contributed papers. Invited papers were to focus on the following topics: an overview of the problems in establishing conceptually equivalent definitions of health across multiple cultural groups, identification of methods for developing conceptually equivalent measures across multiple cultural groups, and coverage of methods for verifying conceptual equivalence of measures across multiple cultural groups.

In the fourth session, the committee desired to establish a big-picture view of where we are with respect to the conduct of general population health surveys in 2003. Hence, the session was aptly titled *How to Conduct Health Surveys in the 21st Century*. Attention was to be given to the declining feasibility of random-digit-dialing telephone surveys for producing credible data and the consideration of Web-based surveys and mail surveys as potential substitutes for or complementary components of dual-mode protocols. These operational considerations were to be further informed by a total survey design perspective, with featured presentations considering costs, the quality of sample frames, the rates and biases associated with nonresponse, and the issues of data quality and data comparability associated with alternative modes of collection data.

Finally, the issues of best practice that ensure the rights and welfare of survey participants are protected is a theme that has grown in importance in recent years. A session on *Security and Privacy* was included to provide a forum for presenters and participants to explore issues related to ethical research standards and informed consent. Emphasis was to be placed on the inclusion of presentations addressing the new challenges posed by the increasing demand for the collection of sensitive information, the inclusion of mature minors, the retention of biological samples, and the availability of regulated and unregulated data.

At the Eighth Conference's inception, we presented the participants with a set of

challenges. One was for participants to help frame the ensuing discussion. Another was to identify overarching themes, common problems, and potential solutions. More specifically, the presenters, discussants, and participants were asked to connect the topics addressed by the set of related papers in each session with recommended strategies to improve the quality of health surveys. In evaluating the effectiveness of new design features and methodological innovations, the following parameters were given particular attention: accuracy, relevance, timeliness, accessibility, clarity, and cost-efficiency.

In addition, to facilitate discussion that could identify and prioritize future efforts to improve the conduct and quality of health surveys, the participants were asked to frame their comments with the following considerations in mind:

- Anticipation of future needs for timely, accurate, and reliable policy-relevant data and best practices to satisfy demand;
- Identification of the greatest challenges faced by survey designers and researchers to provide high quality data in a cost efficient manner and best practices;
- Identification of strategies to improve communications among and between researchers; policy makers; survey designers (statisticians, methodologists); survey operations, field, management, and data processing staff; and
- Identification of future research priorities.

A total of 75 persons attended the Eighth Conference, including researchers from academic disciplines who conduct and use data from surveys, researchers and administrators from federal statistical system agencies responsible for major health surveys, and academic and government health policy researchers who use survey data to help formulate health policies. Almost one-half of the participants had not attended one of the previous conferences. Six had been present at the first conference, although they had not attended every conference. Thus, participants represented a wide range of previous connections to the series, allowing the new members to gain additional insights through interactions with the conference veterans.

In planning for the Eighth Conference, the steering committee members clearly desired to include a more visible representation of new participants who would share their fresh perspectives in dealing with the existing and new challenges faced by the field. Efforts also were made to ensure strong threads of continuity with the inclusion of the leadership from prior meetings, providing a great breadth of collective wisdom from which to draw upon. It appears that the steering committee's careful planning on this front has come to fruition.

Steven B. Cohen
Director, Center for Financing, Access and Cost Trends
Agency for Healthcare Research and Quality

Jim Lepkowski
Research Professor, Institute for Survey Research
University of Michigan

March 2004

ACKNOWLEDGMENTS

Conferences in the series have been possible only through the support of a number of committed federal agencies and individual survey research organizations. The largest share of funding for the Eighth Conference came from the Agency for Healthcare Research and Quality through a conference planning and implementation grant. The National Center for Health Statistics also supported the conference through a financial contribution and through publication services for this volume. Financial contributions also were made by Abt Associates, Inc., the American Cancer Society, the Health Resources Services Administration, the National Cancer Institute, the National Institute on Drug Abuse, the National Institutes of Health, the Substance Abuse and Mental Health Services Administration, and the Survey Research Center of the University of Michigan. We are very grateful for the support of all of these agencies.

The conference is planned and implementing by a steering committee composed of representatives from academic and research communities and the federal agencies that support the conference. The steering committee is a voluntary collaboration of federal researchers and policy specialists and private and academic survey researchers who meet periodically to develop the structure and content of the conferences. The current conference would not have been possible without the contributions of the individual members: Lu Anne Aday (University of Texas at Houston), James Colliver (National Institute on Drug Abuse), Marcie Cynamon (National Center for Health Statistics), William Davis (National Cancer Institute), Brad Edwards (Westat), Floyd J. Fowler, Jr. (University of Massachusetts Boston), Joseph Gfroerer (Substance Abuse and Mental Health Services Administration), Timothy Johnson (University of Illinois at Chicago), Alice Kroliczak (Health Resources and Services Administration), Richard Kulka (RTI International), Colm O'Muircheartaigh (University of Chicago), and Richard Warnecke (University of Illinois at Chicago). We thank each of them for their dedication and enthusiastic support for the conference.

Sessions during the conference were plenary, and all attendees participated in each session. The formal and informal discussions in each session were stimulating and provoking and reflected the genuine participation of all present at the meetings. We thank all the participants, those who presented papers, those who gave formal discussions, those who contributed in floor discussion, and those who served as rapporteurs at each session.

As has been the case at the last two conferences, Diane O'Rourke of the Survey Research Laboratory (SRL) at the University of Illinois was the most instrumental in bringing the conference together. Diane handled the organizational details with great skill and kept the co-chairs and steering committee on task and moving ahead, despite a number of schedule setbacks. She arranged the steering committee planning meeting and conference calls, posted the Call for Papers, collated and distributed abstracts for invited and contributed papers, selected the site, handled travel arrangements, oversaw the production of these proceedings, and much more. Somehow she managed to serve the needs of more than 70 people, making everyone happy just about all the time. The conference would not have happened without her care and superb management. Diane was assisted by SRL colleague Kris Hertenstein, who demonstrated grace and patience in the midst of more than one moment of panic when things didn't go just right. We thank both of you for making the conference not only possible

but also a success. We also would like to acknowledge Lisa Kelly-Wilson from SRL; she and Kris designed and produced the conference program and compiled, formatted, and edited these conference proceedings.

We as co-chairs have enjoyed working with so many wonderful and committed individuals who share an abiding interest in conducting health surveys better. It was a pleasure to be able to contribute to continuing this remarkable series of conferences.

Steve Cohen
Jim Lepkowski

INTRODUCTION TO SESSION 1: Capturing Diversity and Change in a Dynamic Population

Graham Kalton, Westat

Many surveys are single-time and cross-sectional, aiming to measure the characteristics of the surveyed population at the time the survey was conducted. However, there also is considerable analytic interest in examining changes over time, and there are various design strategies that can be used to examine such changes.

A key distinction to be made is between *net change* (change at the aggregate level) and *gross change* (change at the individual level). Measures of net change can be obtained from a *repeated survey* design in which the same cross-sectional survey is carried out at different points of time, with fresh samples at each time point. Examples in the health field include the National Health Interview Survey and the National Survey on Drug Use and Health (NSDUH). Note that net changes in estimates (for example, estimates of drug use) from a repeated survey reflect a combination of changing characteristics of the population and change in population composition (e.g., births, immigrants, deaths, and emigrants). A repeated survey design cannot provide measures of gross change unless retrospective questions are asked about sampled individuals' past characteristics or behaviors. The limitation of retrospective questioning is, of course, that respondents may have forgotten or misremembered the information sought. The paper in this session by Gfroerer et al. takes advantage of the fact that the NSDUH is a repeated survey that asks questions about both current and retrospective substance use. The authors thus are able to examine the ability of the retrospective questions to provide estimates of substance use for earlier times that are comparable with the corresponding cross-sectional survey estimates (with adjustments made for changes in population composition).

The fallibility of memory rules out retrospective questioning as a means of studying gross changes for most phenomena of interest in health surveys. As a result, some form of *panel survey* design is needed to study gross changes. One form of panel survey selects a sample from a specified cohort and follows the sample over time as, for example, is the case with the British 1946 Birth Cohort Study described in Wadsworth's paper in this session. The strength of the cohort design is its ability to identify the time ordering of the experiences of the sample members and hence analyze the relationships of earlier experiences with later health outcomes. The results from a cohort design strictly apply only to those whose life experiences coincide with the given time period and the conditions pertaining in that period (e.g., the early years for the British 1946 birth cohort were ones of post-war food rationing and no national health service). However, repeated cohort studies can address this limitation and provide valuable cross-cohort analyses. In Britain, similar national birth cohort studies were started in 1958 and 1970, and a millennium cohort recently has been introduced. A major U.S. birth cohort study—the National Children's Survey—currently is being planned with a focus on the effects of environmental influences on children's health and development. In his paper in this session, Correa describes the current state of the planning for this survey, which is expected to begin in full in late 2006 to early 2007 and will follow a sample of around 100,000 births through to early adulthood.

A different type of panel design selects a representative sample of the total population and follows that sample over time. This is the design used in Canada's National Population Health Survey, as described by Wolfson and Rowe in this session. Their paper demonstrates how longitudinal data from a panel survey are needed for the kinds of microsimulation modeling that they conduct to predict future numbers of frail elderly who lack the possibility of close family support.

A variant on this design is described by Ezzati-Rice and Cohen in their paper. The Medical Care Expenditure Survey (MEPS) also starts from a representative sample of the population but follows the sample for a limited duration of two years. With new panels being started every year, the design is what is termed a *rotating panel survey* design. An important consideration in choosing a panel design for the MEPS was to be able to aggregate medical costs over time, as distinct from measuring change. (This consideration also applies to the Survey of Income and Program Participation.) A rotating panel design affords a number of analytic possibilities, such as combining two or more panels in the analysis for the periods of overlap. Composite estimation, as used in the Current Population Survey, also may be applied to improve the precision of survey estimates by borrowing strength from data collected in other panels.

In summary, the papers in this session bring out the importance of the time dimension in health survey research. They illustrate various approaches to incorporating that dimension into a survey design and the strengths and limitations of alternative designs. Panel surveys present a number of methodological challenges, particularly in retaining sample members in the panel. However, panel designs provide the longitudinal data needed to examine the precursors of health outcomes and possible cause-effect relationships. They therefore have a great deal to offer to health survey research.

FEATURE PAPER: Life Course Health Research: The British 1946 National Birth Cohort Study

Michael Wadsworth, University College London

INTRODUCTION

The foremost strength of the life course design is its information on sequence and chronology. Its greatest dilemmas are how to sample appropriately at the outset for future scientific and policy requirements and how to know, ahead of time, what information to collect in order to study, at later times, an individual's age-related change.

This paper describes first these dilemmas and how they have been approached in a British national life course study that began at the birth of its sample members in 1946 and continues still. Then conclusions are drawn from the experience of this study about the design and value of future birth cohort studies.

THE BRITISH 1946 NATIONAL BIRTH COHORT STUDY

The Initial Study

Four questions of both scientific and policy relevance initiated this investigation (Wadsworth, 1991). Why had fertility fallen continuously over the previous 100 years? What was the cost of birth to the family? What was the availability and distribution of specialist obstetric care? What was the uptake of prenatal care? These were questions of relevance to planning a national health service, which began two years later.

Sampling had to represent all regions in England, Wales, and Scotland, and because health care professionals were to collect the data, the period of collection was concentrated into one week, and all babies born during that time were included in the sample (N=16,687). Community nurses collected information from medical records and through home interviews with mothers up to eight weeks after the birth. Nurses collected information on labour and the delivery from the whole sample, and information on costs of the birth and on prenatal dietary supplements was collected from a random half of the sample. Information was collected on about 82% of all births in the chosen week.

Sample Design for the Follow-Up

The follow-up could not include the entire original sample because of the costs and limitations of contemporary data handling methods. Therefore, a sample was taken, which was designed to maintain geographic representation and reduce sample size to one-third. Sample reduction was achieved by randomly selecting one in four of the largest SES group (the manual social class) and all of those in the smallest SES groups (the nonmanual and agricultural classes), resulting in a sample of 5,362. The regional distribution and the clinically unselected and representative nature of the sampled population each have proved of great value. Weighting is used to compensate for the effect of the sampling procedure.

Data Collection

During the sample members' preschool years, the scientific questions were concerned with mortality, growth, development, and morbidity and their social variation. Policy questions focused on the value of maintaining the national network of community nurses who provided (and still provide) clinical services to mothers in the early postnatal period (Figure 1). When sample members were age 2 years, and again at 4 years, community nurses extracted information from clinic records, measured the children, and interviewed mothers at home to collect information on family circumstances, and at 4 years, they collected information on the child's diet during the previous day.

During the school years, the health science questions remained essentially the same. Health policy questions were concerned with the distribution of ill health and the risks of exposure to atmospheric pollution from coal

Figure 1. Social and Policy Questions Addressed by the 1946 British Birth Cohort

Years	Cohort ages	National policy problems addressed
1946	Birth	Costs of maternity, reasons for falling fertility, distribution of obstetric patients, uptake of prenatal care.
1947–1950	1–4 years	SES differences in maternal and child mortality and morbidity. Value of community nurses' work.
1951–1961	6–15 years	Increasing the national level of educational attainment. The "waste of talent problem."
1962–1976	16–30 years	Outcomes of education in terms of occupational choice and skills. Delinquency.
1976–	30 years onwards	Aging processes, self care of health, receptivity to health promotion.

burning. Educational policy questions were about the efficacy of the new national system to select the most able children for entry at 11 years to schools that would prepare them for university entrance. Social policy questions were about delinquency and career and employment selection (Figure 1). These questions required new methods of data collection. When the children were ages 7, 8, 11, and 15, school physicians and nurses measured and examined them and asked mothers/caregivers about their health, and reports of hospital admission were checked with hospital records. National area-based information on atmospheric pollution from coal burning was used to assess each child's exposure. Teachers gave information about their schools, about communication with parents, and about the sampled children's attendance at ages 7, 10, 11, 13, and 15. Teachers also administered tests of cognitive function and educational attainment to sample members at ages 8, 11, and 15 (that is, before and after the nationally-administered tests at age 11). At ages 13 and 15, teachers assessed the children's temperament and behaviour. Results in national examinations were checked with awarding authorities. The frequency of data collection in this period was determined by rates of developmental change.

In the adult years, the study concentrates on health. Of scientific concern are the progression, precursors, and distribution of physical and cognitive aging, and secondarily, the precursors and distribution of mortality, morbidity, disability, and health-related behaviour. Health policy questions concern aging; use of health care services; dietary, smoking, and exercise habits; and alcohol consumption. Social science questions are about the precursors of income, employment, and fertility histories. Social policy questions are about the returns of education, in terms of employment, skills, and income (Figure 1). The study concentrates on specific topic areas (cardiovascular, respiratory, musculoskeletal, and mental health, as well as cognitive function). Health is measured primarily in terms of function (e.g., blood pressure, memory) and body shape and size, as well as morbidity. Case ascertainment is by clinically validated questions and examination of hospital records and death certificates. When sample members were age 53, a source of DNA was collected. Risk exposure has been measured by information on diet, smoking, reports of alcohol consumption, exercise habits, and home and work circumstances. Data is collected through at-home visits by research nurses trained for the study and by postal questionnaire (Kuh & Hardy, 2003).

Intervals between data collections have been longer in adulthood because functional aspects of health in middle life change fairly slowly, and cost has been a factor as well. In later life, the intervals between collections will decrease as events happen more often and memories become less reliable.

Response

Response was high during the preschool and school years (89%–95% of those alive, resident in Britain, and not refusals) because health and educational professionals collected the data. Also, since the study covered all regions, migrants within England, Scotland, and Wales were not lost to follow-up. Birthday cards, together with an annual check on

contact details and summaries of work and publications, help to maintain response, and a study Web site has been established (www.nshd.mrc.ac.uk). No incentives have been offered. The anticipated large increase in numbers of refusals expected with the introduction of blood sampling and DNA source collection at age 53 did not materialise (the response rate, as defined above, was 83%), possibly because interest in health is rising as age increases. Response rates to the seven annual postal questionnaires on women's health (Kuh & Hardy, 2003) and to the two home visits for a study of first-born offspring (Wadsworth, 1991) were high (each over 90%), probably also because of perceived relevance.

Data collections were designed not to overburden the sample. Although we would like to measure a wider range of indicators, we have resisted the temptation to develop a multipurpose study and restricted our concerns to particular health topics (cardiovascular, respiratory, and musculoskeletal health and cognitive function) and things that affect them. We concentrate on age-appropriate topic areas at different visits, in order to reduce demands on respondents and to maintain high quality measurement.

Representativeness, Loss to Follow-up, & Missing Data

Comparison with census data shows the responding sample at 53 years to be representative of the married, to under-represent in varying degrees those in lower SES groups, those who do not own their accommodation, men with university level qualifications, the never married, and separated or divorced women (Wadsworth et al., 2003). Avoidable losses by age 53 years are through refusal (12% of the original 5,362 sample) and inability to trace (6%). Recovery rates from temporary loss are generally high. At the most recent data collection (at age 53 years), avoidable loss was greater among those who in childhood had experienced health problems, who had low cognitive scores, and who were disruptive at school. Adult characteristics of those classified as avoidable losses by this age included manual SES, low or no educational and training qualification attainment, and earlier obesity (Wadsworth et al., 2003).

Of the 3,035 successfully contacted at age 53, 36% had been successfully contacted at all the earlier 19 data collections from the whole sample, and a further 37% had been successfully contacted at 17 or 18 of the 19 possible contacts. Only 7% had been successfully contacted at 10 or fewer contacts. Imputation techniques are being explored to infer information that is missing (Longford, Ely, Hardy, & Wadsworth, 2000).

The Scientific Value of the Data

The primary asset of many years of follow-up is the archive of life course data. That allows us to show, in particular, how the childhood endowment of physical and mental health is strongly associated with the social and economic environment of fetal development and early life, and how that endowment is the beginning of lifetime trajectories of health and SES (e.g., Jones, Rodgers, Murray, & Marmot, 1999; Richards, Hardy, Kuh, & Wadsworth, 2002; de Stavola et al., 2004). This study also has contributed to understanding functional change with age in adulthood and the pathways from childhood to adult health.

The Policy Value of the Data

In policy terms, the study has investigated effectiveness of health and education services, principally through its prospectively collected data that show health or attainment before and after interventions and through comparisons with data from other studies, both longitudinal and cross-sectional, about children born at other times who were exposed to other treatments, environmental and social risks, and diets (Ely, Richards, Wadsworth, & Elliot, 1999; Ferri, Bynner, & Wadsworth, 2003; Prynne et al., 1999). There is additional policy relevance for the future, in that the study sample represents the early post-war baby boom that is soon to become the beginnings of the boom in those of retirement age. Since we show continuities and trends in body shape and health-related behaviour and have mapped

baselines and changes from them in functional terms (e.g., in memory, blood pressure and respiratory function, and soon also in bone mineralisation and other musculoskeletal measures), we have a good picture of the state of health of this cohort as it arrives at the threshold of later life. Without a representative sample, the policy value of the study would be greatly reduced.

NEW BIRTH COHORT STUDY DESIGNS FOR HEALTH STUDIES

The Sample

The initial opportunity to select the sample for a proposed long-term follow-up study is a vital and irrevocable decision and can only relate to the science of its time. In a very long-running study, sample decisions can become a constraint.

Sampling is of people and time. In sampling individuals for a life course study, decisions must be made at the outset about sample size, sample units (e.g., individuals or families), geographic distribution, and whether to stratify by clinical and/or sociodemographic criteria. Those decisions have to be made in light of the study's scientific and policy objectives. Sample size decisions have to take account of whether relatively low prevalence conditions are to be studied, as well as estimates of likely loss through death, compliance failure, and migration. It may be appropriate to oversample in some population and geographical sectors, such as ethnic minorities or poor rural areas. Sampling of more than one historical time may be appropriate for some aims and can offer the opportunity for natural experiments through comparison. In Britain, second and third national birth cohort studies (Ferri et al., 2003) were initiated originally to determine if the introduction of the National Health Service and improvements in health care and population health were associated with reduced regional and socioeconomic variation in perinatal mortality and its associated risks.

Sample size considerations usually are driven by expected numbers of illness events and anticipated losses through death, migration, and refusals. Sample size also may need to be adequate for gene/environment interaction analyses. In the 1946 cohort, a larger population size would have provided greater numbers of events and allowed us to study some illnesses of low prevalence, such as multiple sclerosis. But the price for that would have been fewer data collections at longer intervals (because of cost) and a reduction in data quality (because of the increase in numbers of data collectors). Data collections in the two later-born British cohorts that each follow-up *all* the births in one week, as compared to our follow-up of *a sample* of all the births in one week, have been much more widely spaced (Ferri et al., 2003), with a consequent loss of prospective measures and greater reliance on memory.

Response & Representativeness

The value of a representative sample for both scientific and policy purposes is high.

If the sample is based on selected centres, the study should plan that those migrating within the country will not be lost to follow-up. Sample retention methods should be planned. In the experience of the British 1946 Birth Cohort Study, clear and frequent feedback of findings in accessible language helps to maintain sample compliance. It also is helpful to inform health care and educational professionals about the study, since sample members are likely to discuss participation with them. Further, feedback about test results and findings about individuals is likely to be via these professionals.

Selection of Data to Collect

It is useful to consider whether it would be desirable to measure and differentiate *health* as well as illness, since the majority of the population will be healthy. We have found it invaluable to measure adult function and its change (e.g., blood pressure and respiratory and musculoskeletal function) and temperament; we wish we had done so in childhood. We did not, because the science of the time concentrated on the search for early signs of disease. The cohort's childhood predates measures of temperament and

depression that we would now wish to use. Also, we wish we had collected and stored blood and other biological samples in childhood. It was of great value to study education as well as health, because they are interrelated, and because childhood measures of cognitive function (collected for educational study purposes) have been of value in the study of adult cognitive function.

In a longitudinal study, it is necessary to design current measures for their present value as outcomes and their future value as precursors and indicators of positions on pathways to health later in life. Today we seek to measure what we judge to be necessary to continue the database of measures of functional aging and morbidity, selecting measures appropriate to that end and in the hope that they will be of value at later cohort ages.

A longitudinal health study's measures have to be sufficiently fine grained and repeatable so that change with age and intra-individual diversity may be accurately measured. Collecting and coding data to the finest practicable detail maintains its usefulness in later times when scientific needs will be different. So, for instance, we have collected dietary data by retrospective and prospective diary methods, including all foods consumed, rather than frequency of consumption of selected foods. We have coded the data both as nutrients and as food sources (Prynne et al., 1999) and collected blood analyte data.

The Scientific & Policy Value of a New Study

Both scientific and policy objectives are essential, and the study must deliver under each of these headings if it is to continue. Although all aspects of these objectives cannot be described at the outset of the study, there are some general aspects that should be decided at that time.

It is important to have a unifying theme. That may be, for example, a common cause hypothesis to show the lifetime physical and mental endowment of health that is established in prenatal and early postnatal growth, and in early cognitive function and temperament. It is relevant to ask whether the data already exist to address such questions. Preliminary studies with existing data are usually appropriate before deciding on the need for a new life course study. For policy purposes, it is likely to be valuable to show the SES and geographic distribution of the mental and physical health endowment of a sample of children and to include information on their exposure to environmental risks.

The sampling design for that kind of purpose may not be appropriate for a study of morbidities of all kinds, including those of rare prevalence. It can be argued that it is not appropriate to study diseases of rare prevalence in a prospective long-term follow-up investigation, because the large numbers required have adverse consequences for the frequency of data collection (which is necessary during the early years) and for the quality of data collected. In any event, a probability sample is desirable for both scientific and policy aims.

CONCLUSIONS

The British experience, like that in some Scandinavian countries and the U.S., shows that long-running follow-up studies of health and social context are sustainable (Elder, Modell, & Parke, 1993; Ferri et al., 2003; Friedman et al., 1995; Giele & Elder, 1998; Laitinen, Pietilainen, Wadsworth, Sovio, & Jarvelin, 2004).

The large-scale life course studies in Britain and elsewhere in Europe currently are considering how to develop and improve interstudy comparability in order to prepare for meta-analysis and data pooling that will enhance their value as a resource for future gene-environment interaction research. We hope our American and Canadian colleagues also will wish to take part in the development of interstudy comparability.

REFERENCES

Barker, D. J. P. (1998). *Mothers, babies and health in later life.* Edinburgh: Churchill Livingstone.

de Stavola, B. L., dos Santos Silva, I., McCormack, V., Hardy, R. J., Kuh, D.J., & Wadsworth, M. E.

J. (2004). Childhood growth and breast cancer. *American Journal of Epidemiology, 159,* 671–682.

Elder, G. H., Modell, J., & Parke, R. E. (Eds.). (1993). *Children in time and place.* Cambridge University Press.

Ely, M., Richards M. P. M., Wadsworth, M. E. J., & Elliott, B. J. (1999). Secular changes in the association of parental divorce and children's educational attainment: Evidence from 3 British birth cohorts. *Journal of Social Policy, 28,* 437–455.

Ferri, E., Bynner, J., & Wadsworth, M. E. J. (Eds.). (2003). *Changing Britain, changing lives: Three generations at the turn of the century.* London: Institute of Education Press.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., Tomlinson-Keasey, C., Martin, L., Wingard, D. L. et al. (1995). Psychosocial and behavioral predictors of longevity. *American Psychologist, 50*, 69–78.

Giele, J. Z., & Elder, G. H. (1998). *Methods of life course research.* Thousand Oaks, CA: Sage.

Jones, P., Rodgers, B., Murray, R., & Marmot, M. (1994). Child developmental risk factors for adult schizophrenia in the British 1946 birth cohort. *Lancet, 344,* 1398–1402.

Kuh, D., & Hardy, R. (Eds). (2003). *A life course approach to women's health.* Oxford University Press.

.

Laitinen, J., Pietilainen, K., Wadsworth, M. E. J., Sovio, U., & Jarvelin, M.-J. (2004). Predictors of abdominal obesity among 31 year old men and women born in Northern Finland in 1966. *European Journal of Clinical Nutrition, 58*(1), 180–190.

Longford, N. T., Ely, M., Hardy, R., & Wadsworth, M. E. J. (2000). Handling missing data in diaries of alcohol consumption. *Journal of the Royal Statistical Society, Series A, 163,* 381–402.

Prynne, C. J., Paul, A. A., Price, G. M., Day, K. C., Hilder, W. S., & Wadsworth, M. E. J. (1999). Food and nutrient intake of a national sample of four year old children in 1950: Comparison with the 1990s. *Public Health Nutrition, 2,* 537–547.

Richards, M., Hardy, R., Kuh, D., & Wadsworth, M. (2002). Birth weight, postnatal growth and cognitive function in a national UK birth cohort. *International Journal of Epidemiology, 31*, 342–348.

Wadsworth, M. E. J. (1991). *The imprint of time: Childhood, history & adult life.* Oxford University Press.

Wadsworth, M. E. J., Butterworth, S. L., Hardy, R. J., Kuh, D. J., Richards, M., Langenberg, C. et al. (2003). The life course prospective design: An example of benefits and problems associated with study longevity. *Social Science & Medicine, 57,* 2193–2205.

FEATURE PAPER: Planning the National Children's Study

Adolfo Correa for the Interagency Coordinating Committee of the National Children's Study*

BACKGROUND

The National Children's Study (NCS) is a joint effort of the Department of Health and Human Services (DHSS) and the U.S. Environmental Protection Agency (US EPA) to study environmental influences on children's health and development. The NCS had its origins in the President's Task Force on Environmental Health Risks and Safety Risks to Children, which in 1997 was charged with the development of strategies to reduce risks from environmental exposures to children. The Task Force, co-chaired by the Secretary of DHHS and the Administrator of US EPA, concluded that many environmental and safety risks to children were not clear or quantified and proposed a longitudinal cohort study of the effects of environmental exposures (broadly defined) on the health and development of children.

In January 2000, the Developmental Disorders Work Group of the Task Force convened an expert panel to provide advice regarding the Task Force's proposal. The panel considered the experiences of a number of experts from past or ongoing major longitudinal studies and discussed the feasibility of embarking on such a large national study. The panel's discussions resulted in a strong endorsement of the proposed study and a number of recommendations:

(1) Specific hypotheses should be developed and applied.
(2) Families should be included along with index children.
(3) Planning must address ethical issues of collection, storage, and distribution of information, including biologic specimens, genetic material, and environmental samples.
(4) The study should be a collaborative effort among many Federal agencies.
(5) Modern information technology and bio-analytic and environmental monitoring techniques should be incorporated.
(6) New funds would have to be appropriated from Congress to carry out the study.

The panel's final message was to think boldly in the planning for such a study.

Planning for the NCS was mandated by the Children's Health Act of 2000, which authorized the National Institute of Child Health and Human Development (NICHD) "to conduct a national longitudinal study of environmental influences (including physical, chemical, biological, and psychosocial) on children's health and development." It instructed the Director of the NICHD to "establish a consortium of representatives from appropriate Federal agencies (including the Centers for Disease Control and Prevention, and the Environmental Protection Agency) to 1) plan, develop, and implement a prospective cohort study from birth to adulthood to evaluate the effect of both chronic and intermittent exposures on child health and human development; and 2) investigate basic mechanisms of developmental disorders and environmental factors, both risk and protective that influence health and developmental processes."

To lead the planning and implementation of the study, staff and funds have been allocated by the NICHD, the National Institute for Environmental Health Sciences (NIEHS), and the Centers for Disease Control and Prevention (CDC), all in DHHS, and by the Office of Research and Development of the U.S. Environmental Protection Agency (EPA). Investigators from each of these four lead

* A. M. Branum (CDC), G. W. Collman (NIEHS), A. Correa (CDC), S. A. Keim (NICHD), W. Kessel (DHHS), C. A. Kimmel (US EPA), M. A. Klebanoff (NICHD), (NIEHS), P. Mendola (US EPA), S. Newton (NIEHS), J. Quackenboss (US EPA), S. G. Selevan (US EPA), P. C. Scheidt (NICHD), K. Schoendorf (CDC), M. Yeargin-Allsopp (CDC).

entities serve on an Interagency Coordinating Committee (ICC) that has further developed the conceptual framework for the study, as well as an administrative structure and process for planning the study.

CONCEPTUAL FRAMEWORK

The rationale for the NCS stems from several observations:

(1) Exposures to some environmental agents (e.g., alcohol, lead) in utero and postnatally have been associated with serious developmental effects.
(2) Children experience frequent low-level exposures to a number of agents (e.g., pesticides, plasticizers) whose chronic or cumulative effects remain unknown.
(3) Existing studies are limited in size and scope.
(4) There is a need for studies to identify effects or assure effects from exposures to environmental agents.
(5) The optimal design to evaluate the relationships between multiple exposures and multiple outcomes is the longitudinal design.

Planning efforts for the NCS are based on the following principles.

(1) The NCS will be a high-quality longitudinal study of children, their families, and their environment.
(2) It will be national in scope.
(3) It will define "environment" broadly to include chemical, physical, behavioral, social, and cultural factors.
(4) It will study a range of common environmental exposures and less common outcomes.
(5) It will evaluate the relationship between environment and gene expression.
(6) It will use state-of-the-art technology for tracking, conducting measurements, and data management.
(7) It will involve a consortium of multiple agencies and extensive public-private partnerships.
(8) Finally the NCS will become a national resource for future studies.

Inclusion of other family members is desirable to facilitate studies of gene-environment interaction and the social environment. A total sample resulting in approximately 100,000 children has been proposed with follow-up to 21 years of age.

ADMINISTRATIVE STRUCTURE

The administrative structure for planning the NCS consists of the following organizational components:

(1) The NICHD Director, responsible for overall guidance and strategic decisions;
(2) The ICC, responsible for strategic planning and operational decisions;
(3) The Program Office at NICHD, responsible for day-to-day operations, administration of pilot studies, and protocol development;
(4) A Federal Advisory Committee, chartered under the Federal Advisory Committee Act, that manages the working groups and provides advice;
(5) Working Groups (n=22), comprised of federal and non-federal scientists (approximately 300), charged with development of potential hypotheses and proposals for measures and consultation;
(6) Federal consortium of agencies, comprised of representatives of federal agencies providing strategic guidance; and
(7) The Study Assembly, which includes all interested parties, meets periodically to receive updates on study planning, and provides a forum in which to discuss issues related to the study.

PRIORITY OUTCOMES AND EXPOSURES

Planning activities have identified a set of priority outcomes and exposure factors for inclusion in the NCS. *Priority outcomes* include undesirable outcomes of pregnancy, neurobehavioral development, injuries, asthma, obesity, and physical development. Anticipated *outcome measures* include fetal growth and outcomes of pregnancy; birth defects and newborn examinations; growth and physical development (e.g., weight, height, obesity, pubertal development);

information on medical conditions and history of illnesses (e.g., asthma, injuries); cognitive and emotional development; and mental, behavioral, and other developmental conditions.

Priority exposures and other factors include physical environment, chemical exposures, biological environment, psychosocial exposures, and genetics. Anticipated *exposure measures* include environmental samples of air, water, soil, and dust; biomarkers of exposures and genetic factors in blood, breast milk, hair, tissue, and urine; interview and history data on occupation, dietary intake, use of medications, supplements, and herbals; and information on housing and living

Table 1. Completed Methods Development/Pilot Studies for the National Children's Study

Design Issues

- Literature review of cohort studies (Lewin Group, 2000)
- Systematic review of potential hypotheses
- Methods of eliciting community involvement, subject recruitment, and retention
- Feasibility of using primary care practices for the NCS
- A systematic analysis of possible sampling strategies (Westat, 2002)

Exposure Issues

- Alternative exposure measurement design
  - *Predictors of exposure (questionnaire analyses)*
  - *Analysis of temporal variability*
  - *Efficient exposure measurement design*
- Methods studies (Sampling, Analytical)
  - *Long-term integrated sampler (SPMD); simple rapid methods for SPMD*
  - *Literature review for integrated samplers*
  - *Evaluation of disposable diapers for measuring pesticide metabolites in urine*
- Low-cost, low-burden exposure monitoring strategies
  - *Recruiting and retaining participants*
  - *Self-completed sampling, online questionnaire completion*
- Exposures and health of farm workers' children in California
  - *Monitor pesticide exposures for 24-month-old children via air, dust, surface/toy, food, breast milk, urine samples*
  - *Identify pathways, predictors, and algorithms*
- Evaluation of exposure assessment methods and approaches (White paper – Chemical Exposures Working Group)

Health-Related Issues

- Noninvasive collection methods and storage of samples for genetic testing
  - *Hair, nail, buccal swab DNA, validated with blood DNA*
- Biomarkers for assessing potential sensitivity of children
  - *Sensitivity to DNA-damaging agents*
  - *Surrogate tissues for genomic analysis of exposures and future disease states*
- Developmental neurotoxicity
  - *Develop a practical field-ready test system for neurobehavioral assessment*
  - *Related animal model*
- Childhood injuries
  - *Validity and reliability of parental reports*

Cross-Cutting Issues

- Review of new and emerging technologies applicable to the NCS
  - *Collection of health data, questionnaires, exposure information*
- Database of biomarkers for children's environmental health research
  - *Focus on asthma, pediatric cancer, injury, and neurodevelopment*
  - *Focus on air pollutants, pesticides, "exposure"*

characteristics, family and social experiences, and neighborhood and community characteristics.

Since no single hypothesis or research question can possibly reflect the intent of the Children's Health Act of 2000, and since hypotheses are necessary for framing the core protocol prioritizing measures and other costly elements of the study, criteria were developed for the purpose of selecting a set of core hypotheses. These criteria are (1) importance for child health and development (i.e., prevalence, severity, morbidity, mortality, disability, cost); (2) reasonable scientific rationale; (3) require a large sample size (~100,000); and (4) require longitudinal follow-up. Some example hypotheses by priority outcome are

- **Undesirable outcomes of pregnancy:** Infection and mediators of inflammation during pregnancy are major causal factors associated with preterm birth.
- **Neurobehavioral development:** Low-level pesticide exposure in utero is associated with impaired neurobehavioral and cognitive performance.
- **Injury:** Repeated head trauma without anatomic damage is a causal factor for cumulative adverse effects on neurocognitive development.
- **Asthma:** Experience with early bacterial and microbial exposures is associated with asthma (hygiene hypothesis).
- **Obesity and physical development:** Impaired glucose metabolism in pregnancy is associated with obesity and altered physical development (e.g., timing and progression of puberty).

PILOT STUDIES & WORKSHOPS

Methods development and pilot studies have been and are being conducted to assist in the development of the core protocol. Early in fiscal year 2000, these studies were initiated by the lead agencies and addressed general questions about exposures and outcome measures, as well as technology related to information gathering. Since then, several methods development/pilot studies have been initiated and completed on study design, exposure, health-related, and cross-cutting issues (Table 1).

In addition, workshops have been planned to address general measures expected to be included in the NCS. Table 2 lists completed and planned workshops. Products from these workshops include reports and white papers.

The choice of sampling design will need to take into account a number of complex issues, such as the following:

- Should the cohort be representative (i.e., probability-based sample)?
- Should women be enrolled before pregnancy or early in pregnancy?
- Should certain subpopulations (e.g., agricultural, industrial, economically disadvantaged) be oversampled?
- Which sampling design (e.g., center vs. home-based) enables the collection of reliable and accurate measurements based on physical examinations (e.g., prenatal glucose levels, fetal growth and development, neurodevelopment) or of biological specimens (e.g., placenta)?
- Which sampling design will ensure optimal recruitment, response, and retention rates?
- Which sampling design will ensure that a sufficient range of exposures and proportions of outcomes are represented in the cohort?

A workshop to consider various sampling approaches, including hybrid approaches, was held in March 2004. It brought together a panel of expert statisticians and epidemiologists to discuss various sampling approaches and will result in a report on leading sampling strategies.

Future pilot studies for the NCS will consist of the feasibility of the leading sampling strategies, reviews of the literature, state of the science, available instruments and measures, lessons learned from other studies (e.g., Children's Environmental Health Centers), pilot of specific measures, and pilot of the full protocol at vanguard centers.

Table 2. Workshops Completed and Planned for the National Children's Study

Completed

- Community engagement
- Fetal and neonatal growth and development assessment methods
- Medicine exposure: collection, coding, and classification
- International consultation on longitudinal cohort studies
- Innovative technologies for remote collection of data for the NCS
- Ethical issues in longitudinal pediatric studies: "Looking back, thinking forward"
- Assessing the incidence and outcomes of mild traumatic brain injury
- Placental measurements
- Psychosocial stress and pregnancy and the infant
- Physical activity
- Herbals and dietary supplements
- Effects of the media

Planned

- Impact of rural environment
- Sampling design
- Measures of social environment
- Growth and development
- Day-specific probabilities of pregnancy
- Questionnaires and diary-based methods for the early assessment of asthma-related health outcomes
- Gene expression and behavior
- Measurement of maternal and fetal infection and inflammatory response
- Assessing dietary intakes and patterns in women and young children
- Measures for health care processes and outcomes

CURRENT & FUTURE ACTIVITIES

In addition to the planning and conduct of pilot studies and workshops, a major focus of current planning activities is the development of the core protocol by the NCS Program Office staff. With respect to future activities, it is anticipated that the initial centers and pilot testing of the core protocol will occur in late 2005 or in early 2006, the full study will start in late 2006 or early 2007, and follow-up will be for at least 21 years, with the first preliminary results available from pregnancy in 2009–2010.

A number of issues related to children's health have gained increasing importance in the past decade in the public health and risk assessment communities. These issues include the following:

- What is the role of environmental factors, including diet, in children's health and development?
- Does exposure to particular environmental agents increase the burden of disease in children?
- What are the effects of aggregate exposures to a chemical or to cumulative exposures to mixtures?
- Are there long-term effects from early exposures of children to environmental factors (e.g., asthma, obesity, diabetes, cardiovascular diseases, or neurological diseases)?
- What genetic factors alter the susceptibility of children to the effects of environmental agents?
- What are the differences in response to environmental exposures and susceptibility by age or life stage?
- Are there disparities in children's health due to race/ethnicity, poverty, housing, income?

It is anticipated that the NCS will provide the information to address these issues. The NCS also is expected to provide the following:

- Data to determine harmful, harmless, or beneficial effects of exposures;
- A longitudinal framework for determining risk factors for a number of diseases and conditions of children;
- Information on how multiple exposures interact to result in multiple outcomes;
- Information on the role of gene expression in the effects from environmental factors;
- Identification of early life factors that contribute to many adult conditions; and
- A national resource of stored biological and environmental samples and extensive interview data to answer future questions for decades to come.

Additional information on the NCS can be obtained by visiting the NCS Web site (http://NationalChildrensStudy.gov), joining the listserv for news and communication, or by e-mail (ncs@mail.nih.gov).

REFERENCE

The National Children's Study Interagency Coordinating Committee. (2003). The National Children's Study of environmental effects on child health and development. *Environmental Health Perspectives, 111*(4), 642–646.

FEATURE PAPER: Disability and Informal Support: Prospects for Canada

Michael Wolfson and Geoff Rowe, Statistics Canada

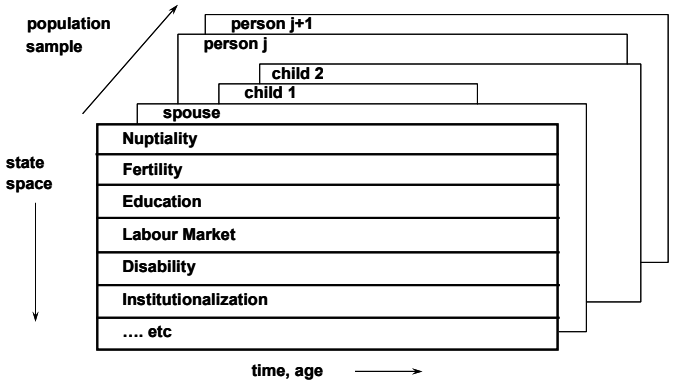
INTRODUCTION

Population aging in Canada and other countries continues to raise questions regarding who will provide care for an increasing number of frail elderly. Broadly, there are two main types: informal care most often from close relatives and formal care provided through government programs. The expected demand for publicly funded care will depend in part on the availability of informal care. The objective of this analysis is to project the need for and supply of close family support for Canada's elderly over the next two decades.[1]

These projections require extrapolations based on historical trends in a range of key demographic events, such as marriage and fertility, as well as the relationships like that between education and fertility, to take account of underlying trends, such as increasing levels of educational attainment. As a result, the methods used for this analysis are based on computer simulation, in particular, the Statistics Canada LifePaths model.

In the following sections, we first sketch the basic concepts of microsimulation as implemented in the LifePaths model. The following section describes the analysis of the longitudinal National Population Health Survey (NPHS) used to estimate transitions among disability states. These statistical descriptions of disability dynamics were then built into the LifePaths model and used with demographic projections to explore the likely future joint prevalence of disability and availability of informal family support.

Figure 1. State Space and Longitudinal Micro Data Sample Generated by a LifePaths Simulation



OVERVIEW OF LIFEPATHS STRUCTURE

LifePaths[2] is a computer simulation model that produces, with each run, a representative microcosm of the Canadian population. It is microanalytic—the basic units of observation are individuals—and is focused on microlevel dynamics—how individuals move among various mixtures of socioeconomic states over their life courses.

Empirically, LifePaths is metasynthetic—drawing upon multiple data sets, covering diverse subject matters, and using each in order to assemble the best possible overall estimate of the information of interest.[3]

The basic unit of analysis in LifePaths is an individual life history, as Figure 1 shows. The "state space" of attributes or individual characteristics is shown along the vertical axis, with age and calendar time along the horizontal. The third axis indicates a representative sample of individuals in the population of interest. These are not all unrelated individuals; rather, they are juxtaposed to show that family structure also is included.

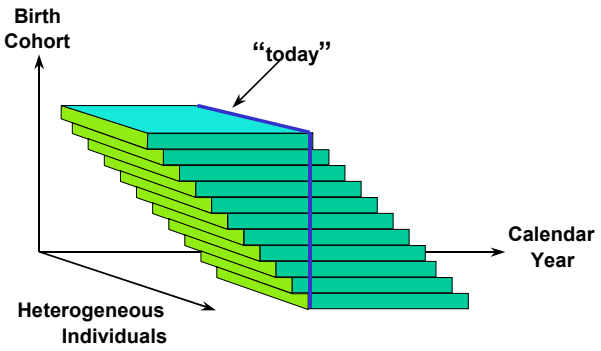
Earlier versions of this paper were presented at the International Conference on Health Policy Research, October 2003, and the United Nations Statistical Division, January 2004, and have benefited from the comments and discussion that ensued. A fuller version of this paper is available on request. The authors remain responsible for any errors or infelicities.

[1] These projections are being undertaken in a research mode and are not official projections of Statistics Canada.

[2] More details on LifePaths are available at www.statcan.ca/english/spsd/LifePaths.htm.

[3] The term "metasynthesis" is used in contrast to the epidemiological term "meta-analysis," which refers to the combination of results from a number of data sets, all of which pertain to the same question.

Figure 2. Overlapping Birth Cohorts with Heterogeneous Members



Given these microlevel life histories as the basic building blocks, LifePaths assembles large representative samples of individuals grouped into nuclear families in a sequence of overlapping birth cohorts (Figure 2). Each "layer" in the diagram represents one birth cohort, while the sequence of layers represents successive birth cohorts. A typical population pyramid showing age structure by sex at a point in time corresponds to a vertical slice through the overlapping birth cohorts along the line for "today." [4]

LifePaths essentially creates a large sample of representative individual life histories, where the individuals have been born throughout the 20th century in accord with historical population data. The historical reconstruction and projection processes proceed by data synthesis using longitudinal microsimulation: Each individual's life history is synthesized, starting at birth and then recursively generating the events and characteristics shown along the vertical axis of Figure 1 until death. Then another family of individuals is synthetically generated, and again, and again, until a very large sample (e.g. 1,000,000) is generated. The result is our "fitted" population microcosm (for years prior to "today"), plus microlevel extrapolations of each life history beyond "today" (if still alive) over coming decades. The result is a very large longitudinal sample of synthesized individuals that reproduces a diversity of observed data, such as population characteristics from censuses and mortality and fertility rates dating back to about 1900, age- and sex-specific employment/population ratios since the 1970s, and 1990s disaggregated disability prevalences.

BASIC DYNAMICS

Underlying a LifePaths simulation is a detailed set of empirically based state transition dynamics. In this analysis, essentially all of the characteristics are categorical or discrete (e.g., marital status and disability level). As a result, dynamics are represented by transition probabilities. The first main group of transitions relates to sociodemographic status: nuptiality, fertility, and educational attainment.

For example, the nuptiality transitions explicitly modeled are shown in Figure 3. The different states are given by the boxes, while the arrows indicate the possible transitions. For each arrow, there is an empirically estimated transition probability, which in turn is a function of time-varying covariates. The transition probability functions have been estimated initially from survey data and then (where possible) adjusted so that LifePaths as a whole will reproduce the distribution of families by marital status observed in Canada's 1996 population census.

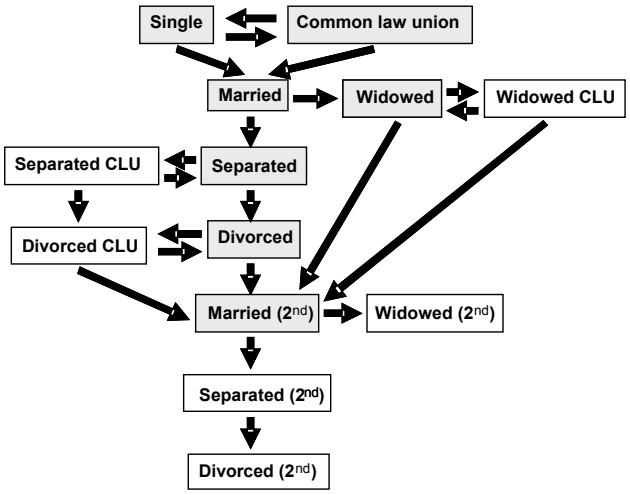
DISABILITY STATES

The second main group of transition probabilities is for disability and institutionalization. Due to limited data availability, these transitions are not as robust as those for the sociodemographic transitions. Reasonable disability data are available mainly from the 1990s. In particular, the NPHS (Statistics Canada, 1998) provides longitudinal data from 1994 to 2000, as well as cross-sectional prevalences. The NPHS covers both the household population and those in institutions, and the survey included several measures of disability. As a result, the NPHS has provided the basis for both prevalence distributions of disability by level of severity and microlevel estimates of transition dynamics.

Following Carrière et al. (2003), we have defined disability in terms of the characteristics most likely to be associated with the need for assistance in performing everyday activities, concentrating on four activities: everyday housework, grocery shopping, meal preparation,

[4] Although the diagram implies that time is discrete, LifePaths represents and models all events in continuous time.

Figure 3. Nuptiality States and Transitions



and personal care. In turn, these activities were posited to depend primarily on the following kinds of disabilities or impairments: mobility, dexterity, cognitive capacity, and pain, each of which is a specific subscale of the McMaster Health Utility Index Mark 3 (HUI3; Grootendorst, Feeney, & Furlong, 1999). Five levels of severity of disability were defined, as shown in Table 1, where the level assigned to an individual corresponds to the most severe item in his profile

ANALYZING DISABILITY STATE TRANSITIONS

Transition probabilities for movements between each possible combination of these disability states were estimated using longitudinal NPHS data spanning the 1994, 1996, 1998, and 2000 cycles of the survey.

The choice of specification for the statistical analysis is driven in part by the nature of the LifePaths modeling architecture, which is very open and invites possibly richer specifications than would be typical in epidemiological or other statistical work. In these latter contexts, the focus is usually on whether a given covariate is statistically significant with regard to some outcome, and if so, the direction and relative magnitude of the influence. In this case, we are using a simulation model to integrate results from a range of empirical analyses, so there is no penalty for being expansive in the specifications. Moreover, LifePaths' architecture, in particular its continuous time, discrete event, explicit competing risks, non-parametric character invites as detailed a statistical description of disability dynamics as the available data allow and suggest is warranted.

To account for the ordinal character of the disability states, we assume the underlying process has a phase-like structure (Aalen, 1995; Aalen & Gjessing, 2001): a process in which individuals make repeated transitions up or down in a health status continuum. Thus, we take the observable features of the process to be whether an individual's health has improved or deteriorated between one interview and the next.

LifePaths directly implements the notion of competing risks. An individual in an intermediate disability state might improve her health status, or her disability may worsen, for example. LifePaths reflects such competing risks by drawing two waiting times: one for an improvement, and one for a deterioration. The event actually simulated to occur is the one with the shorter waiting time.

To allow for transition probabilities varying not only with the current disability state but also with the individual's prior disability trajectory, the estimation draws on both lagged and current disability status. With four waves of the panel survey data, we can evaluate transitions between times 't+1' and 't+2' conditional on disability status at times 't' and 't+1', as well as evaluating transitions between times 't+2' and 't+3' conditional on disability status at times 't+1' and 't+2'. As a result, most of these transition

Table 1. Definition of Disability States

- **No disability**
- **Mild disability**
  - Mobility problem but does not need any help
  - Dexterity problem but does not need any help nor special equipment
  - Somewhat forgetful & little difficulty in thinking
  - Moderate &/or severe pain prevents performing some or few activities
- **Moderate disability**
  - Requires wheelchair or mechanical support to walk
  - Dexterity problem & needs help to perform some tasks
  - Very forgetful & a lot of difficulty in thinking
  - Severe pain prevents performing most activities
- **Severe disability**
  - Cannot walk or needs help from others to walk
  - Dexterity problem & needs help for most or all tasks
  - Unable to remember or think
- **Living in an institution**

"triples" will be in pairs, with two (overlapping) triples collected from the same respondent. Thus, person-specific terms are added to the equations to account for the correlation between observations drawn from the same respondent. These terms represent otherwise unobserved person-specific factors and, as such, reflect the effects of (some) omitted variables.

We used the SAS procedure NLMIXED for our estimation, which permits specification of a conditional Poisson probability for binary transition events (i.e., *y:* representing either improvement or deterioration in health*)* jointly with an unobserved person-specific Gaussian random effect *b.* In the following expressions, 'i' denotes a respondent and 't' a time-period. PY_{it} represents approximate person-years-at-risk (i.e., the two years between interviews for respondents who experienced no 'event' and one year for respondents who reported either improvement or deterioration in health at the second interview, assuming for simplicity that any events occurred midway between sample interviews).

$$y_{i,t} = PY_{i,t}\exp\left(X_{i,t}\beta + \sigma\, b_i\right) + \varepsilon_{i,t}$$

$$p\left(y_{i,t} \mid PY_{i,t}, X_{i,t}, \beta, \sigma\right) = \int p\left(y_{i,t} \mid PY_{i,t}, X_{i,t}, \beta, \sigma, b_i\right)\, p(b_i)\, db_i$$

$$where\ y_{i,t} \mid PY_{i,t}, X_{i,t}, \beta, \sigma, b_i \sim Poisson\left(PY_{i,t}\exp\left(X_{i,t}\beta + \sigma\, b_i\right)\right)$$

$$and \quad b_i \sim Gaussian\left(0, \sigma^2\right)$$

Explicit estimates of b_i can be obtained given multiple observations from most respondents. Such estimates resemble averaged respondent-specific residuals and so must be determined simultaneously with the fixed regression parameters. Parameter estimation is carried out by maximizing a marginal likelihood obtained by integrating the b_i's out of the expression.

We estimated separate equations for each initial disability state. We also incorporated an explicit variance equation (making the logarithm of the standard deviation be a linear function of covariates):

$$\sigma_i = \exp\left(Z_i\theta\right)$$

and by adding the terms Z_i and θ to the likelihood. The variance equation may be interpreted as identifying factors associated with heterogeneity within the population.

The NPHS uses a multistage survey design, meaning there is no simple formula that can be used to estimate variances. Instead, bootstrap survey weights are provided as part of the NPHS to permit variance estimation of most statistics of interest (Yeo, Mantel, & Lie, 1999). Cross-validation was used for evaluating different possible specifications for the hazard regressions, by directly assessing the prediction error of each fitted equation. The available sample was split, and one part used to fit the equation (model construction), while the other part was used for an assessment of predictions (model validation; Picard & Cook, 1984). This was straightforward given the availability of the 500 bootstrap sample weights. (Each bootstrap subsample randomly excluded some respondents, assigning them a weight of zero.) Each regression was estimated 500 times, and each time predictions were made for the portion of the sample whose weights were zero (i.e., not used in the regression). As a result, 500 goodness of prediction statistics were generated for each equation, and the choice among competing equation specifications generally was based on the median prediction statistic.

The implementation in the LifePaths model of the disability status transition equations involved the following steps:

(1) At birth, each simulated individual is deemed to have no disability.
(2) Persons born outside of Canada (predestined to become immigrants) are not subject to mortality or to disability transitions until they arrive in Canada.
(3) Each simulated individual is assigned a single random number at birth that will correspond to the random terms (*b*'s) in the estimated equations (drawn from a Gaussian distribution with a mean of 0.0 and a variance of 1.0 and remaining fixed throughout an individual's life).
(4) The magnitude of the influence that the random terms have is determined by the estimated terms in the variance equation (i.e., a function of the simulated individual's current disability state, education level, marital status, and immigration status).
(5) As each simulated individual progresses through life, the chances of a disability improvement or deterioration are determined by current age and disability state, disability state 24 months previously,

time-varying covariates (e.g., age), and the fixed random term assigned at birth.

(6) Each time one of the right-hand-side variables changes (and at least once every month when the lagged disability state is updated), a random waiting time to disability improvement and a competing random waiting time to disability deterioration are generated. If either is less than a month, the corresponding disability status transition will occur at the scheduled time (unless an intervening event occurs first that changes another of the right-hand-side variables).

(7) Given 'Severe Disability' as the current simulated disability state, a transition involving deterioration of health is taken to imply 'Institutionalized.'

(8) LifePaths contains separate, average baseline mortality schedules for each sex and for each cohort born after 1871 (projected into the future as necessary). Age-sex specific relative risks of mortality (estimated from NPHS data) for each disability status grouping also were introduced to reflect the relatively low mortality risks experienced by those with no disability compared to the relatively high mortality risks experienced by the severely disabled or the institutionalized.

These modules have been validated by a range of detailed comparisons with historical census, NPHS disability, and other data.

MAIN RESULTS

Table 2 presents the results on the current prevalence of disability and counts based on one scenario projected to 2021. The first set of columns shows the numbers of individuals estimated for 2001 in Canada by age group and disability status. For example, among the 65–69 age group, about 17% were moderately or severely disabled or institutionalized (about 190,000 of 1.1 million), rising to about 27% in the 75–79 group and to almost 43% in the 90+ group.

The second set of columns shows one set of projected *changes* in counts between 2001 and 2021. The number of persons age 65+ would grow by about 2.6 million by 2021; the numbers with moderate/severe disability or institutionalized grow by about 770,000. In other words, over two-thirds of the added numbers of seniors could well be either mildly disabled or (more likely) not disabled at all.

Table 2. Population Counts by Age Group and Disability Level, Both Sexes (in thousands) /

5-Year Age Group	COUNTS, 2001 — Level of Disability					CHANGES IN COUNTS, 2001 TO 2021 — Level of Disability				
	None	Mild	Moderate	Severe or Institution	**All**	None	Mild	Moderate	Severe or Institution	**All**
65	716	188	93	96	**1,093**	626	182	85	110	**1,003**
70	553	189	98	112	**953**	383	117	74	100	**675**
75	401	155	90	118	**764**	164	71	41	77	**352**
80	230	103	62	90	**485**	75	44	26	62	**206**
85	110	59	36	57	**262**	58	28	23	55	**164**
90+	61	35	22	50	**168**	67	46	32	95	**241**
Ages 65+	2,073	728	401	523	**3,725**	1,374	488	282	498	**2,642**

Table 3. Projected Changes in Prevalences of Close Family Members by Age Group (in thousands) /

5-Year Age Group	COUNTS, 2001					CHANGES IN COUNTS, 2001 to 2021				
	No spouse, no children	Spouse only	Spouse & children	Children only	**All**	No spouse, no children	Spouse only	Spouse & children	Children only	**All**
65	88	99	623	284	**1,093**	143	186	392	282	**1,003**
70	91	79	470	312	**953**	86	110	270	210	**675**
75	85	58	298	323	**764**	47	47	136	123	**352**
80	71	24	133	257	**485**	19	20	75	92	**206**
85	52	11	48	151	**262**	5	6	38	115	**164**
90+	44	5	12	107	**168**	28	5	29	178	**241**

Table 4. Projected Changes in Joint Prevalences of Moderate or Severe Disability or Institutionalization and Presence of Family* (in thousands)

	COUNTS, 2001				CHANGES IN COUNTS, 2001 to 2021			
5-Year Age Group	Not or mildly disabled	At least moderately disabled & family available	At least moderately disabled & no family available	All	Not or mildly disabled	At least moderately disabled & family available	At least moderately disabled & no family available	All
65	904	174	15	**1,093**	808	169	26	**1,003**
70	742	192	18	**953**	501	153	22	**675**
75	556	186	22	**764**	234	101	17	**352**
80	333	131	21	**485**	119	76	11	**206**
85	169	73	19	**262**	86	73	5	**164**
90+	96	52	20	**168**	114	111	16	**241**

* "Family" is here defined as a spouse or (adult) child.

A basic concern with any projected increase in the numbers of disabled elderly like that shown in Table 2 is the pool of close family members who might be available to provide informal support. Table 3 shows corresponding projections of the numbers of elderly with and without spouse, adult children, and with neither.

In 2001, the proportions of those who had no spouse or children alive ranged from under 10% in the 65–69 age group to just over 25% in the 90+ group. Of the increase of about one million individuals age 65–69 shown for 2021, almost 15% would be without spouse or children, while in the 90+ age range, this would be just over 10%. Thus, notwithstanding the "baby bust" fertility rate decline after the mid 1960s and the sharp increase in the divorce rate after the late 1960s, in this scenario we do not see a growing proportion of the oldest old who have no close family in 2021. The main reason is not the unimportance of these major demographic changes but rather the fact that they will not have their greatest impacts until later decades.

Table 4 puts the two perspectives together to indicate the numbers of individuals who could be both at least moderately disabled and without close family members who might provide informal care.[1] Over 90% of those age 65–69 are not moderately or severely disabled or institutionalized in 2001 (based on the specific definition of disability that has been used). This proportion falls with higher age, so that in the 75–84 age group, the proportion is about two-thirds, and in the 90+ age range it is about 57%.

Among those at least moderately disabled in 2001, over 90% in the 65–74 age range at least had the potential of calling on a spouse or child for informal care and support. This proportion falls to about 72% in the 90+ age range. Overall, the vast majority of Canadians age 65+ in 2001 either was not seriously disabled or had living close family relatives.

The right half of Table 4 shows one scenario for projected changes in these counts for 2021. The population age 65+ with at least moderate disability and no close family members is shown growing by about 100,000, out of a total growth in the population age 65+ of more than 2.6 million.

CAVEAT: SOURCES OF UNCERTAINTY

The results just presented draw on a very large data synthesis exercise, including the specific analysis of disability dynamics from Canada's National Population Health Survey. All of the underlying data embody errors of various sorts, including both sampling and non-sampling error in individual data sets, as well as possibilities of bias in the causal stories that have been elicited from the data due to missing variables. They also depend on many assumptions, including the functional forms for various statistical relationships that have been

[1] This is a lower bound on the numbers of "needy" elderly because we have not taken any account of whether the adult children live nearby nor whether the child or spouse is in sufficiently good health that he or she could in fact provide support.

fitted to the data and the way they have been assembled into the larger whole that is the LifePaths model.

We have taken a number of steps to ameliorate concerns about the robustness and reliability of the presented results. Sufficiently large samples have been generated in each simulation so that Monte Carlo error is negligible. For the disability dynamics, systems of equations have been used, disaggregating where the data or other evidence suggests it is appropriate, and alternative specifications have been assessed using cross validation. Generally, simulation results also have been compared and, if necessary, "aligned" so that results fit the historical data as well as possible.

Nevertheless, there is a rather pervasive assumption of "conditional independence" underlying these results. For example, the demographic dynamics have been estimated largely without accounting for disability status, mainly due to data limitations. It is plausible that individuals with congenital disabilities are less likely to marry and have children; this has not been taken into account here. Judging from Table 2, though, the numbers of such cases are likely small relative to the counts that have been the focus of discussion. And in the case of disability dynamics, marital status data, for example, were available and included among the candidate covariates.

The other major source of uncertainty derives from the fact that the key estimates are projections, where errors are intrinsically unknowable. In particular, given the novelty of the disability analysis and the still unresolved debate in the broader health literature as to whether population aging is being accompanied by a "compression of morbidity" (Fries, 1980), disability dynamics is likely the area of greatest uncertainty in the projections. Other factors are also important, such as assumptions about fertility, migration, and mortality rates. However, for this analysis, it is clear that fertility rates are not of direct import, because all those who will be 65+ in 2021 are alive today and probably have had all the children they are going to have. Immigration rates could be important over the next two decades, but this is likely of second order importance. Mortality rates are certainly important, and sensitivity to these assumptions will be discussed in a moment. Similarly, projections of union formation and dissolution rates have considerable uncertainty, particularly with the growing importance of common-law unions. But given the relative counts of individuals projected to be at least moderately disabled and to have close family members, disability is likely to be the more important factor.

As a result, the sensitivity of the results in Table 4 has been assessed by constructing two alternative scenarios for disability dynamics: one a compression of morbidity where disability dynamics is posited to be delayed five years, and the other an expansion where disability dynamics are advanced five years. The main results are quite sensitive to these alternative assumptions, though we have no good basis for judging whether this assumed range is plausible—consistent time series data on disability in Canada are unavailable for long-term trends.

Comparing these scenarios generates a range of about 500,000 in the size of the 65+ population in 2021, a change of about 9% up or down in the increase in the population age 65+ by 2021. (Recall that mortality rates depend on disability as well as age and sex.) Further, the scenarios suggests that the projected increase of 27% in the numbers of elderly at least moderately disabled in 2021 could range from 17% to 37%.

SUMMARY AND CONCLUSIONS

A growing concern in societies with aging populations is who will provide care for the frail or sick elderly. Much of the care for these individuals to date has been informal, relying substantially on friends and family. With the decline in fertility rates and increased rates of marriage dissolution, however, future elderly could have fewer close family members available to provide informal support. On the other hand, declining mortality rates would suggest that future elderly will be more likely to have surviving children and spouses. An important question is the expected numbers of frail elderly without close family who could provide informal support and hence the relative magnitudes of these trends.

We have drawn on a range of highly multivariate longitudinal microdata sets and microsimulation modeling embodied in Statistics Canada's LifePaths model to estimate and project the joint patterns of disability levels and potential availability of informal support for a representative sample of individuals' life cycles to 2021. LifePaths incorporates detailed and pre-existing work on a range of factors, including educational and demographic transitions. In this analysis, we added a disability state transition submodel where a nonlinear function of age, educational attainment, living arrangements, age at immigration, and recent disability history, as well as unobserved person-specific factors, all were statistically important. The statistical specification represented competing hazards of progressive deterioration or improvement in health, including institutionalization and death.

Overall, Canada's population age 65 and over is projected to grow by about 2.6 million from 2001 to 2021, based on a middle scenario. The number of these individuals with at least moderate disability is projected almost to double, from about 925,000 to about 1.7 million. In proportionate terms, however, the growth in the prevalence of disability could be much smaller, increasing among the 65+ population from about 25 to 27%. Sensitivity analysis suggests a considerable range of uncertainty around these projections. Based on the scenarios examined for compression or expansion of morbidity, this could be from 17 to 37%.

In 2001, about 115,000 of those age 65+ could be considered "needy"—they were at least moderately disabled and had no living spouse or children who might potentially provide informal care. This number is projected to increase by about 100,000 in 2021. Again, this is likely a lower bound on the numbers of such individuals, since a living spouse or child could be too ill or live at too great a distance to provide any care.

These kinds of results are fundamental to planning for the aging of Canada's population. For example, there is great concern about the sustainability of Canada's health care system and, to a lesser extent, public pensions. At the same time, the caveats in the previous section should be borne in mind. While the analysis underlying these projections has been very thorough and made unprecedented use of the widest range of available data, it still rests on incomplete and imperfect data and a range of assumptions.

In the end, perhaps the best way to assure that these kinds of results are "fit for use," for example in public policy analysis, is for others to build on this version of the LifePaths model, try alternative modules, and use the model to generate a wider range of scenarios.

REFERENCES

Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics, 22*(4), 447–463.

Aalen, O. O., & Gjessing, H. K. (2001). Understanding the shape of the hazard rate: A process point of view. *Statistical Science, 16*(1), 1–22.

Carrière, Y., Légaré, J., Keefe, J., Rowe, G., Martel, L., Lin., X. et al. (2003, June). *The use of microsimulations to better understand the effect of changing family structure on the needs for formal home care services.* Paper presented at the annual meeting of the Canadian Population Society, Halifax, Nova Scotia.

Fries, J. (1980). Aging, natural death, and the compression of morbidity. *New England Journal of Medicine, 303,* 130–135.

Grootendorst, P., Feeney, D., & Furlong, W. (1999). *Health Utilities Index Mark 3: Evidence of construct validity for stroke and arthritis in a population health survey* [Department of Economics Working Paper 1999-09]. Hamilton, Ontario: McMaster University, Department of Economics.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79,* 575–583.

Yeo, D., Mantel, H., & Liu, T.-P. (1999). Bootstrap variance estimation for the National Population Health Survey. In *1999 Proceedings of the Survey Research Methods Section* (pp. 778–783). Alexandria, VA: American Statistical Association.

FEATURE PAPER: . Design and Estimation Strategies in the Medical Expenditure Panel Survey for Investigation of Trends in Health Care Expenditures

Trena M. Ezzati-Rice and Steven B. Cohen, Agency for Healthcare Research and Quality

INTRODUCTION

Population-based health care surveys are an important resource to inform health care policy and practice. To be cost effective, such surveys often are designed to provide a cross-sectional "snap-shot in time" of, for example, the population's health status and health care access, utilization, and expenditure experience. While aggregate cross-sectional health care statistics are extremely important, surveys that include a longitudinal feature allow for a better understanding of both the incidence and duration or transition in certain health status or use states, such as lacking insurance coverage or how the distribution of health care expenditures changes with time. Moreover, longitudinal studies can provide information for the study of current and emerging policy issues, such as the persistence of exceptionally high or inadequate levels of medical services use and associated health care expenditures and how individual characteristics, behavioral factors, financial incentives, and institutional arrangements affect health care utilization and expenditures in a rapidly changing health care market.

The design of and analytical requirements for a national health care and expenditure survey present a unique set of challenges in terms of sample design, survey content, and estimation strategies. In this paper, we discuss how the demand for essential longitudinal health care and expenditure information is translated into the design requirements of a national health survey. Specifically, we summarize (1) the design features of the Medical Expenditure Panel Survey (MEPS) and (2) the estimation strategies used to support the measurement of health care expenditures and related time trends. Selected examples that demonstrate the analysis potential of MEPS to inform both survey design and health care policies are provided using annual, panel-specific, and longitudinal data.

MEPS SURVEY DESIGN

The capacity to conduct detailed analyses of the health care expenditure experience of the population each year and to examine patterns over time are two major design features of the MEPS. In particular, a major strength of the MEPS is its capacity for longitudinal analyses, but this feature adds additional complexity for the survey's design, achievement of acceptable response rates, and estimation and analytic strategies. Sponsored by the Agency for Healthcare Research and Quality (AHRQ), MEPS is an ongoing longitudinal panel survey of the U.S. civilian noninstitutionalized population. Since its inception in 1996, its primary analytical focus has been on health care access, coverage, cost, and use. MEPS also provides estimates of measures related to health status, demographic characteristics, employment, income, access to health care, and satisfaction with health care. Estimates can be produced for individuals, families, and selected population subgroups.

The MEPS includes a family of three interrelated surveys: the Household Component, the Medical Provider Component, and the Insurance Component. The Household Component (MEPS-HC) is the focus of this paper. Each year a new sample (panel) of households is selected for the MEPS-HC. This set of households is a subsample of those households participating in the previous year's National Health Interview Survey (NHIS), an ongoing annual household survey of approximately 42,000 households (109,000 individuals) conducted

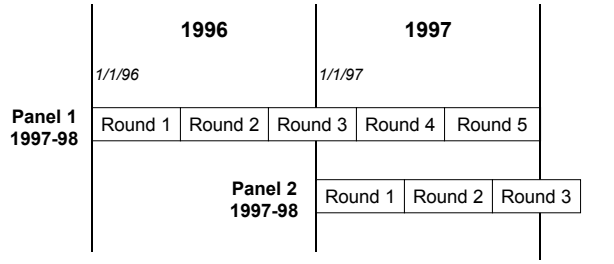
The views expressed in this paper are those of the authors and no official endorsement by DHHS or the Agency for Healthcare Research and Quality is intended or should be inferred.

by the National Center for Health Statistics, Centers for Disease Control and Prevention (Botman, Moore, Moriarity, & Parsons, 2000). Because Hispanics and African Americans are oversampled for the NHIS, these minority groups are efficiently identified for a targeted oversample each year in MEPS. In addition to the cost savings achieved by eliminating the need to independently list and screen households, selecting a subsample of NHIS participants provides enhanced analytical capacity for the resultant survey data. Specifically, use of the NHIS data in concert with the data collected in the MEPS provides additional capacity for longitudinal analyses not otherwise available. Also, the large number and dispersion of the primary sampling units (195 PSUs) in the current MEPS has resulted in improvements in precision over prior expenditure survey designs (Cohen, 2003).

To achieve a longitudinal component, the MEPS-HC consists of an overlapping panel design. Health care and expenditure data are collected for each new MEPS sample, which is interviewed five times in person over 30 months to yield annual use and expenditure data for two calendar years (Figure 1). Each computer-assisted personal interview takes place with a family respondent who reports for him/herself and for other family members. To produce estimates for a calendar year, data are pooled across two distinct nationally representative MEPS samples. More specifically, the annual estimates are based on combined data from the second year of one panel and the first year of the subsequent panel (Figure 1 and Cohen, 2000).

To help inform an annual profile of the quality of health care and to assess improvements over time as part of AHRQ's National Healthcare Quality Report (NHQR), the MEPS implemented content enhancements and a sample expansion in 2002. A significant sample expansion was made to ensure an adequate number of individuals with certain illnesses of national interest in terms of quality of care and burden of disease (Cohen, 2003). Similarly, to inform the National Healthcare Disparities Report (NHDR), a

Figure 1. MEPS Household Component Overlapping Panel Design



larger MEPS sample was needed to allow for greater representation of population subgroups that included racial and ethnic minorities and low-income individuals (Cohen, 2003). Thus, for the 2002 MEPS and subsequent years, the overall sample size was increased to 15,000 households and 39,000 individuals. The sample increase allowed for an oversample of two additional subgroups of the population (in addition to African Americans and Hispanics): Asian Americans and persons predicted to have family income below 200% of the federal poverty level. MEPS sample sizes by year are provided in Table 1. The design enhancements to inform the NHQR and NHDR with the larger overall national sample and greater representation of selected population subgroups had the added attraction of improving the capacity of the survey to produce health care expenditure estimates at the national level and for policy-relevant population subgroups, such as the poor, the elderly, children, and minority subgroups (Hispanics, African Americans, Asians) with much greater levels of precision.

A major MEPS design feature is the ability to support both cross-sectional and longitudinal analyses to address current and emerging health care policy issues. As discussed previously, improved statistical power for the annual estimates is achieved by combining data from two panels to produce health care estimates for the calendar year. Also, as a consequence of the design and the five rounds of interviews covering two calendar years, data exist for examining person-level changes in selected variables, such as expenditures, health insurance

Table 1. MEPS Household Component: Annual Sample and Panel Specific and Pooled Annual Response Rates

Year	Families	Persons	Yr1, Current Panel R1-R3 x NHIS (%)	Previous Panel (Yr 2) R3-R5 x NHIS (conditioned on response to R1-R3) (%)	Final Pooled Annual Response Rate (%)
1996	8,655	21,571	70.2	–	–
1997	13,087	32,636	69.2	63.5	66.4
1998	9,023	22,953	70.8	65.0	67.9
1999	9,345	23,656	65.5	63.1	64.3
2000	9,500	24,000	68.3	63.7	65.8
2001	13,500	35,000	66.8	65.4	66.1
2002+	15,000	39,000	–	–	–

coverage, and health status from the two-year panel of the MEPS. For example, researchers can assess the persistence of high health care expenditures by examining whether individuals with high expenditures in one year have high expenditures in the following year or shift to a higher or lower expenditure level percentile. The MEPS panel design also allows the assessment of the impact of survey attrition on the resultant survey estimates. Specifically, analysts can compare the national health care estimates produced from the first year of a sample panel (with a higher response rate) with the estimates derived from the second year of a MEPS sample panel (with a lower response rate) covering the same time period. Further, with the linkage of MEPS and NHIS files, longitudinal analyses of transitions in health insurance coverage and health status characteristics over a three-year period for the total population and population subgroups are feasible. The longitudinal design of the MEPS also provides a foundation for estimating the impact of changes affecting access to insurance or medical care on economic groups or populations of interest.

ESTIMATION STRATEGIES TO REDUCE NONRESPONSE BIAS & SUPPORT ANALYSES OF TRENDS IN HEALTH EXPENDITURES

In panel designs with multiple rounds of data collection, the overall survey response rate is a multiplicative function of the round-specific response rates. To produce annual health care and expenditure estimates for a full calendar year, data from the first three rounds of MEPS data collection for the given calendar year (i.e., current panel) are pooled with data collected in Rounds 3 through 5 of the previous panel (i.e., Year 2 of the previous year's panel). The response rates calculated for the MEPS annual estimates likewise follow this overlapping panel design. Response rates are calculated separately for each panel with the response rate at the end of the appropriate round for the specific panel factoring in the response rate in previous rounds. The panel-specific response rate also includes the NHIS response rate. To obtain the overall annual response rate, a pooled response rate is calculated by taking a composite of the panel-specific response rates. Panel-specific and pooled (combined) annual response rates by year are shown in Table 1. The response rates reflect response to both the NHIS and the MEPS interviews. Response rates are highest for the first year of a new panel (about 66–71%) and lowest for the second year of the previous year's panel (about 63–65%). The overall response rate for annual MEPS estimates averages about 66%.

In MEPS, analyses are conducted at both the person level and the family level; thus, both person-level and family-level weights are produced. To support the unique feature of the MEPS for longitudinal analyses, longitudinal weights also are produced. The adjustment strategy used to compensate for survey nonresponse includes (1) an adjustment for dwelling unit (DU) nonresponse to account for household nonresponse after Round 1 among those households subsampled from NHIS for

inclusion in MEPS and (2) a nonresponse adjustment to account for survey attrition at the person level. Based on previous analysis (Cohen & Machlin, 1998), the following variables (from the NHIS) available for MEPS responding and nonresponding DUs were determined as the most important in reducing bias in the survey estimates resulting from nonresponse and are used in forming the DU nonresponse adjustment classes:

- Age, sex, race/ethnicity (reference person)
- Marital status (reference person)
- Employment classification of reference person (item nonresponse)
- DU level personal help measure (limitations)
- Propensity to cooperate: telephone number provided during NHIS interview
- Size of DU
- Family income
- Metropolitan Statistical Area (MSA) size
- Census region

To inform the nonresponse adjustment strategies to correct for survey attrition, previous studies identified the characteristics that distinguish MEPS survey participants across waves from those that participate only in initial rounds and then discontinue their survey participation. The prior study findings revealed that nonrespondents in the first round of the survey were more likely to be from single- or two-person households located in large metropolitan areas with a higher level of income and were more likely to include healthy elderly members (Cohen & Machlin, 2000). Reluctant respondents in the first round of the survey were significantly more likely to become nonrespondents in the second round. As with nonrespondents in the first round, MEPS nonrespondents in subsequent rounds were more likely to reside in large metropolitan areas. They also were more likely to reside in households with five or more members, be elderly, and be either married or separated relative to individuals who were never married. These findings informed the specification of weighting class adjustments to compensate for person-level nonresponse in the survey (Cohen, DiGaetano, & Goksel, 1999). The variables used to adjust for survey attrition in the MEPS after Round 1 include an indicator for initial refusal to the Round 1 interview, family size, age, MSA, marital status, race/ethnicity, and sex.

One of the primary analytic advantages of a panel survey is the ability to conduct longitudinal analyses on variables for the sampled units measured at different time periods. To facilitate longitudinal analyses with the MEPS, a special longitudinal weight is constructed for each panel. For example, the two-year longitudinal file for MEPS Panel 4 contains a weight applied to those persons who participated in both 1999 and 2000 to allow analysts to examine national estimates of person-level changes in selected variables.

DESIGN & ESTIMATION CONSIDERATIONS FOR THE MEASUREMENT OF HEALTH CARE EXPENDITURES

Health care expenditures represent nearly one-seventh of the U.S. gross domestic product. Findings from the 1996 MEPS show that the top 1% of the health care expenditure distribution was associated with 27% of the total health care expenditures incurred by the civilian noninstitutionalized population (Berk & Monheit, 2001). Further, the top 5% percent of the population by magnitude of health care expenditures accounted for 55% of the total. Thus, additional attention and prioritization has been given to data collection procedures and estimation strategies to help improve the quality of the survey estimates that characterize this policy-relevant subgroup. First, there is a prioritization employed in the fielding of the medical provider sample to prioritize efforts to enhance response rates for the sample associated with decedents and other cases likely to incur high levels of medical expenditures (e.g., cases with inpatient care and long lengths of stay). Second, the MEPS population estimates of decedents are poststratified to national mortality counts, and comparable adjustments are implemented for individuals entering nursing homes in a given year. Then, in the

expenditure imputation procedure, donor records are required to match on decedent status for event-level records with missing expenditure data. Adopting these special procedures for the subgroup characterized by high levels of medical expenditures improves the accuracy of the overall national expenditure estimates.

The longitudinal design feature of the MEPS provides the capacity for both methodological and analytical evaluations. For example, panel-specific and longitudinal data can be used to evaluate the quality of the highly skewed expenditure estimates. Beginning with the 1997 MEPS, national estimates for a given calendar year can be derived from the following four data files:

(1) Full year (FY) file that combines the second year of a given MEPS panel with the first year of a new MEPS panel. This file has the largest sample size (~22,000 in FY96; ~33,000 in FY97; and ~23,000 in FY98);
(2) Panel-specific file for the first year of a new panel. This file (PUF97 and PUF98 in Figure 2) has the highest survey response rate with response rates of about 69–71%;
(3) Panel-specific file for the second year of a given MEPS panel. This file (Long. 96 [Year 2, Panel 1]), Long. 97 [Year 2, Panel 2] in Figure 2) has the lowest response rates of about 63–65%; and
(4) First year of MEPS longitudinal file (Long. 96 [Year 1, Panel 1], Long. 97 [Year 1, Panel 2], Long. 98 [Year 1, Panel 3] in Figure 2) with response rates ranging from 63–65%.

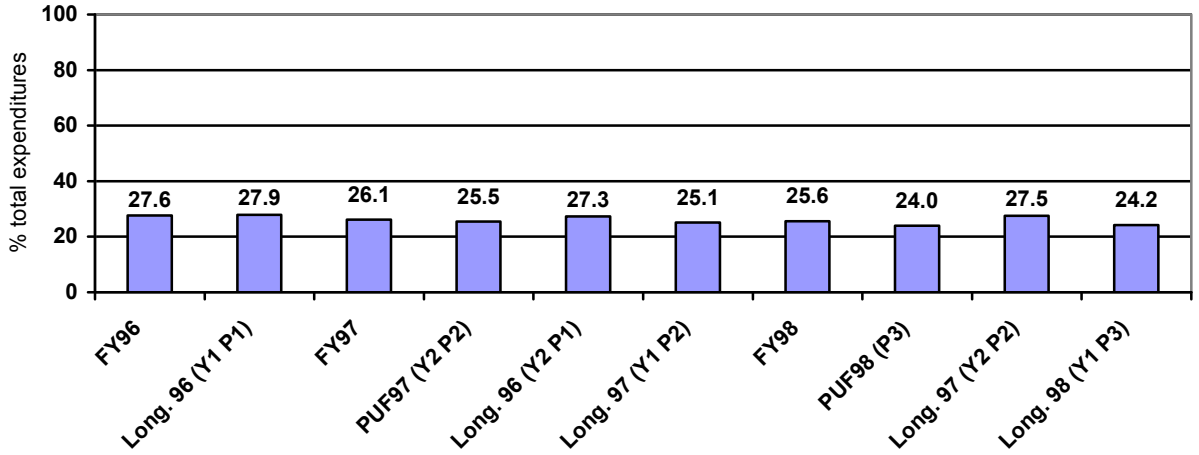
The latter two files permit longitudinal estimates over a two-year period and are subject to the lowest response rates relative to the first two files, given the five rounds of data collection that characterize the estimation time period. Figure 2 shows the percentage of total health care expenditures accounted for by the top 1% of the population as calculated from each of the files described above. The lack of major differences in estimates are noted across the files when compared to the first year of a new panel (which always will have the highest response rate) provide evidence that the nonresponse adjustment strategies employed to minimize the impact of nonresponse bias in survey estimates due to sample attrition have been effective. Further, these results provide incentive for pooling consecutive years of MEPS data to improve the precision of the highly skewed distribution of medical expenses and to expand the types of expenditure data analyses for population subgroups.

CAPACITY FOR LONGITUDINAL ANALYSES USING MEPS DATA

The nationally representative two-year longitudinal panel sample of MEPS is extremely useful for addressing a broad range of health policy issues. For example, the MEPS longitudinal panel samples can be used to develop predictive models for determining a set of correlates that best and consistently predict medical care expenditures for the total population and for population subgroups. These prediction models could be a useful statistical tool to inform oversampling strategies for ensuring adequate coverage of this policy-relevant subgroup of the population in sample surveys. This would have the benefit of improving the analytic power of MEPS for more in-depth health policy analyses for current and emerging issues related to health care use, delivery, cost containment, etc. Other examples of methodological and analytical research that can be done taking advantage of the longitudinal nature of MEPS are as follows:

- What factors predict the persistence of high drug expenditures?
- For individuals with specific chronic conditions, what factors influence spending levels from one year to the next?
- How is the burden of out-of-pocket spending for health care distributed among population subgroups when examined over a two-year period?
- How do transitions in health insurance coverage over a two-year period affect health care utilization?

Figure 2. Percentage of Total Health Care Expenditures for Top 1% of the U.S. Civilian Noninstitutionalized Population, by Year, MEPS



FY = Full Year File; Long = Longitudinal File; PUF = Public Use File; Y1P1 = Year 1, Panel 1; Y2P2 = Year 2, Panel 2
Source: Center for Financing, Access, and Cost Trends, Agency for Healthcare Research and Quality

- Does the use of certain prescription medicines one year affect the use of certain health services in the subsequent year?

SUMMARY

This paper has provided an overview of the design and estimation strategies employed in the MEPS to ensure the accuracy and quality of a key analytic component of the survey: the health care expenditure estimates. This paper also has discussed methodological as well as current and emerging health policy issues that can be investigated using the two-year nationally representative longitudinal panel samples in MEPS. Specifically, use of MEPS data to identify persistence of high levels of health care expenditures is an important health policy and statistical sampling tool. While limitations may exist in the evaluation of expenditure persistence and transitions from one year to the next and which may not represent what might happen for a longer time period, the MEPS is nevertheless a valuable resource for the study of health care and cost issues and to inform health care policy and practice. As additional panels of data become available, the analysis potential will likewise expand and through an ongoing research and analysis program, new survey design and estimation innovations can be identified to further enhance the quality and analytical utility of the resultant MEPS survey data.

REFERENCES

Berk, M. L., & Monheit, A. C. (2001). The concentration of health expenditures, revisited. *Health Affairs, 20,* 9–18.

Botman, S., Moore, T., Moriarity, C., & Parsons, V. (2000). Design and estimation for the National Health Interview Survey, 1995–2004. *Vital and Health Statistics*, Series 2, No. 130. Hyattsville, MD: National Center for Health Statistics.

Cohen, S. B. (2000). *Sample design of the 1997 Medical Expenditure Panel Survey Household Component* [MEPS Methodology Report No. 11, AHRQ Pub. No. 01-0001]. Rockville, MD: Agency for Healthcare Research and Quality.

Cohen, S. B. (2003). Design strategies and innovations in the Medical Expenditure Panel Survey. *Medical Care, 41*(7), 5–12.

Cohen, S. B, & Machlin, S. R. (1998). Nonresponse adjustment strategy in the household component of the 1996 Medical Expenditure Panel Survey. *Journal of Economic and Social Measurement, 25,* 15–33.

Cohen, S. B., & Machlin, S. (2000). Survey attrition considerations in the Medical Expenditure Panel Survey. *Journal of Economic and Social Measurement, 26,* 83–98.

Cohen, S. B, DiGaetano, R., & Goksel, H. (1999). *Estimation procedures in the 1996 Medical Expenditure Panel Survey Household Component* [MEPS Methodology Report No. 5, AHRQ Pub. No. 99-0027]. Rockville, MD: Agency for Healthcare Research and Quality.

FEATURE PAPER: . Estimating Trends in Substance Use Based on Reports of Prior Use in a Cross-Sectional Survey

Joseph Gfroerer and Arthur Hughes, Substance Abuse and Mental Health Services Administration
James Chromy, David Heller, and Lisa Packer, RTI International

INTRODUCTION

Substance use trends in the United States have shown dramatic shifts since the 1960s. Among youths age 12 to 17, the rate of past-month marijuana use was less than 2% in the early 1960s, increased to 14% by 1979, then decreased to 3.4% in 1992 before rising to 8.2% in 1995. Major shifts in prevalence at different points in time and for different age groups have been observed for other substances, including cocaine, LSD, Ecstasy, opiates, cigars, and cigarettes (SAMHSA, 2003). Accurate measurement of these trends is critical for policy makers making decisions about targeting limited resources efficiently toward emerging problems. Trend data are also used for assessing the impact of prevention and treatment programs.

The typical method used for measuring substance use trends is comparing prevalence estimates across repeated cross-sectional surveys. An alternative approach is to collect data about prior substance use within a cross-sectional survey and construct prevalence estimates for prior years based on these data. Besides the cost advantages, these retrospective estimates have some analytic advantages. When data are obtained for different periods from the same respondents, trend analyses are more powerful, due to the positive correlation between estimates, as is the case in a longitudinal study. Retrospective estimates also may be the only alternative if estimates are needed for periods for which direct estimates are not available. However, retrospective estimates do have important limitations. Bias due to recall decay, telescoping, and reluctance to admit socially undesirable behaviors could cause underestimation or distort trends (Johnson, Gerstein, & Rasinski, 1998; Kenkel, Lillard, & Mathios, 2003). Bias also could result from coverage errors affecting the capability of the sample to represent the population of interest for prior time periods, due to mortality, immigration, or other changes in the population.

This paper discusses several types of retrospective estimates and presents analyses of data from the National Survey on Drug Use and Health (NSDUH) to assess biases in these estimates.

METHODS

Description of NSDUH

The NSDUH collects data in face-to-face interviews, employing self-administration for substance use questions. Formerly called the National Household Survey on Drug Abuse (NHSDA), it was conducted periodically from 1971 to 1988 and annually since 1990. The survey covers the civilian noninstitutionalized population age 12 and older in the U.S. Annual sample sizes were below 10,000 prior to 1991, 17,000–33,000 during 1991–1998, and about 70,000 after 1998. Methodological changes to the survey in 1994, 1999, and 2002 affected prevalence levels, limiting comparability across time periods. Based on a split sample in 1994, adjustment factors have been developed to improve comparability of 1974–93 data with 1994–98 data (SAMHSA, 2000).

Retrospective Estimates Produced from NSDUH

Three types of retrospective estimates were assessed, the first being "incidence" estimates, reported as the number of persons who used a substance for the first time during a year. These estimates are based on responses to the question "How old were you the first time you used [substance]?" If first use was recent (age at first use equal to or one less than age), the month and year of first use was ascertained. Combining these data with the interview date and respondent's date of birth, a date of first

use is determined for each respondent who has ever used the substance (Packer, Odom, Chromy, Davis, & Gfroerer, 2002).

The second type of retrospective estimate is rates of lifetime use of each substance. These estimates also are based on age at first use data and give the percent of the population in a specific year that had ever used the substance at that time. Because NSDUH interviews take place nearly uniformly throughout the year, the estimates reflect an average prevalence over the entire year (SAMHSA, 2003).

The third type of retrospective estimate is based on a new question introduced in the NSDUH for the first time in 2003:

> *Earlier questions were about the past 12 months. This question is about the year before that, that is, from [date1] to [date2]. During that year, beginning [date1] and ending [date2], did you use marijuana or hashish, even once?*

Information from this question can be used to make estimates of current use for the prior year but also provides "longitudinal" information on patterns of continuation or quitting among marijuana users.

Analysis Approaches for Assessing Retrospective Estimates

Several analyses were undertaken to assess bias in retrospective estimates:

Impact of immigrants. One known source of bias in retrospective estimates is the inclusion of data from immigrants who were not living in the U.S. in some prior years. We compared estimates of incidence and lifetime use (for those age 12–17 and 18–25) for the full 2002 NSDUH sample with estimates based on the sample excluding these immigrants, according to questions on country of birth and years in the U.S.

Long-term trends in incidence. Trends in incidence estimates for 1965–1990 based on 1991–93 data (shortest recall), 1994–98 data, 1999–2001 data, and 2002 data (longest recall) were compared. For the 1991–97 period, trends based on the 1994–98 data, 1999–2001 data, and 2002 data were compared. Consistency was assessed through visual inspection of curves and with correlations. Because of methodology changes, comparisons of levels from different sets of surveys were not made.

Long-term trends in lifetime use. We compared 2002-based retrospective lifetime use estimates (excluding immigrants) with direct lifetime use estimates from earlier NSDUHs (for those age 12–17 and 18–25) and from the Monitoring the Future (MTF) Study, a survey of high school seniors (Johnston, O'Malley, & Bachman, 2003). To reduce the effect of sampling error, we combined several years of data, depending on availability, and generated average annual lifetime prevalences for specific time periods. Because 1999 and 2002 survey changes resulted in increased reporting of lifetime use, we expect retrospective estimates to be greater than the direct estimates for years before 1999.

Short-term trends in lifetime use. We compared retrospective lifetime use estimates for 2002, based on 2003 NSDUH data (first six months of data currently available), to direct 2002 lifetime use estimates, from the 2002 NSDUH (first six months of data, for consistency). Comparisons were made for 19 substances for those age 12–17 and 18–25.

Retrospective prior year annual marijuana use. To assess the accuracy of these estimates, we compared January–June 2003-based retrospective estimates of past year use in January–June 2002 to direct past year estimates from the January–June 2002 data, by age group.

RESULTS

Impact of Immigrants

Including immigrants results in small bias for most retrospective incidence estimates. For example, marijuana incidence estimates for 1965–2001 were 2.5% higher when immigrants were included. For most other illicit drugs, the bias was smaller, indicating that very little initiation for these drugs occurs among immigrants prior to their entry to the U.S. However, biases for alcohol and cigarette incidence estimates were larger (8% for alcohol, 7% for cigarettes). In general, they

were largest for the years 1979–1994 (3.5% for marijuana, 11% for alcohol, 10% for cigarettes) and smallest for years after 1997 (1% for marijuana, 3% for alcohol, and 3% for cigarettes).

For lifetime prevalence rates, bias due to including immigrants is negative for nearly every substance because of the low rates of substance use among immigrants. For youth estimates during the period 1979–1990, the inclusion of immigrants resulted in biases of about -14% for marijuana, -15% for cocaine, -9% for cigarettes, and -9% for alcohol. Bias was generally worse for estimates for those age 12–17 than for those age 18–25, and there was very little bias in any estimates for years after 1997. Estimates of alcohol use for those age 18–25 including immigrants showed very small but positive bias (1.5%) for the period 1982–1993 and a larger positive bias (7%) for 1965–81.

Table 1. Percent Differences[1] Between Direct and 2002-Based Retrospective Estimates of Lifetime Use of Selected Drugs Among Persons Age 12–17 and 18–25

	Time Period					Annual Change[3]
Substance/Age	1974-1977[2]	1979-1982	1985-1988	1990-1993	1995-1998	
Marijuana						
12–17	-23.7	-26.6**	-15.4*	6.9	10.8**	-0.85*
18–25	-9.4	-6.8	2.1	6.5*	11.1**	-0.15
Cocaine						
12–17	-76.6	-76.3**	-64.3**	-45.5**	-15.8*	-3.39**
18–25	-50.6	-46.9**	-21.7**	-18.7**	13.5*	-1.82**
Hallucinogens						
12–17	-33.9	-48.9**	-25.9	-17.3	-23.7**	-1.72**
18–25	-23.1	-30.8**	12.7	3.0	12.5*	-0.63
Inhalants						
12–17	-66.8	-55.3**	-54.6**	-49.6**	-26.4**	-2.89**
18–25	-41.9	-41.0**	-10.3	14.6	23.4**	-1.21
Pain relievers						
12–17	-75.3	-56.9**	-73.1**	-69.0**	-32.3**	-3.40**
18–25	-51.2	-30.5**	-21.4*	-10.1	30.0**	-1.48*
Tranquilizers						
12–17	-54.4	-70.5**	-80.0**	-72.2**	-10.1	-3.24*
18–25	-41.8	-44.4**	-36.9**	-22.7**	15.8	-1.80**
Stimulants						
12–17	-47.6	-47.3**	-56.7**	-48.9**	3.3	-2.36**
18–25	5.0	15.2*	3.99	55.1**	80.9**	1.00
Sedatives						
12–17	14.5	-15.2	-71.1**	-79.3**	-60.4**	-1.43
18–25	-22.1	-23.2**	-10.1	-25.3*	0.9	-0.97*
Alcohol						
12–17	-44.5	-50.9**	-37.3**	-32.0**	-23.0**	-2.14**
18–25	-7.5	-14.9	-10.2*	-8.7[a]	-4.6	-0.51**
Cigarettes						
12–17	-26.4	-32.4**	-30.3**	-21.7**	-5.3**	-1.37**
18–25	3.4	-10.0	-6.8*	-2.6	0.4	-0.17

*$p < .05$; **$p < .01$.

[1] Percent difference = (Retrospective - Direct)/Direct.

[2] Tests of differences were not computed for estimates presented in this column due to the unavailability of standard errors corresponding to the NSDUH direct estimates.

[3] Annual change is based on the estimated slope for a no-intercept regression model fitting % differences against years of recall counting back from 2002.

Table 2. Percent Differences[1] Between Direct and 2003-Based Retrospective Estimates of Lifetime Use of Selected Substances Among Persons Age 12–17 and 18–25 as of January–June 2002

Substance	Age 12–17	Age 18–25
Marijuana	-0.2	0.1
Cocaine	-4.2	-2.6
Crack	-37.9**	8.4
Heroin	-11.2	17.3
Hallucinogens	-18.1**	0.3
LSD	-24.5**	3.4
PCP	-13.5	14.3
Ecstasy	-16.3	-2.3
Inhalants	-19.4**	-6.0
Pain relievers	-20.3**	-2.2
Tranquilizers	-3.0	4.7
Stimulants	-11.9	0.1
Methamphetamine	-8.4	3.2
Sedatives	-20.9	-13.2
Alcohol	-4.9*	-1.2
Any cigarette	0.6	-1.5
Daily cigarette	9.5	1.7
Smokeless tobacco	-13.7**	-5.4
Cigars	-13.1**	-2.0

*$p < .05$; **$p < .01$.
[1]Percent difference = (Retrospective - Direct)/Direct.

Long-Term Trends

For each of 12 substances, the four separate data sets produced similar incidence curves. Periods of increase and decrease and maximum and minimum periods matched across data sets for virtually every substance. Most correlations were above 0.8. While there were some indications of reduced correlation with longer recall period, this pattern was not consistent across drugs.

Comparison of the retrospective lifetime estimates with direct estimates from prior surveys suggests substantial bias, which generally increases with length of recall (Table 1). Despite the expectation that retrospective estimates should be higher than direct estimates due to the increased reporting of lifetime drug use in 2002 attributed to methodological changes, for youths age 12–17, retrospective estimates for periods prior to 1985 were lower than direct estimates by 23% to 80% for every substance except sedatives. Retrospective estimates for those age 18–25 showed more consistency with direct estimates, but the percent differences also suggest significant bias for several substances, and the bias increased with length of recall. The comparisons with MTF direct estimates showed a similar pattern of increasing bias with longer recall.

The relationship of length of recall to percent differences was examined using a simple linear regression model. Assuming a recall period of 0 years for 2002 corresponds to no recall bias or a small positive bias due to the 1999 and 2002 methods changes, a "no intercept" model was used. The slope for this model is an estimate of the expected annual change in percent relative bias per each additional year of recall counting back from 2002. Fourteen statistically significant slope estimates (most for those age 12–17) were all negative, indicating a negative percent difference increasing in magnitude with increasing years of recall. Since the retrospective measures are based on the respondents' recall of their age of first use, it may be that with increasing length of recall, respondents tend to move their age of first use to an older age, resulting in a larger impact on retrospective lifetime use measures for younger age groups (Johnson & Mott, 2001).

Similar models relating retrospective measures to MTF estimates were run. Since MTF estimates typically differ from NSDUH estimates due to methods differences, an intercept was included in these models. A negative bias was indicated by the estimate of the intercept, but only two of the six substances showed a statistically significant intercept under these model assumptions. None of the six substances studied showed a statistically significant annual change in percent relative difference, but the direction and general magnitude of the estimated annual change estimates for MTF-NSDUH retrospective comparisons was generally consistent with the results in Table 1.

Short-Term Trends

Table 2 compares retrospective and direct estimates of lifetime substance use in 2002. For those age 12–17, retrospective estimates were significantly lower than direct estimates for eight of the 19 substances, while none of the retrospective estimates were significantly greater than their corresponding direct estimate. Overall, 17 of 19 substances had lower retrospective estimates, and 12 of these were more than 10% lower. Correspondence was good (1% to 5% difference) for cocaine, tranquilizers, and alcohol and was excellent for marijuana (0.2% difference) and cigarettes (0.6% difference).

Retrospective estimates of past year marijuana use corresponded very well with direct estimates. The retrospective estimate for those age 12 and older was 10.9%, while the direct estimate was 11.4%. Rates were similar at every age group, except for the youngest examined (age 12–13). Although correspondence for those age 12–17 was good (16.0% vs. 15.5%), for 12- and 13-year-olds the retrospective estimate was significantly higher than the direct estimate (4.7% vs. 2.6%). For persons age 14–15, rates were 15.9% and 16.0%.

CONCLUSIONS

The analyses of long-term trends must be interpreted with caution because some of the differences between retrospective and direct estimates could be the result of differences in survey methods. Nevertheless, the results strongly indicate that prevalence estimates for distant past years based on retrospective reporting substantially underestimate the true prevalence for most drugs. The underestimation is worse for youth estimates than for young adult estimates and is positively correlated with length of recall for most substances. Although we did not compare them with direct incidence estimates, we can conclude that the retrospective incidence estimates also are biased downward, since they are based on the same underlying data as the lifetime retrospective estimates. However, retrospective estimates appear to be a valid tool for identifying past periods of increasing and decreasing initiation and use, as well as the points in time when shifts in trends occurred.

Retrospective estimates for recent time periods (one year ago) exhibit less bias than estimates for earlier years. Estimates for young adults showed little bias, but the youth estimates were biased downward for most substances. Marijuana and cigarette estimates show the best correspondence with direct estimates, and cocaine and alcohol are reasonably good. Except for those four substances, it probably is not valid to use retrospective estimates to draw conclusions about recent shifts in youth substance use.

The inclusion of immigrants in retrospective estimates introduces bias that is not consistent across measures or time periods. For some kinds of retrospective analyses from cross-sectional data, it may be appropriate to include immigrants in the sample. However, if the purpose is to estimate some characteristic of the U.S. population for a prior point in time, persons who were not in the U.S. at that time should be excluded.

Besides immigration, other population coverage changes could bias retrospective estimates. These include mortality, emigration, entering or leaving military service, and entering or leaving prisons, nursing homes, or other institutions (the NSDUH sample excludes active military and the institutionalized). For the estimates we analyzed, the impact of these biases is probably small. First use of most substances typically occurs before age 20, so even the estimates of incidence as far back as 1965 from the 2002 NSDUH are based primarily on reports among the sample age 56 and younger, for whom mortality rates are not high enough to have a significant impact on the sample representativeness. However, studies involving older populations or longer recall periods could be significantly biased due to mortality. Similarly, undercoverage due to incarceration should be minimal for most estimates, since only about 2% of the U.S. population age 18–39 and 1% of the population age 40–54 was incarcerated in 2002 (Maguire & Pastore, 2003). While the impact of incarceration varies by demographic group

(e.g., rates are higher for males and African Americans) and probably by substance, these rates still are not high enough to account for the large differences between the direct and retrospective estimates we found in this study. Finally, changes in active military status and other institutionalization should be small enough to have little impact on these retrospective substance use estimates.

This study focused on substance use estimates based on retrospective data. However, a broader issue is whether cross-sectional surveys can obtain useful data on a variety of past history health and behavioral variables to allow more in-depth epidemiological analyses. Longitudinal data are important in research on health and health care utilization, especially for substance abuse issues. Factors associated with substance abuse occur throughout the lifetime and affect the pathways of use and consequent health problems. Early childhood personality and experiences, mental and behavioral problems, family interactions, school experiences, marriage and divorce, parenthood, aging, and employment all have been shown to affect transitions from nonuse to use, to problematic use, and to treatment and recovery and relapse (Bachman, Wadsworth, O'Malley, Johnston, & Schulenberg, 1997; Glantz & Pickens, 1992). Although longitudinal studies are generally the best way to obtain these kinds of data, they can be expensive and take many years to obtain complete data. Where feasible, less expensive and timelier cross-sectional surveys should be used to obtain longitudinal data on substance abuse and other health issues. Further study of the reliability and validity of retrospective substance abuse and other health related data is needed to guide researchers in the collection and analysis of this data.

REFERENCES /

Bachman, J. G., Wadsworth, K. N., O'Malley, P. M., Johnston, L. D., & Schulenberg, J. E. (1997). *Smoking, drinking, and drug use in young adulthood: The impacts of new freedoms and new responsibilities*. Mahwah, NJ: Lawrence Erlbaum Associates.

Glantz, M. D., & Pickens, R. W. (1992). *Vulnerability to drug abuse*. Washington, DC: American Psychological Association.

Johnson, R. A., Gerstein, D. R., & Rasinski, K. A. (1998). Adjusting survey estimates for response bias: An application to trends in alcohol and marijuana use. *Public Opinion Quarterly, 62,* 354-377.

Johnson, T. P., & Mott, J. A. (2001). The reliability of self-reported age of onset of tobacco, alcohol, and illicit drug use. *Addiction*, *96*(8), 1187–1198.

Johnston, L. D., O'Malley, P. M., & Bachman, J. G. (2003). *Monitoring the Future national survey results on drug use, 1975-2002. Volume I: Secondary school students* (NIH Publication No. 03-5375). Bethesda, MD: National Institute on Drug Abuse.

Kenkel, D., Lillard D. R., & Mathios, A. (2003). Smoke or fog? The usefulness of retrospectively reported information about smoking. *Addiction*, *98*(9), 1307–1313.

Maguire, K., & Pastore, A. (Eds.). (2003). *Sourcebook of criminal justice statistics,* 2002 (NCJ 203301). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics. [Updated online at http://www.albany.edu/sourcebook/]

Packer, L., Odom, D., Chromy, J., Davis, T., & Gfroerer, J. (2002). Changes in NHSDA measures of substance use initiation. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (DHHS Publication No. SMA 03-3768). Rockville, MD: Office of Applied Studies, Substance Abuse and Mental Health Services Administration.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2000). *National Household Survey on Drug Abuse: Main findings 1998* (DHHS Publication No. SMA 00-3381, NHSDA Series H-11). Rockville, MD: Author.

SAMHSA. (2003). *Results from the 2002 National Survey on Drug Use and Health: National findings* (Office of Applied Studies, NHSDA Series H-22, DHHS Publication No. SMA 03–3836). Rockville, MD: Author.

SESSION 1 DISCUSSION PAPER

Joseph Waksberg, Westat

PLANNING THE NATIONAL CHILDREN'S STUDY

I am not competent to comment on the medical aspects and the importance of this bold and ambitious contemplated project, but there seems to be a consensus that it will provide extremely valuable information on many sources of juvenile and early adult health conditions and probably sources of many health problems in mature adults. My comments relate to two issues involved in the operations and planning for the study: the methodology for recruiting the pregnant woman whose children will be observed, and the timing of preliminary activities.

Recruitment

Westat, the statistical research company with which I am associated, prepared a report about a year and a half ago on three possible methods of identifying pregnant women for the sample. The first involved selecting a sample of doctors, and with their help, recruiting a sample of their patients who are pregnant. The physicians or their nurses would perform the initial measurements.

The second entailed carrying out brief interviews at a representative sample of households to identify those containing women of childbearing ages and recruiting for the study those who are pregnant. The nonpregnant women who are not sterile would be contacted at regular intervals (probably quarterly) over the following three years to check on their pregnancy status, and if pregnant, would be added to the panel. Who would carry out the medical and other measurements and recontact the panel members over the life of the project was left open.

The final method involved contracting with a group of generally recognized medical "centers of excellence" to recruit pregnant women. The centers would be informed of the importance of diversity in the recruitment. The centers then would be responsible for carrying out the required medical, environmental, and social measurements, as well as recontacting the mother and child at regular intervals over the 20-year life of the study.

The first alternative did not seem practical and was dropped from further consideration. Low response rates were anticipated, as well as other sources of potential bias. In addition, the stage of pregnancy at which recruitment would take place appeared to be very erratic, whereas the study plans emphasized the need for data early in pregnancy.

The other two approaches were considered at several meetings, and there were strong differences of opinion. Not surprisingly, those with medical backgrounds generally favored using centers of excellence, while the statisticians and social scientists preferred the household sample. The arguments for centers of excellence were based mainly on their competence to perform complex medical tests and the fact that they would add intellectual resources in both planning and analysis. The proponents also believed that although the panel would not be a completely random sample, the centers could establish sufficient diversity to satisfy the goals of the study. The household sample proponents were more concerned with using a method that insured both achieving a representative sample and enrolling women at an early stage of pregnancy. The household sample also would permit measurements to be taken prior to pregnancy, although this would come at a considerable cost because they would have to be administered to most women in childbearing ages in the household sample, the majority of which would not become pregnant in the course of the study. The description of the household sample method in the Westat report did not address the issue of who would administer the medical, environmental, and other measurements over the course of the 20-year course of the study and attempt to locate

movers. Presumably this could be done through contracts with local physicians or by setting up a central medical staff and mobile units for medical and other measurements, such as used in NHANES.

An article in a July 2003 issue of *Science* reported that an outside advisory committee disagreed with federal scientists on the recruitment method and voted in favor of a representative sample. Dr. Correa indicated that a hybrid design was under consideration. I have difficulty picturing what this would look like.

I would be interested to know whether a decision has been reached on this issue, and if so, what it is. It is basic to most of the future planning. For example, it is unlikely that the medical centers will be able to provide reliable data on rural populations and possibly other groups of interest not located primarily in metro areas (e.g., Native Americans).

Timing of Activities

If the household sample is chosen, considerable advance work is needed, not so much on sample selection but on the integration of household interviews with environmental, social, and medical measurements. Plans for tracing households and persons who move also will be needed; Census data indicate that mobility rates tend to be fairly high for new parents. Arranging with physicians and other staff to perform the medical and environmental tests is likely to be time consuming. If mobile laboratories are to be used, work on their development and associated software are needed. If the project is to start in 2005, planning of operations should begin almost immediately.

I cannot visualize exactly how the centers of excellence will operate, but they will have to plan the same operations as the household sample except that hospital facilities will be used, eliminating the need for recruiting local doctors or building mobile testing equipment. Somewhat less planning time probably can be tolerated. However, with this method, additional work will be necessary for quality control, in particular to make sure all instruments are uniformly calibrated and that there is standardization of all measurements among the centers. I recall that the decision to use traveling units in NHANES interviewing, examination, and testing was greatly influenced by the difficulty of standardizing measurements with local doctors. Considerable advance work will be needed and should begin soon for a 2005 startup.

DESIGN & ESTIMATION STRATEGIES IN THE MEDICAL EXPENDITURE PANEL SURVEY

The panel structure is a sensible sample design for MEPS. As Ezzati-Rice pointed out, it has the capability of providing both cross-sectional and longitudinal data (at least for pairs of years), and the cost is probably lower than choosing a fresh sample each year. The reduction in the response rate in the second year is fairly modest (3–5%), and I think it is a reasonable trade-off. With the estimation procedure used in MEPS, the variances for cross-sectional estimates should be about the same as would have occurred without the use of a two-year panel, but variances of year-to-year change are lower due to year-to-year correlations. This alone would justify the two-year panel. The ability to carry out longitudinal analyses is, of course, another important advantage.

The Census Bureau introduced rotation panels in CPS about 50 years ago, and since then, their properties have been studied fairly intensively. It is useful to list their main impacts on MEPS.

(1) The correlations over time that tend to exist among sample units usually will create reductions in variances of estimates of period-to-period change over an equal size nonpanel sample. Without special action, variance of cross-sectional data is the same for panel and nonpanel samples. With rotating panels, there is a loss in precision of estimates of averages over two or more years.
(2) It is possible to improve variance of both cross-sectional and longitudinal data through use of a more sophisticated estimation method, such as the one used in the CPS. In MEPS, if the year-to-year

correlations for identical persons are fairly high, such estimation could produce useful gains in precision.

(3) As noted in the paper by Gfroerer and his colleagues, the population of the U.S. in two adjoining years is not identical because of births, deaths, immigration and emigration, entrances and exits to military service and institutions, etc. Thus, the second year of a panel is slightly biased. Losses in population are automatically taken care of—they presumably are reported as having zero expenditures. However some types of additions are missed. My understanding is that immigrants and other persons who become part of the U.S. population in a given year and who become members of existing households are represented in the second year of a panel, but those who start their own households are missed. Intuitively, I would not expect the bias resulting from this omission to be serious. However, AHRQ should be aware of this. Perhaps it would be useful to include several additional items to the questionnaire used in the first-year panel of MEPS that identify such new additions to the population and give them a double weight. This would eliminate or sharply reduce the bias.

(4) Ezzati-Rice discussed the potential biases in the second year of a panel arising from attrition in the sample. It would surprise me if the modest reduction in response rates had an important effect. However, there is another source of difference between the two panels that could be more significant: the possibility that respondents report differently in the two years. In CPS, this is referred to as "rotation group bias." It has been studied intensively, without detection of the cause or any insight on which reports are more accurate. The phenomenon occurs in other longitudinal studies. I suggest that MEPS add this to items for future research.

On a different subject, the skewness of the estimates is startling. The fact that persons with very high expenditures mostly differ from year to year probably means that the ability to identify them in advance so they can be oversampled is very limited. I think AHRQ is right to explore the issue, but I would be surprised if an effective method of oversampling is found.

ESTIMATING TRENDS IN SUBSTANCE USE

Gfroerer and his colleagues described some interesting research designed to explore whether collection of data on substance use a year ago will improve measurement of year-to-year change. As they note, its success is partially dependent on respondents' ability to recall year-ago events.

There is a fair amount of research on problems associated with attempts to reconstruct past events. The chief concerns are probably memory loss, telescoping, and general confusion of when events in a person's life occurred. My personal experiences in research on such topics seem to indicate that the success rate in using retrospective information is highly variable. Telescoping can be controlled reasonably well, but its use is limited to situations where adjoining periods of time are involved. The SAMSHA methodology being considered is quite different. About 35 years ago, the Census Bureau explored a methodology for CPS very similar to the one described in the Gfroerer et al. paper. Respondents in a subsample of CPS were asked about their employment status in the prior month as well as the current month. The prior months' reports then were compared to the reports the same respondents gave during the prior month. There was very poor correspondence, and the research was stopped. Of course, that study was done a long time ago, and there has been considerable work by survey researchers on methods to jog memory. Also, it is likely that subject matter would affect recall ability in different ways. Intuitively, I would expect a year's recall on substance abuse would be reasonably good for long-term users but likely poor for persons who first used substances

about a year ago—that is, started any time between six months and a year and a half ago.

If the purpose of obtaining longitudinal data is to improve estimates of change over time rather than studying individuals who have changed status, an alternate way of accomplishing this is to use a sample panel approach, with partial overlap from year to year, similar to the design for MEPS, described by Trena Ezzati-Rice. The MEPS experience indicates that there would be only a minor loss in response rate. I believe the reduction in response would cause fewer problems than the uncertainty over the quality of reporting.

SESSION 1 SUMMARY

Daniel Kasprzyk, Mathematica Policy Research
Joanne Pascale, U.S. Census Bureau

INTRODUCTION

"Panel" or "longitudinal" surveys, in which individuals are tracked over time, provide rich datasets for measuring social and economic dynamics. The studies discussed in this session help to illustrate a number of general measurement issues that occur in the implementation of studies that measure change over time. Three of the studies employ a panel design. One type of panel design is a "cohort" study, in which the same group of individuals is followed over a long period of time, such as the 1946 British Birth Cohort Study and the National Children's Study (expected to begin data collection in 2006). Another uses panel data from Canada's National Population Health Survey (NPHS), along with simulated cases, to make long-term projections. A different design approach to capturing change is through a rotating panel, in which a group is followed at specified intervals for a relatively short period of time, with a new group of sample units introduced at specified intervals, as in the Medical Expenditure Panel Survey (MEPS). Yet a third approach is a series of cross-sectional surveys conducted periodically, as in the National Survey on Drug Use and Health (NSDUH).

Panel surveys, while coming into wider use recently, pose a number of challenges: Their complexity is often an impediment to analysis; the constant cycle of collection, processing, and evaluation can be exhausting to staff and take on the appearance of a "treadmill" operation; and the loss of sample through attrition can raise questions about the representativeness of the study. Authors, discussants, and session participants highlighted nonsampling error issues associated with surveys designed to measure change over time during the floor discussion.

IDENTIFICATION & MEASUREMENT OF NONSAMPLING ERROR

Coverage & Representation

The use of cross-sectional survey designs that employ retrospective questions to obtain "longitudinal" estimates must be interpreted with caution. These comparisons are made between two points in time using data obtained from a population at one point in time; in other words, the inference and comparisons made can be misinterpreted because measurements are taken and reported on essentially different populations at different times. For example, the NSDUH draws a sample of individuals to represent the civilian noninstitutionalized population in 2002 and asks questions about their drug use behavior as far back as 1968. Under this design, inferences can safely be made about the 2002 population, but the same cannot be said for the 1968 population because immigrants who moved to the U.S. since 1968 will be included in the sample in 2002. In other words, comparisons of the civilian noninstitutionalized population estimates at two time points cannot easily be made because the sample in 2002 will not represent the population at an earlier date. Comparisons of cross-sectional estimates using the retrospective data will be biased if the comparisons are meant to measure change between two cross-sectional populations at two points in time. The comparisons do represent, though, the change experienced by the individuals represented in the 2002 sample.

Another challenge facing all surveys but posing special challenges for panel surveys is the recruitment of individuals in "difficult" populations into the survey and retaining these individuals over the duration of the

panel. People who are difficult to find initially—that is, those tenuously attached to a household and the mobile and transient populations—also are extremely difficult to follow over time. These populations often have higher missing interview rates and higher item nonresponse rates than other groups. The difficulty is that important analytic variables, such as having health insurance and an individual's use of drugs, often are most prevalent in these hard-to-find/hard-to-track populations. The differential nonresponse in these subgroups may result in biased estimates if the adjustment method does not adequately compensate for the survey's inability to locate these people. The data collection's cost to retain these populations in the sample can be prohibitive. Multiple follow-ups, continuing efforts to locate the individuals, and extensive refusal conversion efforts can be very expensive, and so fieldwork must be stopped at some point due to budget considerations.

A proposal from the floor to consider implementing a "side survey," a "National Survey of Difficult Populations," is conceptually appealing. The idea requires the cooperation of several survey programs to work together to identify common key variables that could be used to estimate bias due to lack of coverage. The survey would be an add-on supplement to a large data collection program, such as the American Community Survey or the Current Population Survey, and would require considerable effort. It would require substantial follow-ups in the field to identify and obtain information on the difficult-to-reach populations. This type of effort is expensive, but if several interested parties contributed to the study, it may be feasible to collect a few crucial variables that could be used to assess bias through statistical models. If implemented on a regular basis, survey research and the survey programs, in particular, will benefit. While the practicality of the idea is open for debate, the idea highlights an ongoing problem in all surveys and panel surveys in particular—the potential for reduced representation of the population over time and our lack of information related to a survey's critical variables about the difficult-to-reach populations.

As in any survey, survey objectives, the target population, and the population of inference must be clearly stated, and decisions with respect to design and inference must be made on that basis. The British Birth Cohort Study recognizes the fact that it is a study of one cohort, and analyses follow from that fact. Surveys that try to achieve multiple objectives often meet theoretical and practical obstacles that make correct inferences from the survey difficult. For example, with regard to representation, the National Children's Study (NCS) is in the midst of a decision concerning the sample design—options being considered are a household-based probability sample, a "Centers for Excellence" design, or a "hybrid" design that implements the best characteristics of the two former designs. Design decisions must be informed by the survey objectives and the desired statistical inference intended to be made; this includes both the long-term and short-term analytical objectives. If the analytic objective is to produce analyses at the national level, and, in particular, for subpopulations, a nationally representative design that covers the U.S. population is necessary. Other designs may provide reduced coverage of the U.S. population but may be desirable if full coverage is not seen as critical. The point is that when initiating any study, particularly a panel study where the implications extend for years and sometimes decades, an assessment must be made of the population being covered and its relationship to the desired inference.

Questionnaire Design & Construct Measurement

Through the repeated interviewing of panel members, individual measurement errors of panel members can change over time. Many factors contribute to this, particularly the many aspects of the complex field operations. Interviewers and data collection mode can be different, and even the

respondent can be different (in surveys that accept a proxy response). Questionnaires also can be different, giving rise to context effects, even when the questions are exactly the same. In addition, the meaning and interpretation of questions and terms may change over the course of a panel survey, resulting in a failure to measure the same construct over time. So while there is a tendency to maintain standard question wording over time for comparability purposes, this can be at odds with real world change. For example, the drug Ecstasy may have been characterized and identified by several different names in the past and could very well take on different names in the future. The challenge in panel surveys is in trying to measure the same construct at two (or more) different points in time, and this may or may not require a change in question wording.

Related to this is the issue of "harmonization" of cross-sectional data over time. Apart from coverage and representation issues discussed above, appropriate methods to analyze historical data and use them meaningfully in current analyses are an important consideration for analysts. The analyst must recognize that the historic data are likely to have been collected using very different data collection methods. The differing methodologies may not be limited to questionnaire design; for example, interviewer training, mode, and other aspects of the survey conditions may vary from one survey to another. Furthermore, these types of issues may go beyond survey measurement of constructs reported by a survey respondent (e.g., technologies for measuring blood pressure have changed over time). It is important to examine differences in measurement across time to ensure that observed changes are related to the phenomenon of interest and are not an artifact of changing measurement methods.

Retrospective data pose other challenges, particularly when measuring duration of a behavior or time between events. Many analysts are skeptical of a respondent's ability to accurately report on events 10–20 years ago. Techniques such as "landmarking" or anchoring behaviors of interest (e.g., first time use of marijuana) to major life events (entering middle or high school) may serve to aid more accurate recall. Literature from cognitive psychology on memory suggests using shorter recall periods to promote more accurate recall, as is implemented in the Medical Expenditure Survey as described by Ezzati-Rice. The development of well-measured panel data requires substantial thought and consideration. Recognition by analytic staff of panel survey programs of the fact that individual measurement error changes over time will help improve the analysis of panel data; furthermore, ongoing research by panel survey programs to improve the measurement of change over time also will be helpful. An important step in this regard is to allocate more time and effort to ensuring questionnaires are well crafted and that constructs retain their interpretation over a panel's life.

Sample Retention & Nonresponse

Nonresponse is an important topic in the survey research literature, both in the development and implementation of methods to reduce it and adjustment methods to compensate for it. Operations statistics, such as the nonresponse rate, have become one of the regularly reported survey measures that indicate the quality of the survey operations. Nonresponse, however, is exacerbated and complicated in panel surveys when compared with cross-sectional surveys. Multiple interviews, complicated patterns of response, and the inevitable growth in the nonresponse rate over the course of the panel increase the significance and importance of this source of error. Not surprisingly, floor discussion highlighted these issues.

Recruitment & retention

Maintaining the representativeness of the sample by implementing field operations and procedures aimed at emphasizing the reduction of nonresponse and maximizing sample retention are critical to a successful

panel survey. The 1946 Birth Cohort Study, for example, uses annual birthday cards to stay in touch with respondents from wave to wave and to encourage them to stay in the sample. While the methods and procedures used vary from survey to survey, other methods of building rapport with the respondents ought to be explored. RTI conducted focus groups on methods of recruitment and retention for panel surveys. Participants were likely to stay in a survey if they trusted the researchers and understood the survey goals, how the data would be used, and how the data benefit their local environments. This finding was particularly prevalent among minorities. Data collection programs need to recognize the value of research on what motivates sample members to participate in panel surveys; this research will result, ultimately, in a better use of survey resources. Panel survey designs require significant presurvey planning to ensure that survey design priorities are reconciled with the analytic objectives and that the appropriate field priorities are identified to ensure sample loss is minimized.

Attrition

Panel surveys usually result in complicated patterns of nonresponse. Missing data at the sampling-unit level can result in a variety of missing data patterns. The most common missing data pattern usually is associated with a monotone nonresponse situation—where a sample unit stops participating in the survey and continues, thereafter, to be a nonparticipant for each data collection cycle. While all panel surveys are subject to this type of nonresponse, the effect can be exacerbated in certain situations. For example, in studies that measure exposure to harm and the effects of that exposure, sample members may be identified as being exposed and at risk from those harms and are subsequently notified of that risk. The notification of being "at-risk" may cause the sample members to stop participating in the survey or to change behavior to avoid the risk. The change in the behavior of the respondent will introduce bias into change estimates or estimates of the results of being exposed. Statistical adjustment methods to correct for their nonparticipation or change in behavior are needed.

Nonresponse adjustments

Attrition is the most typical pattern of unit nonresponse in a panel survey. Other patterns of nonresponse are generated by sample cases not participating in one or more waves of data collection but not dropping out of the panel forever. As in all surveys, there is the ever-present need to understand how respondents and nonrespondents differ and how the observed differences affect estimates obtained from the survey. While the additional layer of nonresponse generated from one or more missing waves of data poses additional challenges, it also provides opportunities for more sophisticated adjustments than those available in cross-sectional studies. Statistical procedures that account for the patterns of missing data and use information available from the waves of collected data can be used to develop statistical models that adjust survey weights to compensate for the various patterns of panel nonresponse. The MEPS study provides a good example of a data set that could benefit from such an approach. The study now reviews information from all five rounds of data collection and uses the information to adjust sample weights for respondents who participate in all five waves of the survey. An alternative strategy for the study to consider is the development of methods to include sample respondents who miss one or more interviews in the analytic data set by using their reported information to develop improved statistical models to account for patterns of nonresponse.

Benchmarking/Validation

Survey estimates are subject to a wide variety of sources of error. Quantifying each individual source of error is not possible. In the absence of a rigorous, comprehensive identification and measurement of the sources of error in a survey, researchers often compare survey estimates to estimates from

independent data sources. The data sources are usually administrative record data or data from other sample surveys. The key to such comparisons is the analyst's ability to identify a comparable data set or make adjustments to a data set that renders the comparisons valid. The 1946 British Cohort Study has made efforts to compare findings with data from similar studies conducted in Finland, Sweden, and Greece, while recognizing different cultural circumstances. The Canadian study is monitoring the World Health Survey (out of Geneva) in an attempt to ensure cross-cultural comparability. The study will use more current data as they become available to ascertain whether assumptions made about parameter estimates have changed; furthermore, the study encourages the development and use of alternative modules to validate results. The point is that cross-validation of estimates from multiple data sources and sensitivity analyses play a critical role in establishing the policy analytic usefulness of complex data sets. The effort should not be taken for granted and should be given a high priority by survey program managers.

FUTURE ISSUES

The increased collection of biomedical and environmental data in panel surveys raises implementation and ethical issues. Serious questions can be raised about how survey researchers efficiently collect such data, obtain cooperation, and maintain high response rates over time. Others are concerned about the ethics of funding the type of data collection that identifies "biomedical at-risk subjects." As panel designs continue to be discussed, the emerging trend to capture biomedical data and their role in panel designs requires multidisciplinary participation in the discussions.

Efficiency in the conduct of sample surveys, particularly panel surveys, is important. Survey integration has become an important design characteristic associated with the National Center for Health Statistics and Agency for Healthcare Research and Quality surveys. Taking advantage of existing large-scale survey systems and their sample cases can be an important design consideration; for example, the American Community Survey (ACS) will be in the field about the same time as the National Children's Study, suggesting that the use of ACS as a vehicle for screening sample cases to identify eligible sample may be efficient. The larger issue, of course, is the consideration of options in the design of panel surveys that build on existing knowledge, data, and field structures.

RESEARCH AGENDA RECOMMENDATIONS

As in all surveys, matching the study design with the research objectives is critical. Both cross-sectional and panel surveys today face some of the same challenges: declining response rates, emerging technologies that affect the mode of data collection (e.g., decreasing access to respondents by telephone, expansions in Internet-based data collection), new legislation protecting privacy and confidentiality, and rising costs associated with all these changes. Panel surveys are faced with the additional dimension of time. Research goals may change over time in response to emerging technologies and issues, societal values, and threats to public health. These types of changes, of course, are difficult to predict, but researchers may benefit by considering survey design factors at the outset of the study that build in as much preparation for such changes as possible.

Coverage & Representation

The importance of a reconciliation of the survey objectives, research questions to be answered, target population, and the population of inference is critical. Multiple objective surveys will fall short of their goals if adequate attention is not given to a rigorous assessment of the population being covered and the desired population of inference. Issues of population coverage must be addressed prior to the conduct of the survey.

Measuring change over time can be a problem if careful attention is not given to the issue. The use of retrospective data from a cross-sectional survey to measure change between two points in time must be approached cautiously since the two points in time represent different populations. However, if such measurements are deemed important, the identification of some of the population differences can be built into the study design. For example, it may be possible to identify variables that characterize sample members as eligible or ineligible at various points in time and use that information to segment the sample cases for analysis or adjustment. Such an approach requires substantial forward planning.

Panel surveys expect reduced representation of the population over time and difficulty in reaching/finding certain subpopulations. The unknown effects on critical variables of the reduced representation can be a problem. Presurvey planning to maximize sample retention and/or identify variables that can improve statistical adjustments to the survey are desirable. As a component of maximizing panel retention, more research is needed to understand what motivates survey participation.

Construct Measurement

Questionnaire design and the measurement of survey constructs at several points in time are ongoing issues in survey research. A plan for testing the validity of construct measurement at each point in time of the data collection must be made and implemented, and sufficient time must be built into the planning process to accommodate changes. Survey program managers should develop standards or criteria to establish the nature of what constitutes adequate communication of the intent of the question to the respondent.

Nonresponse Adjustments

The difficulty of understanding the differences in characteristics between respondents and nonrespondents is a continuing problem in survey research and often dependent on the content of the survey. The complicated nature of patterns of missing data in a panel survey suggest further study is necessary to examine patterns of nonresponse across waves of data, taking advantage of panel waves in which data are available to make adjustments through weighting and imputation. The missing data problem in panel surveys is serious enough that staff and financial resources ought to be explicitly allocated to investigate nonresponse, its components, and its correlates. Special attention should be given to identifying auxiliary variables for use in nonresponse adjustment models.

Benchmarking/Validation

Complex survey data must be evaluated regularly to ensure their continuing use for policy analytic applications. Cross-validation of survey estimates with independent data sources and sensitivity analyses should be an important aspect of every survey data collection program and given high priority by survey program managers.

SESSION 2 INTRODUCTION AND DISCUSSION: Community Participation and Community Benefit

Chair: Marsha Lillie-Blanton, Kaiser Family Foundation
Rapporteur: Judith D. Kasper, Johns Hopkins University
Organizer: Lu Ann Aday, University of Texas

INTRODUCTION

While much of the interest in new forms of community participation may be driven by concern about downward trends in response rates for surveys, particularly among subsets of the population that are of great interest from a public health and health policy perspective, this is not just a technical problem. It has real implications for the quality of what we know about who we are and the magnitude and distribution of health problems we face as a nation.

Increasing survey research participation in specific "communities" has to be considered in the context of several challenges. One is the changing and greater complexity of the demographic characteristics that we associate with definitions of populations or communities—most obviously, the aging of the population, and its increasing racial/ethnic diversity. For example, in 50 years, one in two persons in the U.S. will be a person of color (U.S. Bureau of the Census, 2000). Not as obvious are the diversity within these groups and diversity on other dimensions, such as gender and sexual orientation.

If researchers wish to reach into populations and define communities, they need to understand them. There are expectations and histories within communities, and to succeed, researchers must understand both. For example, the Tuskegee experiment ended in the early 1970s, but it remains current and has real implications for the people who remember that experience (Jones, 1991). More recent instances of medical errors in the context of research (or simply the provision of services), which receive widespread publicity in today's world, serve to increase suspicion of research (Institute of Medicine, 1999). Furthermore, despite the fact that Tuskegee and other of the most harmful examples of research gone awry are medical in nature, all research gets lumped together in the minds of many people. The challenge in striving for greater community participation in the current environment is to balance the demands for scientific rigor—threats to validity and reliability—with involvement in communities to achieve the fullest participation possible.

Subject involvement in research is not a new idea. There always has been "consultation" with potential research subjects or representatives of the communities of interest in a research project. Examples include focus groups and cognitive interviews conducted for the purpose of informing researchers about respondent interpretation of questionnaire concepts and wording. In many instances, this involvement goes beyond consultation to active assistance in the form of letters of endorsement from respected community leaders or organizations to assist in gaining *entrée* and improving receptivity to research in a community. What is different in new models of community-based research is the early and continuous involvement of the community in the research process—in the most faithful adherents to this model, from defining research goals to dissemination of results.

DISCUSSION

Although certain features of research that aspire to "community participation" have been described (Arnstein, 1969; Israel, Schulz, Parker, & Becker, 1998; Lillie-Blanton & Hoffman, 1995), this still is not a well-defined entity, as evidenced by the range of papers presented in this session. The papers from the Universities of Maryland and Minnesota come closest to achieving continuous involvement

throughout the research process of racial/ethnic minority "communities" of interest. CHIS, a statewide survey in California, involved community constituents representing health professionals and research subjects from various subgroups in a somewhat conventional advisory capacity but is unique in seeking to make it possible for all communities to become end users of the data being collected. The NHANES and NSDUH are both national sample surveys. NHANES engages with geographically-defined communities due to its data collection protocol, but both of these national surveys also are engaged in trying to identify a "match" between individual respondents and a "community"—for example, American Association of Retired Persons (AARP) membership—as a mechanism to boost interest in study participation of persons 50 and older. These efforts derive from increased awareness that there are subgroups within the population at large that increasingly pose particular challenges in terms of survey research participation.

Several questions were raised concerning how to achieve greater community participation, its implications for the research process, and its impact on researchers and communities that participate in the research process.

Does Community Participation Interfere with the Scientific Process?

Some evidence suggests that involvement of communities from goal setting through dissemination can improve the science. There are opportunities for qualitative insights that derive from collaborative relationships that cannot be obtained from quantitative data alone and that provide access to rich information on social context that may contribute in substantial ways to interpretation of results based on a sample of surveyed individuals. Community involvement in "agenda-setting" (e.g., articulating research goals) also may have greater potential for actionable results at the community level, one important goal of much public health research. As documented in the paper on the National Survey of Drug Use and Health, the reasons older nonrespondents gave for refusing to participate, such as distrust in the legitimacy and relevance of the study, may be mitigated by fuller engagement of the study populations in the survey design and implementation process.

To date, however, while principles of what constitutes community participation have been laid out, we have relatively little empirical experience and few guidelines for when and how to apply these principles. Are there types of research in which community participation should be strongly encouraged and others in which it is not feasible? How do we know when it is or is not working? What are the limits on obligations of researchers or the community? Who defines the relevant communities, and are these a function of the greatest need, the most resources, or the loudest voices? If community participation represents a new methodology for conducting health surveys, we need to describe its essential components, differentiate it from existing approaches, and begin to develop an empirical base to form criteria for good science under this model.

Does Community Participation Require Conducting Research Differently & in Ways That Create Tension for Academic-Based Researchers?

From the experiences in Maryland and Minnesota in particular, it seems clear that community involvement in all phases of the research process changes it (this should not be a surprise for those of us trained in the social sciences!). By yielding control over aspects of the process and involving more people with varying levels of sophistication and knowledge about research, many aspects of the research process that otherwise could be taken for granted are open for discussion (or negotiation), and the time and energy commitment increases for all involved. Additional challenges for researchers are the potential for greater risk to academic advancement if research products are slowed or reduced and the possibility of clashes with established procedures for conducting

research in university (or government) environments. These are issues that need attention if community participation in its fullest sense is to spread beyond small-scale boutique studies.

Are There Rewards from Community Participation for the Communities That Are Subjects of Research? Does Community Participation Fall Short of Partnership?

Many studies can point to benefits to individual community members from participation in health survey research projects. Individuals may gain temporary employment and learn new skills that translate into other employment opportunities. In addition to individual social capital, community participation has the potential for increasing social capital at the community level by developing community leaders or by contributing to the social cohesion of groups and enhancing their ability to participate in decisions that affect community members (e.g., policy advocacy).

Whether other more concrete benefits accrue may depend on the project. The project in Maryland developed a culturally competent evaluation instrument to be used in mental health settings. Use of this tool should benefit mentally ill minority individuals by more accurately reflecting their needs, thus improving services. In the Minnesota study, the quality of information about barriers to access among minority Medicaid enrollees undoubtedly was improved, and this information was made available to state policy makers, but whether the minority community experienced improved access as a result of the research was not in the hands of the researchers. The CHIS did not aspire to directly alter the variety of health circumstances being studied; rather, the goal is to provide communities with the raw materials to begin to effect policy changes.

The CHIS model does raise the issue of data privacy, and in this study, considerable effort goes into insuring individual anonymity. A full-time confidentiality manager is employed who works with programmers and statisticians to make sure no individual can be identified in micro data files. Interestingly, the only demand for identification of individuals (which was refused) came from health professionals (some public health officers at the county level), not community representatives.

There also are a few examples of community participation initiated by the subject community rather than researchers. The most prominent example is the HIV/AIDS community. The HIV Cost and Services Utilization Study (HCSUS), which focused on HIV/AIDS health care use and costs, was undertaken after major battles had already been fought about involvement of patients in medically-related HIV/AIDS research (Agency for Healthcare Research & Quality, 2004). The HCSUS study included paid community representatives on all study-related committees. This example deals with a community that was impacted by a particular disease and demanded a say in the research process. Not all "impacted" communities are as vocal, educated, and empowered, however.

The question of what constitutes "community benefit" is not without controversy. As one audience member stated, "Does participation mean partnership? If it is a partnership, it is one way. One partner gets paid and promoted, the other doesn't." There are signs that some funders are beginning to recognize the importance of paying community partners, but this is one more area where guidelines are important.

Is Community Participation Primarily a Means to Address Declining Response Rates & Low Participation Among "Difficult Populations," or Is There "Something More" to Be Gained?

A recurring theme concerning the benefit to researchers of the community participation model is that of increasing survey response rates. A decline in response rates among minority communities is of particular concern. Viewing community participation only as one more tool for gaining research access and increasing survey participation runs the risk, however, of limiting this methodology to

racial/ethnic minorities and other difficult populations. It also raises additional questions about the basis for a true "partnership" and the potential limits of the commitment by researchers to a community participation model. While more extensive consultation and involvement of community members may be effective as a means of increasing response rates, it may not be wise to label this as an application of the "community participation model."

At the present time, the importance of community participation in health survey research is unclear. Proponents argue that greater community involvement can improve the quality of the information obtained and increase the chances that the findings will be used for policy and planning purposes. Nonetheless, as one presenter noted, "a vantage point from ten years in the future" likely will provide answers about what is gained from greater community participation in the research process.

REFERENCES

Agency for Healthcare Research and Quality. (2004). *HCSUS: HIV cost and services utilization study.* [Fact sheet.] Retrieved March 1, 2004, from http://www.ahcpr.gov/data/hcsus.htm

Arnstein, S. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners, 35* (July), 216–224.

Institute of Medicine (IOM). (1999). *To err is human: Building a safer health system*. Washington, DC: The National Academies Press.

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.

Jones, J. H. (1991). *Bad blood: The Tuskegee Syphilis Experiment.* Chicago: Independent Publishers Group.

Lillie-Blanton, M., & Hoffman, S. (1995). Conducting an assessment of health needs and resources in a racial/ethnic minority community. *HSR: Health Services Research, 30*(1), 226–236.

U.S. Bureau of the Census. (2000). *Projections of the resident population by race, Hispanic origin, and nativity: Middle series, 2050 to 2070.* Washington, DC: U.S. Department of Commerce.

FEATURE PAPER: Community Participation and Community Benefit in Large Health Surveys: Enhancing Quality, Relevance, and Use of the California Health Interview Survey

E. Richard Brown, UCLA

INTRODUCTION

Participatory research has been recognized as an important method to increase accountability of university researchers to the communities they study, to enhance the relevance and quality of research intended to improve population health, and to empower communities to improve the conditions that affect health (Israel, Schulz, Parker, & Becker, 1998; Minkler & Wallerstein, 2003). Often called "community-based participatory research" or "community-based research," it is a "collaborative approach to research that equitably involves…community members, organizational representatives, and researchers in all aspects of the research process" (Israel et al., 1998). Proponents emphasize the role of community members as agents of change and see participatory research as a means to inform and support problem-solving action by those affected. Participatory research has been shaped by a number of traditions from Kurt Lewin's "action research" to Latin American liberation movements influenced by Paulo Freire (Minkler & Wallerstein, 2003).

Some of the key elements that characterize participatory research include recognition of community (a geographic community or a community of identity) as a relevant research partner; building collaborative partnerships in all phases of the research (from problem definition to data collection to dissemination of results); gathering information that informs action to improve health; disseminating findings and knowledge gained from the research to all partners involved; and contributing to the capacity of community members to work together to improve health (Israel et al., 1998). Participatory research methods have been applied mainly to local community studies, the scale of which enables researchers and community leaders to engage in face-to-face collaboration in planning and conducting studies, analyzing data, and developing publications.

Participatory research methods have not often been applied to large-scale health surveys. Community organizations and leaders seldom are involved in planning and developing sample design and content for surveys sponsored by government agencies or private organizations. Large government agencies conduct surveys to meet specific policy research goals; they also are usually more accustomed to collaborations among government agencies and top-down decision making than partnering with communities to determine who and what is surveyed. Private organizations usually sponsor such surveys to develop population-based information to meet specific research goals, rather than to provide data for community organizations and local agencies.

The development and implementation of the California Health Interview Survey (CHIS) has applied participatory research principles on a scale not previously seen in survey research. CHIS is a large biennial health survey, with a sample of more than 57,000 households in 2001 and more than 42,000 households in 2003. It is conducted as a collaborative activity of the UCLA Center for Health Policy Research (the Center), the California Department of Health Services (DHS), and the nonprofit Public Health Institute (PHI). These partners created a structure and process that involves a broad range of constituencies that participate in multiple phases of the survey. In fact, CHIS's participatory research elements are key components of CHIS's mission to be a valuable public service that is accountable and responsive to community needs.

This paper describes a participatory model for large health surveys, including the limited-participation planning project that led to CHIS, the substantial participation that is involved in planning each CHIS cycle, and the extensive dissemination of data and results back to participating constituencies.

DEVELOPMENT OF CHIS

CHIS was the product of a three-year planning project that included, in addition to a technical assessment component, outreach activities that are consistent with the limited participatory models found in many public health needs assessments (Soriano, 1995). The planning project obtained input from many state and local public health agencies, health care organizations, and advocacy groups through questionnaires sent to potential data users, public meetings, and key informant interviews. Project staff used the outreach results to shape CHIS (Public Health Institute, UCLA Center for Health Policy Research, and California Department of Health Services, Center for Health Statistics, 1997).

The sample design, for example, reflects this input. Many public health and advocacy responders placed a high priority on data for the state's ethnically diverse population and on specific smaller ethnic groups. In addition, local health departments and locally based advocacy groups, as well as many at the state level, emphasized the need for data on small geographic areas. In response, the CHIS sample was designed to yield estimates for most counties in the state and for the state's major ethnic groups and a number of smaller populations of color.

Thus, the planning project followed a needs assessment model with community participation limited to providing input that was structured, analyzed, and used by the project staff. Nevertheless, its results were instrumental in determining the survey's breadth of content and topics, its sample design and size, the populations and geographic areas sampled, and the frequency of data collection.

CHIS PARTICIPATORY RESEARCH MODEL

At the conclusion of the planning project, a new and extensive ongoing participatory process was created to develop policy and content for the new survey. A formal collaboration was established among the three planning organizations that comprise the Governing Board, with the Center as the lead organization and DHS and PHI as partners. The CHIS principal investigator, Governing Board, and CHIS team bear responsibility for designing and managing the survey and dissemination of data, coordinate the process for participation by other components shown in the conceptual model, and raise the more than $12 million required to implement the planning process, conduct the survey data collection, and disseminate data and results for each two-year cycle.[1]

Participatory Planning

The major avenues for participation occur in the planning and development of each CHIS cycle through the CHIS Advisory Board and Technical Advisory Committees.

CHIS Advisory Board

The Governing Board created the CHIS Advisory Board to provide ongoing policy guidance for all aspects and phases of the survey. The Advisory Board, which includes more than 20 members, was chaired initially by the director of DHS and now by the Governor's cabinet-level Secretary for Health and Human Services. It is comprised of directors of three statewide health agencies; CEOs or presidents of statewide associations of local health departments, community clinics, rural health care providers, the medical profession, public health professionals, hospitals, and health plans; the research arm of the legislature; and advocacy organizations for populations of color and low-income populations.

The Advisory Board meets quarterly and recommends issues to address and topics to include in the survey, sampling goals, dissemination goals, and funding strategies.

[1]For more information about CHIS, visit www.chis.ucla.edu

Although the Advisory Board can only recommend policy to the Governing Board, if the Governing Board does not take that advice in any particular instance, it reports back to the Advisory Board at a subsequent meeting why it did not do so. Although such relationships may lead to tensions over the authority and role of the Advisory Board, the responsiveness of the CHIS team to the Advisory Board has avoided such tensions.

Technical Advisory Committees

The Governing Board also established formal Technical Advisory Committees (TACs) to advise the CHIS team on specific content and measurement issues. Separate TACs are focused on the adult questionnaire, the adolescent questionnaire, the child questionnaire, sample design and survey methodology, and multicultural issues. The Multicultural Issues TAC provides advice on specific ethnic groups for which formal translations should be developed, groups that require culturally specific adaptation of the interview process, and measurement issues related to ethnicity, acculturation, and discrimination. Additional work groups are formed as needed. For CHIS 2001, more than 100 individuals from 54 separate scientific, professional, advocacy, and community-based organizations participated in the Advisory Board and five TACs, and another 20 people participated in the more narrowly focused work groups. The planning for CHIS 2003 was the same, but the numbers of organizations and individuals were larger.

Although the TACs by definition deal with technical issues, they incorporate a broad basis of expertise, including data-using staff of advocacy, public health, and health care delivery organizations, as well as researchers affiliated with major universities and research organizations. The TACs meet twice during the planning process, with CHIS paying the transportation costs of those who must travel to attend. As with the Advisory Board, TACs play an advisory role to the CHIS team, but this limitation has not been a barrier to recruiting and retaining TAC members.

Funders

Compared to the CHIS Advisory Board and TACs, major funders play a more determining role in shaping the survey. The survey's high cost makes it dependent on the decisions of government agencies and foundations to provide support to the survey overall or to specific components of content or sample. The commitment of the state to fund only a quarter of the costs meant that multiple funders could have substantial and potentially conflicting influence. Both CHIS 2001 and 2003 received more than one million dollars from each of five to six funders, and both surveys received substantially less than one million dollars each from another five to six funders.

The dependence of CHIS on multiple major funders clearly reduces the autonomy of the survey team to be entirely responsive to the CHIS Advisory Board and TACs. However, because DHS is one of the partners as well as the core funder, its commitment to support the participatory process and to involve its own program and research staff in that process provides substantial latitude for the CHIS team to respond to the broader user constituency within resource limits. In addition, the CHIS team seeks funding specifically to support the topic areas and population groups identified as priorities by the CHIS Advisory Board and TACs, and all funders are required to contribute to overall content and sample objectives.

RECONCILING INTERESTS: EXAMPLES OF THE CHIS PROCESS

Despite the potential for conflicting needs and interests, there has been considerable congruence between the wishes of the Advisory Board and TACs, on the one hand, and the CHIS team and funders, on the other. For example, Asian and American-Indian advocates and researchers on the Advisory Board and the Multicultural Issues TAC strongly recommended oversampling Asian ethnic groups and American Indian/Alaska Natives (AIANs)—recommendations that were unanimously supported by these advisory groups. Two major funders

channeled some of their funding to support these decisions: a federal agency with research interests in both populations and a foundation that separately heard from Asian advocacy groups. The CHIS team also secured additional funding from the Indian Health Service to help support oversampling of AIANs.

The importance of "being at the table" is illustrated by groups that were not oversampled. Asian ethnic groups and American Indians were the only groups oversampled. Because large samples of Latinos and African Americans would be generated without oversampling, advisory group members did not argue for oversampling Latino or African-American ethnic groups. There also was no pressure to target predominantly White ethnic subgroups, such as Armenians or Russians, despite the high proportions of their populations that are immigrants, but there were no advocacy groups from these populations represented on the advisory groups. Thus, those who were at the table respected resource limitations and were satisfied with the anticipated outcomes, while those groups that were not at the table had no opportunity to make their views known.

In sum, the development of the sample and questionnaires for each survey cycle is determined by the CHIS team but guided by recommendations of several key advisory bodies that represent a broad range of user constituencies from government, public health organizations, health care providers, and advocacy groups at the state and local levels. The responsiveness of the CHIS principal investigator and team to these constituencies is limited, however, by available resources and the directed support of multiple major funders. Nevertheless, transparency in the decision-making process, together with explanations of constraints and reasons for decisions that varied from advisory body recommendations, generates a high level of trust between the CHIS team and the advisory groups.

DISSEMINATION OF CHIS DATA & RESULTS: RETURN ON PARTICIPATION

The payback to constituencies that participate in CHIS is the data and results that are produced from each survey cycle. Consistent with CHIS's public service mission and its participatory research model, substantial resources are invested to make CHIS results, analyses, and data available and accessible to a wide range of constituencies. The Center has developed multiple vehicles to maximize access to data, analytic tools that turn data into information, and results that make data highly relevant to key health policy issues—all designed to make CHIS data as useful as possible to a wide range of constituencies with levels of technical capacity ranging from beginner to sophisticated.

Historically, data analysis has been available only to people and organizations with significant technical capacity. Community and advocacy groups and many local health departments face numerous obstacles to using data in their policy and development work, including limited availability of relevant data, lack of technical capacity to analyze data that are available, and limited knowledge of how to apply the data most effectively to support their advocacy and funding requests.

The Center has developed multiple programs to democratize both access to data and access to the analytic tools that enable data to be applied as useful information. CHIS 2001 results and data have been disseminated to state and county health departments, policy makers, health researchers, community-based organizations, advocacy groups, and the public through publications, fact sheets, data files, and an easy-to-use online data query system—all of which show high levels of use.

Publications Using CHIS Data

Publications written for broad policy audiences and directly disseminated to them offer these constituencies easy access to CHIS results that can be adapted to policy development, advocacy, and funding proposals. The Center has published nine major policy research reports based on CHIS

2001 data, as well as a dozen four- to six-page policy briefs and an equal number of two-page fact sheets.

Publications using CHIS 2001 data are being used widely by state- and local-level policy makers and advocates to develop state and local health policy. Such publications are one way to share data and findings with community partners in participatory research processes, and their wide use builds recognition of the survey as an important public health data source.

*Ask*CHIS

One of the most innovative tools is a uniquely user-friendly online query system called "*Ask*CHIS" that provides data estimates to anyone with Internet access. Users can query the CHIS data with their own health questions and obtain detailed descriptive statistics based on CHIS data tailored to their needs. *Ask*CHIS greatly facilitates access of advocates, community-based organizations, policy makers, and public health officials to user-defined survey results for individual counties, as well as statewide and for a variety of population groups defined by race/ethnicity, income, and many other demographic characteristics. This system maximizes availability of detailed geographic and demographic data while protecting confidentiality of respondents through a statistical algorithm that suppresses specific estimates that might inadvertently lead to identification of individual respondents. Outputs are available as tables, spreadsheets, and graphs, including confidence intervals that take into account the survey's complex design. *Ask*CHIS thus offers a sophisticated analytic tool to advocates, analysts, professionals, and researchers at all levels of technical ability.

One of the goals of *Ask*CHIS was to maximize usability for basic community-level users but still provide value to advanced users. Comments volunteered to CHIS staff via a "feedback" button and a recently completed user survey have been enthusiastic, even when offering suggestions for improvements. The 3,000 registered *Ask*CHIS users have averaged 15.8 queries per user, and many have come to consider it "an awesome resource," in the words of one user, and "the trusted gold standard in policy circles," according to another (*Ask*CHIS User Survey, unpublished data). However, despite efforts to make *Ask*CHIS an easy-to-use tool for people with beginner-level technical skills, the results of the user survey suggest that *Ask*CHIS has been used disproportionately by people with midlevel and more advanced technical skills. Enhancements being developed for *Ask*CHIS are expected to make it easier to use for persons with beginner-level technical skills.

*Ask*CHIS is an example of a tool that can democratize access to technical data resources, giving less technically sophisticated community-based advocates and policy makers access to data and analytic methods that have been the exclusive province of researchers. Such data sharing with participating communities is a hallmark of participatory research.

Public-Use Files & Services for Researchers

Electronic public-use data files, including supporting documentation, are available for download free of charge from the CHIS Web site. More than 430 people have completed electronic confidentiality agreements and downloaded data files in the year and a half that they have been available. The public-use files, of course, are useful only to researchers and health policy analysts with statistical analytic skills and equipment. However, some larger advocacy groups employ or work with skilled researchers and analysts; publications using CHIS 2001 data are being developed and published by many state agencies, local health agencies, and a variety of advocacy groups.

The Effects of Multiple Dissemination Modes

In sum, promoting access to and use of CHIS publications, *Ask*CHIS, and data files shares data and results with statewide and local organizations that help plan and develop

each survey cycle. The wide use of these dissemination tools also demonstrates the relevance and importance of the survey to a broad set of constituencies. CHIS results are being used both statewide and at the local level to shape and expand public health insurance coverage, for outreach to potential Medicaid and CHIP enrollees, to track asthma rates and develop policies and programs to control asthma, to promote diabetes prevention and management, and to develop policies and programs to support children's health and development. Results also are used to identify and track disparities in health and access to care based on racial/ethnic, income, geographic, and other social characteristics. In addition, CHIS data are being used in epidemiologic research to better understand individual and environmental factors that influence health and access to health services. The availability of a number of different vehicles to disseminate CHIS data and results assures that constituencies with a variety of technical capacities benefit from the survey.

CONCLUSION

The level of community and advocacy participation in CHIS planning and development is not as extensive as the maximum levels seen in community-based studies that emphasize participatory research. However, the CHIS model offers a way to optimize that participation in large-scale health surveys, which traditionally involve top-down planning processes that reflect the views of only those agencies directly involved in sponsoring the survey. This participatory research model ensures that CHIS is relevant to the communities that participate in planning it, that it appropriately measures factors related to community needs, and that the data and results are available for use by these communities and their advocates. The CHIS model is a viable participatory research approach, consistent with the flexibility suggested by Israel and her colleagues (1998).

The advisory role of advocacy, service, and policy organizations in the planning and design of CHIS gives them a voice that shapes each survey, while the CHIS research team takes responsibility for both obtaining funding and managing the questionnaire development and data collection. The major investments in dissemination provide direct payback to the advocacy, service, and policy groups that participate in the planning process and indirect benefits to the communities and populations that participate as respondents. The use of the data for policy development and advocacy also generate a wide perception that CHIS is a valuable tool for public health that deserves the continued support required to conduct it, a byproduct that closes the circle of benefits that accrue from adapting participatory research models to large health surveys.

REFERENCES

*Ask*CHIS User Survey. (n.d.) Unpublished data.

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.

Minkler, M., & Wallerstein, N. (Eds.). (2003). *Community-based participatory research for health*. San Francisco: Jossey-Bass.

Public Health Institute, UCLA Center for Health Policy Research, and California Department of Health Services, Center for Health Statistics (1997). *California Health Interview Survey: Final report to The California Endowment*.

Soriano, F. (1995). *Conducting needs assessments: A multidisciplinary approach*. Thousand Oaks, CA: Sage.

FEATURE PAPER: Research as a Partnership with Communities of Color: Two Case Examples

Llewellyn J. Cornelius, Thomas E. Arthur, Iris Reeves, Naomi C. Booker, Oscar Morgan, Janice Brathwaite, Teresa Tufano, Kim Allen, Irma Donato, Larry Ortiz, and Lydia Arizmendi, Mental Hygiene Administration/ Maryland Health Partners Cultural Competency Advisory Group

While considerable efforts have been made to focus on the need to reduce health disparities among persons of color, there are several obstacles to the achievement of this lofty goal. In particular, language barriers, perceptions of mistrust of the health system, and challenges in collecting data from ethnic populations may make it difficult to make generalizations to these populations. At the center of this debate are questions regarding whether one should sacrifice scientific rigor in favor of obtaining data that provide a richer discussion of the experiences of persons of African, Asian, or Latin decent. This paper will provide an overview of some of the challenges that are related to the inclusion of persons of color in health services research. There will be an overview of two studies that used a participatory research model to focus on ethnic minorities and a discussion of some recommended considerations for further research in this area.

CHALLENGES RELATED TO THE INCLUSION OF PERSONS OF COLOR IN HEALTH SERVICES RESEARCH

To set the stage for discussing ways to reach out to persons of color, it is important to first talk about some of the coverage issues related to their inclusion in research. The first challenge one faces in conducting studies of persons of color is that there may be measurement error in samples of this population because of an undercount of the total population. In 1940, there was evidence of an undercount of persons of color in the U.S. Census. Darga (1998) estimates that between 1940 and 1990, the undercount of the African-American population declined from 8.4% to 5.7%. Edmondston (2002) reports that there was a higher net undercount in the 1990 and 2000 Census for Native Americans, African Americans, Latinos, and Asian Americans. Several factors contribute to the undercount, including language barriers, refusals from immigrant residents, difficulties in tracking the homeless population, difficulties in tracking mobile households, and mistrust of the government and/or of people from outside the community (Darga, 1998; Edmondston, 2002). Even though these numerical estimates are small, they tend to lead to measurement errors in population counts (Darga, 1998). This means that if there is measurement error in the population count, any sample drawn from this "population" frame is subject to measurement error. This is not a trivial matter, since the Census frame serves as the basis for a multitude of national studies.

However, even if the Census is an accurate representation of the U.S. sample, there are also survey design-related issues that may limit one's ability to target populations of color. For example, some geographic regions, such as the Rio Grande Valley near the U.S.-Mexico border, have inadequate phone coverage (Aday, 1996), which limits the effectiveness of a phone or Internet survey focusing on Latinos living in this region. Next, weather and topography can make it difficult to conduct face-to-face interviews in rural communities, in remote communities in Alaska, or among populations living on Indian reservations.

A third challenge that may be faced in collecting data from communities of color is

Acknowledgments: The authors would like to thank the other members of the Cultural Competence Advisory Group: Pat Bohnet, Selwyn Charles, Raymond Crowell, Swaran Dhawan, Veronica Giddens, Carolyn J. King, Chong P. Lee, Linda Lively, Thu Nhut Ngyen, Guillermo Olives, Joyce White, Tracee Bryant, Linda Coates, Lillian Bowie, David Brown and Gail Porter who helped to lay the groundwork for the development of this instrument.

This study was funded in part by a grant from the Annie E. Casey Foundation and a grant from the University of Maryland School of Social Work.

their mistrust of organizations and institutions as a result of historical and contemporary mistreatment. Recent English-only and immigration policies have created an atmosphere in which some immigrants and their descendents feel unwelcome (Kilty & Vidal De Haymes, 2000). This is compounded by a long history of discriminatory policies against Native Americans, Latinos, Asians, and African Americans that has left some potential survey respondents leery of visits from the government. In some cases, respondents mistrust not only the government but also academia because of the perception that academic institutions take from the community (that is, data and information) without giving back to the community (Israel, Schulz, Parker, & Becker, 1998).

As indicated above, there are several layers of barriers to the inclusion of persons of color in research, including the problem of undercoverage of minorities as well at potential nonresponse issues. This paper discusses the use of one methodology—the Participatory Action Research (PAR) model—to include persons of color in research. The PAR approach is based on the following principles (Israel et al., 1998):

- Recognizes community as a unit of identity
- Builds on strengths and resources within the community
- Facilitates collaborative partnerships in all phases of the research
- Integrates knowledge and action for the mutual benefit of all partners
- Promotes a co-learning and empowering process that attends to social inequities
- Involves a cyclical and iterative process
- Addresses health from both positive and ecological perspectives
- Disseminates findings and knowledge gained from all partners involved

In the section that follows, applications of this model are presented.

CASE STUDY 1: A CONSUMER ASSESSMENT TOOL FOR CULTURAL COMPETENCY

In July 1997, the Mental Hygiene Administration (MHA) of Maryland implemented a managed care mental health system. Maryland Health Partners (MHP) was contracted as an Administrative Services Organization to assist the MHA in managing the new "consumer driven" Public Mental Health System (PMHS). The goals were easy access and choice for consumers of mental health services. From the onset, evaluation, outcome measurement, and consumer satisfaction were considered essential aspects of the new mental health system. Cultural diversity and cultural competence also were considered to be necessary aspects of Maryland's new mental health system. It was this initial charge that led to the development of what is called the Cultural Competency Advisory Group (CCAG). The 20-member CCAG is comprised of clinicians who specialize in providing services to minority populations, consumers from ethnic or minority groups, and administrators with experience in developing services for minority populations. The role of the MHA/MHP CCAG is to provide cultural diversity and competence advice, recommendations, and assistance to the PMHS. They also are charged with the task of developing ways to examine the issue of cultural competency. The group's activities include developing statewide conferences, initiating regional training, collecting and disseminating culturally relevant data, exploring best practices, developing a train-the-trainer project, and other activities designed to improve the PMHS's cultural competence.

Between 1998 and 2001, the CCAG formulated and reviewed a battery of questions that dealt with cultural competency. With facilitation from the National MultiCultural Institute (NMCI) of Washington, D.C., the CCAG held several sessions to methodically develop items related to satisfaction and cultural competence. Broad categories were identified and utilized to

develop survey items the group felt reflected important issues to measure, such as

- Attitude—how consumers feel they are perceived;
- Communication—issues related to language, being talked with and heard;
- Treatment—use of healing practices, family involvement and spirituality;
- Personnel—availability of multicultural staff at various levels;
- Environment—perceptions of feeling welcomed by staff and agency; and
- Outreach—commitment of staff and agency to engage the community.

By March 2000, the combination of these efforts resulted in the formulation of a series of survey questions that the CCAG felt represented satisfaction and cultural competence. Validation that a good battery of items had been formulated occurred in June 2000 at a conference sponsored by the Georgetown University National Technical Assistance Center for Children's Mental Health. At an individual technical assistance session, a prominent researcher from the Research Division at Santa Clara Hospital in San Jose, California indicated that the CCAG instrument of March 2000 was basically a "good tool that just needed some tweaking." Armed with this feedback, the CCAG searched for funding and successfully applied for a grant from the Annie E. Casey Foundation to engage the services of a research consultant to assist with further development of the March 2000 instrument.

The consultants conducted a literature review to examine the relationship between the battery of items developed by the CCAG and the concepts used in the literature to describe cultural competency. Additional literature reviews performed by CCAG members identified research related to provider self assessments or student research but none on consumer assessment of cultural competency.

By October 2001, it was evident that to accomplish a rigorous examination of validity and reliability, a more extensive pilot test sample would be required. With approval from the Annie E. Casey Foundation, the grant period was extended from December 2001 to May 2002 in order to increase the respondent sampling from 50 to 250.

In December 2001, the Institutional Review Board (IRB) of the Department of Health and Mental Hygiene (DHMH) granted approval for pilot testing to examine validity and reliability and to secure feedback about the structure of the instrument from recipients of mental health services. Utilizing a training manual developed by the consultant, nine CCAG members were trained to administer the pilot test. The manual addressed issues such as engaging volunteer respondents, communicating purpose and procedures for the pilot test, discussing confidentiality and risk/discomforts, facilitating group testing, and managing collected data. Of the nine survey administrators, three were proficient in Spanish and one in Vietnamese. Using claims data from MHP, 13 of the 24 jurisdictions in Maryland geographically dispersed across the state were selected to assure a balanced regional sampling of respondents of color. These specific jurisdictions also provided the opportunity to sample respondents from urban, suburban, and rural areas. Between late January and early April 2002, pilot testing occurred at 30 psychiatric programs throughout Maryland. The empirical examination of the validity and reliability of this instrument followed the administration of a 52-item scale to mental health consumers across Maryland in January 2002. Separate analyses of the validity and reliability of the questionnaire items revealed that the scale had good psychometric properties (Chronbach's alpha = 0.92) (Arthur et al., in press; Cornelius, Arthur, Booker, Reeves, & Morgan, 2004).

Comment

Several years of CCAG effort went into developing and pilot testing an inventory examining consumer perception of the cultural competency of mental health providers. This community-centered approach resulted in an instrument that can be used to assess the cultural competency of mental health

Table 1 Participatory Action Research Activities Used by the CCAG in Developing the Cultural Competency Assessment Tool.

Recognizes community as a unit of identity. By definition, CCAG was focused on addressing the needs of the community of color.
Builds on strengths and resources within the community. CCAG used the experiences of the consumers to develop the monthly agenda and to drive the development of the instrument.
Facilitates collaborative partnerships in all phases of the research. CCAG was designed as a community/administrative partnership.
Integrates knowledge and action for the mutual benefit of all partners. CCAG used expert knowledge and information from literature reviews to support the development of the instrument.
Promotes a co-learning and empowering process that attends to social inequities. CCAG panel was made up of consumers, therapists, and administrators who shared with each other information regarding the needs of consumers and how to develop a culturally sensitive mental health practice.
Involves a cyclical and iterative process. CCAG met monthly over a 24-month period with an average meeting attendance of 12 persons. These meetings were used to seek feedback and input from the CCAG regarding how the process was moving.
Addresses health from both positive and ecological perspectives. CCAG focused on collecting information that would empower consumers to decide whether a health setting was appropriate to them.
Disseminates findings and knowledge gained from all partners involved. CCAG was included in the dissemination process (including presentations and publications of articles regarding the research).

practitioners. The CCAG did not know it at the time, but they used a process called either Participatory Action Research (PAR) or Community Based Research to guide their work.

As indicated in Table 1, the CCAG participated in several activities that are reflective of PAR, including having consumers and community practitioners drive the development of the study and empowering them to make key decisions regarding the scope of the project.

This project brings up a multitude of issues regarding how to help communities develop research agendas that are sensitive to the needs of their populations. The most significant issue is the tension between universities and the community. Some see the PAR approach as oppositional to the traditional method of research, where the University community defines the work, recruits the community to participate, studies them, and then leaves. From a PAR perspective, however, the issue is defined first by the community, and then the community invites others to join the process in a way that is beneficial to the community.

The implication of the PAR approach for health services research is that one can use such a model to conceptualize how to place communities of color at the center of the research development process. By first building a common understanding among one's peers regarding the issue that needs to be examined, one can determine the role of outsiders in the development of the activity. While this approach can be used to empower specific communities around a specific project, "community based research in and of itself does not resolve broader social issues such as racism and social inequalities" (Israel et al., 1998, p. 194). Furthermore, because this is still a new method of research, more needs to be published regarding its effectiveness to increase its adoption by scholars (Israel et al., 1998).

CASE STUDY 2: HEALTH CARE ACCESS AMONG RURAL MEXICAN AMERICANS

The present study reports on the health care access of Mexican Americans in rural Texas, focusing on residents in two *colonias* in two border counties. *Colonias*, rural unincorporated subdivisions located along the U.S.-Mexico border, are among the poorest communities in the nation (Arizmendi & Ortiz, 2004). Situated near but outside the boundaries

of most border cities and towns, they often lack potable water, sewer and drainage systems, electricity, and paved roads. The population in *colonias* is almost exclusively Mexican-American, young, and multigenerational, and there are a relatively high number of people per household.

Before the initiation of the community-based survey, several meetings were held in the Rio Grande Valley between community organizers, community-based organization and health clinic program directors, and a U.S. Congressman to articulate the needs of the residents of colonias in two counties along the Texas-Mexico border. The initial meeting began as a result of a long-standing partnership between community activists, a local elected official, and an academic consultant. The initial planning meetings revealed high interest in identifying the barriers in this community but a lack of resources for conducting the study. Following the receipt of a small grant from the University of Maryland, two researchers met with community leaders in the *colonias* over a nine-month period to identify the issues that should be examined and the optimal approach. After a review of several survey instruments, the community leaders recommended that the research adapt components of the 1994 Commonwealth Fund survey of minority health (Lou Harris and Associates, 1994). The modifications were made to reflect the cultural and geographic issues of the target population. The questionnaire was translated into Spanish by two translators using the forward and back translation techniques recommended by Zambrana (1991).

Following the receipt of approval from the University of Maryland IRB, the community-based organizations in the *colonias* hired ten administrators who were trained to conduct face-to-face interviews with respondents at their homes or places of their choosing. Each administrator was given a geographic area in which to conduct interviews. They were familiar with these areas and known in the community either for their work in the community as *promotoras* or union organizers for the United Farmworkers Union. The survey administrators fanned out over the two county areas, concentrating on specific *colonias* with which they were familiar, and they covered four regions in these two counties.

One month before interviewing began, the administrators received several hours of training on the intent of the study, maintaining confidentiality, use of the instrument, and how to locate subjects. Data collection was conducted in September 2002, and administrators used a snowball sampling method, canvassing neighborhoods, churches, community centers, and other social gatherings to locate initial respondents and referrals for other potential respondents. Each survey administrator received $15 per completed interview, and respondents were given a $15 gift certificate to a local grocery chain. Over 90% of the interviews were conducted in Spanish.

This process yielded 271 usable interviews. Following data editing and cleaning activities, the results were presented to the community for their use. This resulted in the incorporation of the study findings into congressional testimony highlighting the need for services in these *colonias* (Rodriquez, 2003).

Comment

As was the case with the first study, community participation in the design and implementation of this study was critical to its successful implementation. This required extensive and sometime protracted interactions before, during, and after the completion of the studies. It also required the development of tangible products for the community. For example, the cultural competency assessment study resulted in a series of publications and presentations co-authored by community members (Arthur et al., in press; Cornelius et al., 2004), along with their involvement in Phase II to refine the assessment tool for use in the evaluation of the Public Mental Health System in Maryland. The study of the *colonias* also resulted in a tangible product—data that were used on the policy level to advocate for programs targeting these communities (*Hispanic Health Care*, 2003). While both studies used the PAR approach to

focus on specific ethnic populations, one was based on data collected from a sample of mental health facilities, while the second study was of a purposive sample of community residents.

DISCUSSION

While this paper does not address all the issues that are relevant to including the community of color in health services research studies, it suggests that there may be ways to include these populations in research. The PAR method was presented as a way to achieve this goal, since it is based on placing the community at the center of the research design. One of the clear limitations of this approach is the ability to use it as a way to launch a national probability study of the U.S. population. This is a particularly sensitive issue because of the costs of oversampling hard-to-locate populations, such as ethnic subpopulations. One solution to this problem would involve advocating for state or local studies in communities where there are significant populations of color (e.g., the California Health Interview Survey); another possibility would be to partner with national organizations (e.g., the Urban League, the NAACP, the National Council De La Raza) in the development and marketing of a national survey. To say it another way, we just might be able to increase the response rate of surveys that target the community of color if respondents believe that such studies are of relevance to organizations about which they care.

REFERENCES

Aday, L. A. (1996). *Designing and conducting health surveys.* San Francisco: Jossey-Bass.

Arizmendi, L., & Ortiz, L. (in press). Neighborhood and community organizing in colonias: A case study in the development and use of promotoras. *Journal of Community Practice.*

Arthur, T. E., Reeves, I., Morgan, O., Cornelius, L. J., Booker, N. C., Brathwaite, J. et al. (in press). Developing a consumer assessment tool for cultural competency: The journey, challenges, and lessons learned. *Psychiatric Rehabilitation Journal.*

Cornelius L. J., Booker, N. C., Arthur, T. E., Reeves, I., & Morgan, O. (2004). The validity and reliability testing of a consumer based cultural competency inventory. *Research on Social Work Practice, 14,* 201–209.

Darga, K. (1998). *Two papers on the Census undercount adjustment: Straining out gnats and swallowing camels, the perils of adjusting for census undercount; quantifying measurement error and bias in the 1990 undercount estimates.* Retrieved November 19, 2003, from http://128.32.135.2/~stark/Census/camel.pdf

Edmondston, B. (2002). *The undercount in the 2000 Census.* Retrieved January 19, 2004 from the Annie E. Casey Foundation Web site: http://www.aecf.org/kidscount/undercount_paper_final.pdf

Hispanic health care. (2003, April 7). Congressional Record—House H2853–56 (testimony of C. Rodriguez).

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.

Kilty, K. M., & Vidal de Haymes, M. (2000). Racism, nativism, and exclusion: Public policy, immigration and the Latino experience in the United States. In M. Vidal de Haymes, K. M. Kilty, & E. A. Seigal (Eds.), *Latino poverty in the new century: Inequalities, challenges, and barriers.* Binghamton, NY: Haworth Press.

Lou Harris and Associates. (1994). *National comparative survey of minority health care.* The Commonwealth Fund. New York: Lou Harris and Associates.

Zambrana, R. E. (1991). Cross-cultural methodological strategies in the study of low income racial ethnic groups. In M. L. Grady (Ed.), *AHCPR Conference proceedings – Primary care research: Theory and methods* (pp.221–227). Rockville, MD: U.S. Department of Health and Human Services.

FEATURE PAPER: Partnering with Communities in Survey Design and Implementation

Kathleen Thiede Call and Donna McAlpine, University of Minnesota
Heather Britt, The Urban Coalition
Valeng Cha, Cha Consulting
Sirad Osman, New Americans Community Services
Walter Suarez, Midwest Center for HIPAA Education
Timothy Beebe, University of Minnesota

INTRODUCTION

Research suggests that African Americans, American Indians, and Latino Americans are more likely to distrust the medical system and health professionals than are their White counterparts in the U.S. (Doescher, Saver, Franks, & Fiscella, 2000; Kao, Green, Zaslavsky, & Cleary, 1998). Distrust may be justifiable, shaped through negative experiences in the health care system and by knowledge of egregious examples of mistreatment of vulnerable groups, such as the Tuskegee experiment. This pattern of mistrust has important implications. First, differences in trust may reflect racial/ethnic disparities in treatment. Second, mistrust may be partially responsible for well-documented disparities in health care treatment and outcomes (Institute of Medicine, 2001).

While most of the attention has been focused on distrust of the medical **system**, there is a growing body of literature indicating that there are racial/ethnic differences in levels of trust in medical **research**. Much of this research has been done in the African-American community and has found that they report lower levels of trust in medical research than do Whites (Shavers, Lynch, & Burmeister, 2002). There are profound implications of lack of trust in medical research. For example, mistrust may be responsible for the low levels of enrollment in clinical trials (Shavers et al., 2002) and low rates of organ donation (Minniefield & Muti, 2002) among African Americans.

Trust in clinical research has received the most attention. While not systematically investigated, it is likely that distrust also extends to health survey research. Response rates of African Americans, American Indians, and Hispanics are typically lower than those of Whites. Anecdotal evidence from our own experiences suggests there is a high level of suspicion and distrust of survey research (and the results of such research) conducted by institutions.

In his apology to victims of the Tuskegee experiments, President Clinton stated, "We commit to increase our community involvement so that we may begin restoring lost trust [in medical research]." Certainly, greater community involvement has that potential. But "involvement" is an ambiguous term, and many have argued instead for full and equal partnership in the research enterprise. Those engaged in research may find themselves simultaneously pushed by the community to allow full participation and pulled toward scientific standards for what constitutes good survey design and process.

Community-based participatory research (referred to as participatory research in the remainder of this paper) is one model for addressing these complex tensions (Israel, Schulz, Parker, & Becker, 1998). This paper describes a model of participatory research employed in the creation and implementation of a survey of Minnesota health care program enrollees (e.g., Medicaid, MinnesotaCare). The

Acknowledgements: This research was supported by the Minnesota Department of Human Services, the University of Minnesota Division of Health Services Research and Policy, and the Allina Foundation. We acknowledge other members of the research team (Charity Kreider, Jennifer Lundblad, James McRae, Betty Moore), and we are indebted to representatives from a number of organizations for their contributions, guidance, and insights (Hispanic Advocacy & Community Empowerment Through Research, Powderhorn Phillips Cultural Wellness Center, and the Red Lake Tribal Council).

benefits and challenges of this participatory approach are described, and the implications for the field more generally are discussed.

WHAT IS COMMUNITY-BASED PARTICIPATORY RESEARCH?

The goal of participatory research is to build partnerships between community and academic or government researchers, giving each ownership of and responsibility for the research process. Participatory research in the field of racial and ethnic disparities in health is founded on the principle that members of communities most affected by disparities must fully participate in the research process to ensure the relevancy and usefulness of study results (Gaventa, 1991; Israel et al., 1998; Whyte, Greenwood, & Lazes, 1991). Active involvement by community researchers in all stages of the study is intended to foster trust in the process and produce results and may better activate members of the broader community to work toward solutions (Schulz et al., 2001). Researchers from academic and government agencies also should benefit by gaining an understanding of communities' strengths and the constraints members face in pursuing wellness. This model contrasts with traditional models of research that sometimes place the needs of researchers ahead of the needs of those affected by disparities. As a result, traditional models often are viewed negatively, as communities are treated as the "subjects" of research and rarely use and/or benefit from the results (Green & Mercer, 2001). Challenges to participatory approaches to research are significant investments of time to the process and the risk of sacrificing scientific rigor in order to attain the support of community members.

THE DISPARITIES IN MINNESOTA HEALTH CARE PROGRAMS STUDY

Study Goals

The primary goal of the "Disparities in Minnesota Health Care Programs" study was to explore racial and ethnic disparities in barriers to health care services use among public program enrollees. The populations oversampled in the survey were from American Indian, African-American, Hispanic/Latino, Hmong, and Somali communities.

Research Process

From the outset, the project was based on a participatory model of research. A group of researchers who had worked together on another participatory project came together to respond to the request for proposals (RFP). This group included several community researchers (Hmong and Hispanic researchers and a research associate from a nonprofit organization), several university-based researchers, and a staff member from Minnesota's External Quality Review Organization (EQRO involvement was recommended in the RFP). This group comprised the Project Management Team (PMT). The intention to conduct participatory research featured prominently in the response to the RFP; the response described the team composition, structure, roles, power relationships, resource allocation, and communication strategies.

When the contract was awarded, the PMT expanded again to include the Project Officer from DHS, as well as Somali and American-Indian researchers. The PMT oversaw all aspects of the project. Five subcommittees were formed to direct specific tasks: (1) focus groups (focus group design, implementation, analysis, and report writing); (2) instrument development (creating English mail and telephone instruments and translations); (3) survey administration (sample design, monitoring data collection); (4) data analysis (outlining and overseeing analysis, interpretation of results, and formulation of recommendations); and (5) dissemination (report and manuscript production, conference submissions and presentations). Each subcommittee was comprised of community and institutional (e.g., university, EQRO, DHS) members.

The PMT met monthly for two hours and was the forum for subcommittee updates and project decision making. Subcommittees met more frequently (sometimes weekly or twice weekly, depending on the phase of the research process), making many task-specific decisions independently. All PMT members were paid for their participation in the project (salary coverage for institutional members, hourly rates for community researchers' effort).

Information presented here about the participatory process comes from PMT and subcommittee meeting notes and progress reports submitted to DHS. In addition, we draw on personal observations of the process and responses to a short open-ended debriefing questionnaire sent to all PMT members.

PRINCIPAL FINDINGS

Constraints to Participatory Research

Balancing the need for process and consensus building against time and budgetary constraints was a significant challenge. For example, time to respond to the RFP was limited, which in turn limited the participation of community researchers with competing work schedules. Coupled with these time constraints, the more limited grant writing experience of several community researchers may have inhibited their participation in this activity. As is the norm, the total budget for this project was capped in the RFP. Because a large portion of the budget was dedicated to data collection (focus groups and surveys), the contributed effort of PMT members was tightly estimated and, in the end, underestimated.

Time pressures only increased once the grant was awarded; the funding agency required that we complete in nine months what we conservatively estimated to be a 15-month project. The PMT tried to address challenges by setting up efficient communication systems and decision rules and structures early on. However, time constraints and the tension between process and product remained throughout the course of the project. Consensus building and careful deliberation of some decisions were limited due to the need to stay on schedule. Yet PMT members' dedication and ability to maintain a collegial work environment while implementing a fast-paced and rigorous study were listed as the project's top successes in the PMT survey.

Structural factors, such as how budgets are administered in academic settings, also threatened the participatory nature of the process. For example, the portion of total funds the university charges as indirect cost recovery (ICR) was a source of frustration among community members. To avoid ICR charges per subcontract, the PMT entered into one contract with a member's home nonprofit organization. This organization administered the subcontract and distributed funds to community researchers on the PMT, subcontracted with focus group trainers and facilitators (a total of nine separate contracts), and made payments to community members participating in data interpretation and development of recommendations. Worry over the university's ability to make timely payments to resource-restricted individuals/organizations was a second reason for creating this arrangement. Although a practical solution to the ICR and payment problems inherent in grants with academic institutions, it was not without some hassle for the organization taking on this responsibility.

Questionnaire Development

The level of participation by community members was far greater than is typical in survey research. It is not uncommon for researchers to seek their input about the content of an instrument, brainstorm with them, or present key community members with a well-developed draft for comment. The process involved in developing the instrument for this project was quite different. Subcommittee members met at least weekly to

discuss domains to be included and individual items. Example questions used in national health studies were presented for review, but the team held lengthy discussions about the wording of questions and their cultural relevance.

There was some tension between the need for scientific rigor in the development of the instrument and sustaining community support. For example, due to length constraints and concern about the flow of the overall questionnaire, not all issues around conceptualization of health and preventive care raised in the focus groups could be addressed in the final instrument. Instrument content was further constrained by priorities set in the contractual arrangement with DHS.

Community researchers had very high standards for assessing the quality of the questionnaire translation, serving as and locating other pretest subjects, as well as participating in the hiring and training of bilingual interviewing staff (details below). Combined, these increased the face and content validity of the instrument across versions. PMT review and finalization of the translated instruments slowed data collection and necessitated a rebudget. The participatory nature of the project garnered several concrete benefits. First, significant resources were conserved based on multilingual members' advice that the mail survey be administered in English only, because those unable to read in English also would be unable to read a translated instrument. The telephone version of the instrument was made available in Spanish, Somali, and Hmong, and call-in lines were publicized in the relevant languages at the bottom of the mail questionnaire's cover page.

A second major contribution was community members' willingness to question the underlying premises of the research project. For example, as originally conceived, the study focused on barriers to preventive care, conservatively defined as care provided by a doctor or clinic. Community members were the first to question the appropriateness of this definition, pointing out that there are other legitimate sources of preventive health care. This resulted in the expansion of items to include sources of care such as "spiritual or traditional healer or shaman" and "an acupuncturist or herbalist."

Community researchers also successfully argued for the inclusion of many questions related to how individuals are treated by the health care system. These included questions about discrimination based on race/ethnicity, ability to pay, or enrollment in public programs. As it turned out, the perception of unfair treatment based on enrollment in public programs was the most common type of discrimination reported by respondents.

Finally, based on focus group results and community researchers' experiences, sections of the instrument were augmented to assess the availability and quality of interpreter services, as well as trust, fear, and views of the role of doctors or other health care providers in the production of health. For example, questions were added to assess respondents' fears of their providers—fear their providers may not do enough to find out what is really wrong, may deliver care that makes them feel worse, tell them they have an illness they do not really have, and/or fail to find an illness they do have.

Conceptual Equivalence & Interviewer Quality

Participation by multilingual members in the design of the English mail questionnaire resulted in early attention to survey content that would transcend cultural and linguistic communities. However, concerns were voiced that attention to finalizing the instrument in English took precedence over the translations into Spanish, Somali, and Hmong. Multilingual members were in an ideal position to review the translations and act as a quality control mechanism for the telephone surveys (including recruitment of skilled interviewers).

Specific translation problems fall into two categories: translations that were too literal in

orientation and errors in translation. In all three languages, there was a tendency toward literal translations that were not correct or changed the meaning of a given statement or question. A Spanish example was the translation of "Indian Health Center" that included a Spanish word for "Indian" that could easily be confused with "from India." The Hmong translation often relied on the use of high Hmong vocabulary that the majority of Hmong do not use conversationally. For example, most Hmong use the English word "doctor" rather than "tus kws kho mob" (the one who cures diseases). Additionally, literal translation of the word "research" implies something very different and frightening; instead, the word for "survey" was used in the revised translation.

Multilingual members also noted errors in the translations (i.e., the translation added text, deleted text, or was simply incorrect). Although adult and child versions of the instrument were virtually identical, it was clear that the Spanish translations were conducted by two separate teams (or individuals): One was of higher quality than the other. A number of questions (nine total) prompt the respondent with the phrase "*Would you say*"—for example, "In general, how would you rate your overall health? *Would you say* it is excellent, very good, good, fair or poor?" In the Spanish translation, the back translation instead yielded "It could be said [response options]…." For one question, the Spanish translators added a fifth response option to a four-option response set in the English version.

Subtleties of the Hmong language did not receive sufficient attention. Specifically, in the Hmong language, there are multiple ways of saying "Yes/No," and the response code selected must correspond to the question asked. For example, sometimes instead of using "Yes/No," the equivalent of "I believe/do not believe" or "can do/cannot do" is more appropriate. Attention to these details improves the experience for both interviewer and respondent. The Somali community members reviewing the instrument felt some of the question translations resulted in loaded phrasing of the question or indicated a misunderstanding of the concept the question was intended to capture.

Dissemination Issues

Finally, the PMT continues to face challenges in the dissemination phase of the project. Although a participatory model encourages ownership and willingness of all members to share findings within their communities, to date dissemination has been initiated by institutional members alone. Even though community researchers are involved in presentations and manuscript production, dissemination has been restricted to professional communities. Time constraints and lack of resources play a role in this. The PMT is pursuing strategies for sharing study results more broadly out of a commitment to the participatory model, but this is contingent upon acquiring additional funds.

CONCLUSIONS

We have argued that the participatory nature of the project (despite the challenges) benefited the quality of the survey and resulting results. However, there are two central questions left unanswered: Did we achieve full participatory research? Should participatory research matter to the survey research community?

Did we achieve full participatory research? No. If judged by the standards outlined earlier, the process fell short. Time and budgetary pressures worked against fully meeting this goal. The implication for future projects incorporating a participatory model is the need to recognize that early identification and active involvement of key community partners is critical to project planning and implementation. Attention should be given to the additional time and resources that participatory research requires. Developing communication strategies and decision-making rules early on will facilitate

movement from process to product. In addition to benefits to the quality of the research, the process provided intangible benefits to the project team. These included the benefits of working together for a project that required learning about each other's cultures, which built trusting relationships between individuals working on the project. These relationships, in turn, have helped build social capital that can be used by all project members. Researchers from the community and the institutional setting continue to work together on a variety of issues including help with student projects, letters of reference, grant assistance and informal exchanges of information about either community or institutional concerns. These opportunities would not have been possible without the initial work toward building a partnership on this research project.

Our continued commitment to participatory research also partially answers the second question—we believe participatory research should matter to the survey research community. It is possible to meet standards of methodological rigor while still responding to community concerns about survey content. Involving communities fully in survey research also serves a larger purpose. Community members cannot always be expected to be the "subjects" of research and not the owners. Although too early to know, it is hoped that community involvement will encourage systems-level change (e.g., DHS) and community application of study results. Small projects such as ours will not result in a sea change, nor do they begin to address the wider problem of distrust in research. However, the team of researchers assembled believes that our experience was another step in a long and difficult process of changing the relationships between institutional researchers and the community, and building trust.

.

REFERENCES

Doescher, M. P., Saver, B. G., Franks, P., & Fiscella, K. (2000). Racial and ethnic disparities in perceptions of physician style and trust. *Archives of Family Medicine, 9*(10), 1156–1163.

Gaventa, J. (1991). Toward a knowledge democracy. In O. Fals-Borda & M. A. Rahman (Eds.), *Action and knowledge: Breaking the monopoly with participatory action research* (pp. 121–131). New York: Intermediate Technology/Apex.

Green L. W., & Mercer S. L. (2001). Can public health researchers and agencies reconcile the push from funding bodies and the pull from communities? *American Journal of Public Health, 91*, 1926–1929.

Institute of Medicine. (2001). *Crossing the quality chasm*. Washington, DC: The National Academy Press.

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.

Kao, A. C., Green, D. C., Zaslavsky, A. M., Koplan, J. P., & Cleary, P. D. (1998). The relationship between method of physician payment and patient trust. *Journal of the American Medical Association, 280,* 1708–1714.

Minniefield, W. J., & Muti, P. (2002). Organ donation survey results of a Buffalo, New York, African-American community. *Journal of the National Medical Association, 94*(11), 979–986.

Schulz, A. J., Israel, B. A., Parker, E. A., Lockett, M., Hill, Y., & Wills, R. (2001). The East Side Village health worker partnership: Integrating research with action to reduce health disparities. *Public Health Reports*, *116,* 548–557.

Shavers, V. L., Lynch, C. F., & Burmeister, L. F. (2002). Racial differences in factors that influence the willingness to participate in medical research studies. *Annals of Epidemiology, 12*(4), 248–256.

Whyte, W. F., Greenwood, D. J., & Lazes, P. (1991). Participatory action research: Through practice to science in social research. In W. F. Whyte (Ed.), *Participatory action research* (pp. 19–55). Newbury Park, CA: Sage.

FEATURE PAPER: National Health and Nutrition Examination Survey: Advance Arrangements and Outreach

Nancy A. Krauss and Kathryn S. Porter, National Center for Health Statistics
Jack Powers, Westat
Glenn D. Pinder

INTRODUCTION

Although many elements contribute to the success of the National Health and Nutrition Examination Survey (NHANES), this paper focuses on community outreach as a method of achieving and maintaining high response rates. NHANES outreach is a nexus of interactions between the community and the survey staff. Here we describe the complex process of advance arrangements, where contact is made with local officials, community leaders, religious leaders, the media, and potential survey participants. The goals of these efforts are to solicit community participation, convey community and individual benefit, and obtain cooperation from the community and the individuals selected for the study.

OVERVIEW OF NHANES

NHANES is one of the major data collection programs of the NCHS. Designed to assess the health and nutritional status of adults and children in the United States, the NHANES has been conducted periodically since 1960. The eighth, most recent, and currently ongoing survey began in 1999. All NHANES are planned and conducted by a team of NCHS and contractor staff.

The NHANES target population is the civilian noninstitutionalized U.S. population. The current NHANES oversamples low-income Whites, adolescents 12–19 years of age, persons age 60 and older, African Americans, Mexican Americans, and pregnant women.

The survey is comprised of a household interview component and a health and nutrition examination component. Every year, approximately 6,000 individuals of all ages in 15 counties across the country are interviewed in their homes. Of these, approximately 5,000 complete the health and nutrition examination. Health examinations are conducted in mobile examination centers (MECs), which provide an ideal setting for the collection of high quality data in a standardized environment. Approximately 300–400 persons are examined during the 4–6 weeks of examinations conducted at each of the 15 sites.

The major goals of NHANES are as follows:

- To estimate the number and percent of persons in the U.S. population and designated subgroups with selected diseases and risk factors;
- To monitor trends in the prevalence, awareness, treatment, and control of selected diseases;
- To monitor trends in risk behaviors and environmental exposures;
- To analyze risk factors for selected diseases;
- To study the relationship between diet/nutrition and health;
- To explore emerging public health issues and new technologies; and
- To establish a national probability sample of genetic material for future genetic testing.

ADVANCE ARRANGEMENTS

During a year's time, NHANES is conducted in fifteen locations throughout the U.S., each site usually encompassing one county. About four months prior to starting field work in a given location, NCHS and

contractor staff initiate face-to-face meetings with local officials and community leaders. Our experience has shown that local health departments are generally the best source of local information and support for the study, and we schedule our first meeting through the local health officer. NCHS and contractor staff develop a meeting agenda shared with all attendees. The meeting may include a number of health department personnel and, in some cases, others invited by the department director. Our primary objectives are to convey the importance of NHANES, to discuss the community's role in the success of survey operations, and to suggest the benefits of participation for individuals selected for the survey. Discussions at the initial meeting provide a wealth of information about the community. Our intention is to establish a good working relationship with the health department staff and to continue that relationship throughout the advance process and the field work.

Subjects discussed at the initial meeting can be wide ranging, but the agenda always includes specific items, including an overview of NHANES, potential MEC sites, local sociodemographic information, and medical and dental referral sources for those survey participants who may need them. Significant time usually is devoted to conveying the important contributions NHANES data have made to improve the nation's health. Officials are given an overview of the sample design so they can understand how their community was selected and the process of determining which individuals will be asked to participate. The benefits of participation also are detailed. For example, all sample persons participating in the MEC examination will receive a report of medical and lab findings in addition to a cash payment for their time and effort. Many of these tests are not routinely performed during regular physical examinations, and sample persons may share results with care providers if they choose.

Because the location of the examination center will be an important consideration in the decisions of those who are asked to be examined, a top priority of initial discussions is to obtain advice on potential desirable sites for MEC placement. Many logistical and engineering considerations are entailed in selection of the MEC site, such as type and condition of ground surface; maneuverability within the space when "parking" the MEC; and availability of electrical, water, sewer, and telephone connections. But of equal importance in the team's evaluation of a possible site are its centrality, accessibility to major thoroughfares, and image in the community. Discussions with health department staff and other local officials usually have provided concrete leads. In the past, the MEC has been successfully located on health department properties, on hospital grounds, on local university campuses, on hotel parking lots, at shopping malls, and at a few quite unique sites.

The advance team asks health department staff for information about local medical and dental referral resources for sample persons who might have a problem identified during the survey exam but who do not have regular care providers with whom to follow up. We believe we have an obligation to refer these people for appropriate care or treatment, and we compile a list of clinics providing free or sliding-scale services. In some areas, referral resources may not be readily available or hard to identify. However, we have found that some way of providing needed medical or dental care referrals to indigent people can be found and that the health department forum is always a good place to determine the best course of action.

The advance team identifies other local resources at this meeting. We request the name of a nutritionist who could act as a resource to the MEC nutritional interviewers or who could refer a sample person for nutritional counseling. Procedures for handling medical emergencies at the MEC or in the home are discussed and determined.

HIV testing facilities and counseling services available in the county also are identified.

Prior to the meeting, survey staff develop a demographic profile of the area. This allows the team to focus on and discuss groups that may require special outreach services, such as translators, same-sex examining staff, religious leaders and other community gate keepers, etc. In addition, the community profile that emerges from discussions with local officials is a valuable reference for the team as it performs the many tasks required to bring the study to the community. Thus, the initial meeting establishes a collaborative process, and health department personnel often feel they are working with the project staff to achieve the best outcome for the community and the study.

Another important part of the advance arrangement and outreach effort is informing a wide range of local officials and area leaders about the study. We believe that this is essential in building credibility in any community and in creating a consciousness of NHANES as a legitimate and important data gathering effort. The NHANES is a high-profile study in many communities, and officials need to know what it is about and to be able to answer citizens' questions when they arise. About three months prior to the start of field work, all local officials are informed in writing about the NHANES survey operations in their area. Specifically, notification letters are sent to the state's Senators and Congresspersons, local mayors, council members, fire department and law enforcement officials, school superintendents, religious leaders, and other community officials. The letter provides the dates the study will be in the area, indicates the expected number of people to be examined, and provides a name and phone number for questions. An attachment provides additional information on survey content and uses of the data. Two weeks prior to fieldwork, these same authorities are sent another letter giving the addresses and telephone numbers of the local NHANES field office and examination center.

MEDIA COVERAGE

Articles about the survey in local print media have proven to be one of the most important items in our interviewers' package of information to encourage participation. Such media coverage is invaluable in gaining the cooperation of potential survey participants. Television and radio spots also enhance the legitimacy of the survey in a community. Some health departments have a public affairs staff, and those persons (or person) can act as an important liaison to local press, television, and radio. However, the primary resource for obtaining media coverage is a full-time NCHS staff member dedicated to that task.

Approximately two weeks prior to opening the field office in the selected area, a media list is constructed using Burrelle's Data Base. The names of daily and weekly newspapers, including foreign language newspapers, are extracted, by dominant market areas. After our initial meeting with the health department, NCHS may obtain the names of local media contacts from the public relations staff. If contact information is not available, calls are made to newspapers to identify contacts. Press kits, which include a cover letter, press release and a folio with brochures, an overview of NHANES, data briefs and data accomplishments, are sent to each contact. Press releases are tailored to local areas. Usually, several follow-up telephone calls are made to ascertain whether or not the newspaper will use the story. Our initial objective is to place an article in one or more local newspapers immediately prior to the beginning of interviewer household contacts. Ideally, the article should describe the study, the contributions to national public health made by NHANES, and the benefits of participation. It is this kind of article our interviewers want to have to show to potential respondents. Articles are frequently

placed in local Area Agency on Aging Newsletters.

When the MEC arrives and is set up, NHANES presents an "open house" to which local officials, selected guests, and the media are invited. This is an opportunity for "live" coverage and often attracts additional print, television, and radio coverage. The field managers on each field team are given media training and act as the primary spokespersons for the survey at each location. They sometimes are given the opportunity to do television and radio interviews and often make presentations to community groups, such as service clubs and other civic organizations. Senior staff members from NCHS also are called on for media interviews if the occasion arises.

AT THE DOOR STEP

All households selected for NHANES are sent a letter briefly describing the survey and explaining that an NHANES representative will be visiting. These letters are mailed within a week or two before household contacts begin. A copy of this letter, along with many other materials that further describe the study and the personal benefits of participation and detail specific and easily recognizable contributions NHANES has made to the health of the nation, are carried by interviewers as they visit households. Although the mailed advance letter is in English and Spanish, interviewers use a language identification card to identify non-English- and non-Spanish-speaking households and are able to provide a copy of the advance letter in eight different languages. The language identification card also identifies the need for a translator.

Interviewers and other field staff have contributed significantly to the development of numerous brochures and letters of endorsement that target specific groups, such as pregnant woman, the elderly, African Americans, and Hispanics. Interviewers have found this material crucial in converting reluctant respondents.

The goal of targeted brochures is to emphasize the specific relevance of selected aspects of the survey to that particular group. When such relevance is emphasized, motivation may be increased and perceived burden reduced. For example, the NHANES brochure for pregnant women highlights the fact that information from past surveys showed that women of childbearing age did not have adequate levels of folate and iron in their diets, which contributed to significant public health programs and measures to improve the health of women and their babies. The NHANES African-American brochure identifies medical conditions for which African Americans are at higher risk, such as hypertension, diabetes, and high cholesterol—conditions that are assessed in the NHANES examination.

Over the years, the survey field staff has compiled numerous letters of endorsement. Regularly, about one month prior to opening a field office, letters are requested of the Directors of the local Health Department and the local Area Agency on Aging. These letters also add legitimacy to the survey; these agencies, which have been briefed on the survey, provide a local resource for people to verify the authenticity of NHANES activities in their community. At times, letters of endorsement also are received from Senators and Congressional Representatives. Endorsements from a number of national organizations, such as NAACP and AARP, are routinely available. Field staff frequently request letters of endorsement from schools of public health and medicine, local universities, local organizations, and community, cultural, and religious leaders. Local letters of endorsement have been particularly useful in eliciting the cooperation of minority populations.

Sample persons are asked not only to participate in a household survey, which may require 60 minutes or more for each household respondent, but also to travel to the MEC site to participate in a health examination and laboratory studies that

require providing blood and urine samples. Interviews, travel to the MEC, and examination time spent at the MEC may add several hours to the time burden of the survey task. To offset this burden, NCHS has developed mechanisms for increasing the individual benefits of survey participation. Interviewers emphasize the opportunity for a free health examination and numerous laboratory studies not usually provided during a routine physical examination. More importantly, sample persons are informed that some results will be available at the time of the examination and lab results will be forwarded to them within a few weeks in an easy-to-read format that they are free to share with their regular health care provider. Sample persons also are told that they can call a toll-free number to discuss abnormal findings with the Medical Officer at NCHS and, if necessary, can be referred to a local clinic for further evaluation and/or treatment. Further, those sampled are informed that they will receive cash remuneration for their time and that travel expenses to the MEC will be reimbursed or transportation will be provided if necessary. We have found that potential individual benefits of the health exam and lab studies have been critical in maintaining high rates of survey participation.

CHALLENGES TO OUTREACH

Community leaders are most interested in health estimates for their county, and they generally ask NHANES representatives for community-specific data. Because the sample is a nationally representative design, area-specific datasets cannot be provided to local authorities. NHANES is working to assist local health authorities to obtain health examination data for their area. This is being accomplished by exploring the use of the NCHS Research Data Center for local-area estimates from other surveys, as well as creating a Community Health and Nutrition Examination Survey to be piloted in selected areas in the future.

Some communities have serious public health problems, and local health officials may believe a Federal survey's data findings would reflect poorly on the health department's abilities to manage public health issues. The NHANES staff continually informs health officials that data are compiled for national estimates. Individuals who live in communities where investigations of disease outbreaks or other public health problems or where situations of possible environmental exposures may have led to negative experiences with Federal health officials may tend to mistrust NHANES survey operations. The NHANES program targets outreach to address specific questions that may arise in such communities.

Locked buildings, gated communities, retirement homes, and university dormitories are a continual challenge to survey operations, and anticipation of these should be incorporated into the advance arrangements process. The NHANES program is currently evaluating ways to identify these in advance of survey operations so contact can be made with building management, homeowners associations, and administrative offices to seek cooperation for entry.

COMMENT

NHANES is a very complex survey and requires the continuous monitoring of response rates. Although we are becoming more sophisticated in developing methods of obtaining cooperation from local government agencies, the media, and ultimately the sample persons, we need to evaluate quantitatively the extent to which outreach actually contributes to participation. Over the next year, we are focusing on research that provides insight into participation rates and the effectiveness of alternative outreach strategies on different populations.

FEATURE PAPER: Nonresponse Among Persons Age 50 and Older in the National Survey on Drug Use and Health

Joe Murphy and Joe Eyerman, RTI International
Joel Kennet, Substance Abuse and Mental Health Services Administration

INTRODUCTION

The National Survey on Drug Use and Health (NSDUH)[1] is an ongoing cross-sectional face-to-face household survey of approximately 150,000 households and 67,500 persons each year. It collects data through audio computer-assisted self-interviewing (ACASI) and covers the U.S. civilian noninstitutionalized population age 12 and older. Response rates traditionally have been highest among the youngest respondents and lowest among the oldest, with the lowest rates found in the 50 and older (50+) age group. The introduction in 2002 of a series of methodological enhancements to the study appeared to improve the response rates for most age groups but had only a small impact on the 50+ age group (Kennet Gfroerer, Bowman, Martin, & Cunningham, 2003).

Because lower response rates make nonresponse bias more likely, and since there is a disturbingly low response rate among the 50+, this paper aims to understand why this may be in order to understand how the problem can be ameliorated. This topic is of increasing importance as the proportion of Americans in this age group increases (U.S. Census, 1999). Obtaining unbiased survey estimates will be vital to accurately assess substance abuse treatment need for older Americans in the coming years. This need is expected to nearly triple by 2020 as the baby boom carries its alcohol and drug use into older ages (Gfroerer, Penne, Pemberton, & Folsom, 2002).

The purpose of this paper is to provide a better understanding of nonresponse among older sample members in the NSDUH in order to tailor methods to improve response rates and reduce the threat of nonresponse error. This paper examines the components of nonresponse (refusals, noncontacts, and/or other incompletes) among the 50+ in the NSDUH. It also examines respondent, environmental, and interviewer characteristics in order to identify the correlates of nonresponse among the 50+, including relationships that are unique to the 50+. Finally, this paper considers the root causes for differential nonresponse by age, drawing from focus group sessions with NSDUH field interviewers on the topic of nonresponse among the 50+. The results show that the response patterns are different for the 50+ age group than for younger age groups and that the difference is probably a function of different perceptions of the interviewing process.

BACKGROUND

In an ideal situation, nonresponse would be consistently low for all demographic groups in the target population. Unfortunately, NSDUH nonresponse is positively associated with respondent age. This relationship has been identified elsewhere in the survey nonresponse literature. Herzog and Rodgers (1988) analyzed data from several face-to-face surveys including the Americans View Their Mental Health study (AVMH) and the American National Election Studies and found a linear decline in response rate with increasing age. Refusal as a proportion of all nonrespondents increased for the middle years (35–74) and then declined, reaching particularly low proportions among the oldest old (75+). The reason for nonresponse among the oldest age groups was less often outright refusal than among the middle age groups. Groves and Couper (1998) suggested that although the elderly are more frequently at home due to their low employment rate and reduced mobility, their poor health may prevent them from survey participation. Others also have noted that increased age of household members negatively affects survey cooperation (e.g., Redpath & Elliot, 1988).

Chiu, Riddick, and Hardy (2001) analyzed data from the National Health Interview Survey (NHIS) and found a different relationship between response and age. They report that difficulties in interviewing are experienced less often in households containing seniors and members with activity limitations when controlling for all other predicting variables. They believe that this is because these people are more

[1] The survey was called the National Household Survey on Drug Abuse prior to 2002.

likely to be home during the day and because the topic of health is viewed favorably among the elderly. Kautter, Khatutsky, Pope, and Chromy (2003) found no significant relationship between age and nonresponse in their analysis of the Medicare Beneficiary Survey (MCBS), another household health survey. The idea that the topic of health is salient to older respondents and that they are more likely to respond to surveys that deal directly with health topics could be important to the tailoring of the NSDUH.

RESPONSE RATE & AGE IN THE NSDUH

Each year, approximately 150,000 households are screened for the NSDUH. Basic demographic information about the household and its members is captured during a short screening interview. When the screening is complete, 0, 1, or 2 sample persons are selected for the full NSDUH interview based on household composition. The screening response rate for the 2002 NSDUH was 90.7%. Sample members are selected for the NSDUH interview with a predetermined proportion of respondents in each of five age groups: 12–17, 18–25, 26–34, 35–49, and 50+. Interview response rates[2] by age for the NSDUH survey years 1999 to 2003 are presented in Figure 1. Response rates were successively lower for each sampled age group in each year.[3] Across all years, response rates were lowest for the 50+ age group and highest for the 12–17 age group.[4] The difference in response rates between these two groups remained around 13% from 1999 to over 18% in 2002 and 2003. Response rates for each age group increased from 2001 to 2002, and these increases were statistically significant for all age groups except the 50+.

The increase in response rates between the 2001 and 2002 surveys occurred about the same time as several methodological changes introduced during this period (SAMHSA, 2003): The name of the survey was changed in 2002 from the National Household Survey on Drug Abuse (NHSDA); incentive payments of $30 were given to all interview respondents beginning in 2002; improved data collection quality control procedures were introduced in the survey during 2001 and 2002; population data used in NSDUH sample weighting procedures were based on the 2000 decennial census for the first time in the 2002 NSDUH; and the pair selection algorithm was changed in 2002 to increase the pairs selected in the 50+ age group. The pattern of change in response rates by age between 2001 and 2002 suggests that these methodological changes had a larger positive effect on the response propensity of younger respondents than older respondents, thereby creating even larger differences in response rate by age. The remaining analyses focus only on the 2002 nonresponse in order to minimize the confounding effects of the methodological changes.

ANALYSIS OF NONRESPONSE COMPONENTS

Nonresponse in the NSDUH can be categorized into three components: noncontacts, refusals, and other incompletes. Because each of these components is the result of a different process in survey participation, it is important to understand which components are driving the overall nonresponse in order to design and implement effective strategies to reduce it (Groves & Couper, 1998). The relative contributions of each of these components to nonresponse may vary with the age of the sampled person. For example, older respondents may be easier to contact because they spend less time out of the home, but they may refuse at a greater rate than younger respondents due to concerns with personal safety. The relative contributions of each of the components to NSDUH nonresponse are examined in Figure 2, which presents the weighted noncontact, refusal, and other incomplete rates for the 2002 survey. Age has been disaggregated into five-year categories to detect any differences that might occur within a particular age group. The 50+ group is not homogenous in terms of reasons for nonresponse; those 65 and older (postretirement age) may be more likely to be at home and therefore may have a lower rate of noncontact than those age 50–64. As shown in Figure 2, nonresponse was below 25% for all age groups under 50 and above 25% for all age groups age 50 and above. The main reason for this difference was a higher rate of refusals among the 50+ compared to those under

[2] Unless otherwise specified, all rates presented in this report are calculated using weighted data.
[3] Because age is not collected until the screening interview, screening response rates by age are not available.
[4] This does not imply that people actually become less likely to respond as they get older. The NSDUH data are cross-sectional and cannot be used to measure such a relationship.

Figure 1. Weighted Interview Response Rate (IRR) by Age: 1999–2003

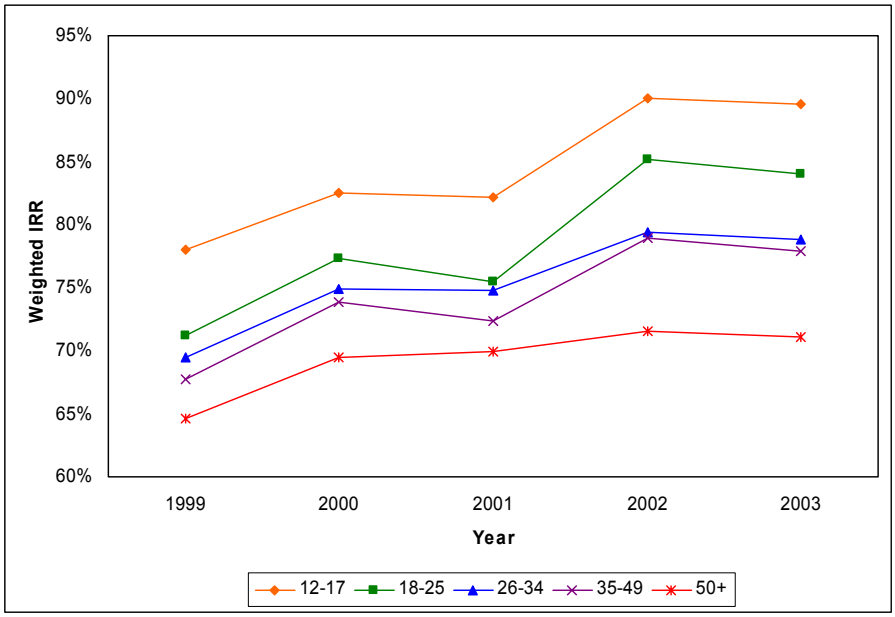
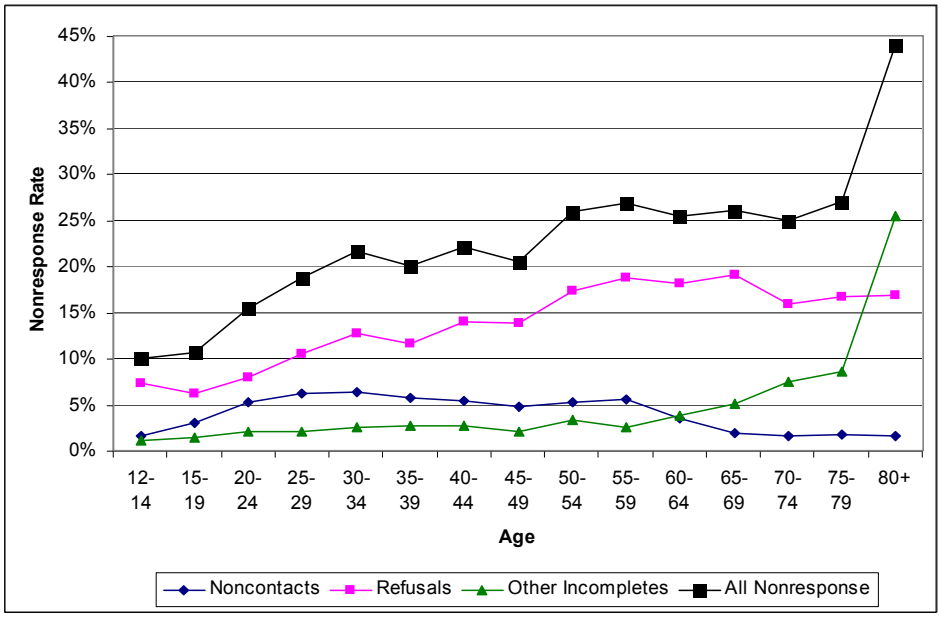


Figure 2. Weighted Nonresponse Rates by Five-Year Age Groups: 2002



50. Refusals for all ages below 50 in 2002 were below 15%, and refusals for all ages 50 and older were above 15%, though the rate of refusals for those 70 and older was only slightly higher than the rate for those age 40–49. The rate of noncontacts was near 5% for the ages 20–59 and then declined starting at age 60. For the oldest age groups, it was almost a nonfactor. The relatively higher rate of other incompletes among the oldest sample members also contributed to higher nonresponse for those age groups.[5]

It appears that lower response rates among older respondents are due to relatively higher refusal rates and, in the oldest age categories, other incompletes, as compared to younger

[5] Nearly one-fourth of the 80+ sample members did not complete an interview because they were physically or mentally incapable.

sample members. Because refusals have such an impact on nonresponse among this age group and because it is a component that may be reduced through a tailored methodological approach, the remainder of this analysis is focused on refusals among the 50+.

ANALYSIS OF REFUSAL CORRELATES

Previous research on the NSDUH and other surveys has demonstrated that, in addition to age, numerous factors affect nonresponse (Eyerman, Odom, Wu, & Butler, 2002; Groves & Couper, 1998). To assess the impact of respondent, household, environmental, and interviewer characteristics on the likelihood of refusal among those 50 and older, the available NSDUH data were analyzed on the possible correlates of gender, race, ethnicity, number of respondents selected per household, household composition, population density, socioeconomic status (SES), region, and interviewer experience. Noncontacts and other incompletes were excluded from this analysis because their inclusion in the refusal rate denominator could result in misleading conclusions. Among the 50+ sample members, refusals were most common for those with these characteristics: in households where two sample members were selected for the interview; in two-person households; in households with no members under age 18; in non-single-parent households; in Metropolitan Statistical Areas (MSAs) with 1 million or more residents; in high-SES segments; in the Northeast region; when the respondent and field interviewer (FI) were of opposite sexes; when the respondent was White/other being interviewed by a Black FI[6]; when the respondent was interviewed by a FI under age 50[7]; and when the respondent was interviewed by an inexperienced FI.

The correlates of refusal were similar for sample members 50 and older compared to all sample members. There were few differences in terms of characteristics of sample members most likely to refuse. Among sample members of all ages, those in households where one respondent was selected for an interview refused more often than those in households where two were selected. The FI/respondent race combination that had the highest rate of refusals was that in which the FI was Hispanic and the respondent was White/other. Refusal rates were significantly lower for all respondents compared to 50+ age group for most types of sample members; exceptions are households of five or more persons and some FI/respondent race combinations.

Logistic regression models were run to simultaneously test the effects of these measures on refusal propensity. The first model was limited to all sample members 50 and older. Refusal propensity was not significantly different among any age group over 50. Sample members in households with one or two members were significantly more likely to refuse than those in households with five or more members. However, the presence of a minor and single parent status were not significant predictors among the 50+. This suggests that older sample members living in small households (one or two members) are much more likely to refuse, regardless of the age of the other household members. Among the 50+, sample members in densely populated areas were significantly more likely to refuse than those not living in such areas. The combination of FI and respondent gender was a significant correlate of refusal in the case where the respondent was male and the FI was female (compared to when both were female). Compared to the scenario in which the FI and respondent were both White/other, refusal was significantly less likely when both the respondent and FI were Hispanic and when the respondent was Black and the FI was White/other. Cases finalized by inexperienced FIs were significantly more likely to result in a refusal than those worked by highly experienced FIs.

Most of these relationships were replicated when respondents of all ages were included in the model. But unlike with the 50+, the full model showed no significant relationship between number of respondents selected in the household and refusal propensity. This difference may be due to older respondents not having time or not being willing to devote their collective available time to the survey. It is possible that the increase in the number of selected pairs containing an older person in 2002 may have had a detrimental effect on response rates among older sample members.

[6] Although the FI Hispanic/R Black combination had the highest refusal rate, there were too few cases of this type to make this a meaningful result.

[7] The highest refusal rate is actually found most often where FI's age was unavailable, but this rate is very close to that for FIs under 50.

FOCUS GROUPS WITH FIELD INTERVIEWERS

To address the question of *why* those 50+ refuse at a higher rate than those under 50, we conducted focus groups with NSDUH interviewers, who have the most direct contact and experience with respondents. The focus group data suggest that fears and misperceptions are factors in the response process for older respondents. A fear of scams among this group may lead to an aversion to inviting unknown persons into their households. Also, apprehension toward the electronic hand-held device used by interviewers to enter screening data and the ACASI laptop may affect participation among the 50+. This is consistent with studies that have found that older adults have significantly higher computer anxiety than younger adults (Laguna & Babcock, 1997). Another commonly reported misperception among older respondents is that they have nothing to offer the study. Interviewers report that many respondents say, "I do not use drugs, so you don't need to interview me," or "My experiences are irrelevant to this study." Hoinville (1983) argues that relevance or interest in the survey topic might be a critical factor in obtaining high response rates among older adults.

Increasing the public's awareness of the study through contact with local police and public health departments as well as press releases to local newspapers could help raise awareness among community residents and enhance the perceived legitimacy of the study. While interviewers reported that the $30 incentive is helpful in gaining the cooperation of most respondents, money was not the prime motivator for this group, and in some cases, the incentive actually raised suspicions of fraud or scams.

All interviewers agreed that a great deal of patience and friendly professionalism is needed to gain the cooperation of 50+ sample members. Gaining the respondent's trust is an important step that needs to be taken before attempting to complete a screener or interview. Interviewers also reported that the survey provides an opportunity for parents or grandparents and children to communicate on the subject of drugs and provides a positive shared experience and that 50+ respondents may be motivated by their concern for children and society in general.

DISCUSSION

To combat the effects of lower response rates among older sample members, a variety of methods have been implemented on other surveys including tailoring the questionnaire for older respondents (Jobe, Keller, & Smith, 1996); providing mode options and allowing proxies to respond for older sample members; employing interviewers with strong interpersonal skills (NESC, 2002); developing a special interviewer training module (NESC, 2002); using a slower pace in the interview to increase comfort (NESC, 2002); alleviating respondent fears (Moorman, Newman, Millikan, Tse, & Sandler, 1999); and converting refusers with a financial incentive[8] (NESC, 2002).

Several protocol changes and methodological enhancements have been considered on the NSDUH to improve the response rates for the 50+ age group. These possible changes include adjusting training modules to better cover the concerns of the 50+ age group; altering the lead letter and refusal conversion letter to emphasize concepts that are salient to the older population, such as civic duty, the problems of drug-related crime, or the potential benefits for the younger generation (e.g., grandchildren); developing alternative modes for interfacing with the ACASI interview, such as a larger keyboard, a keypad tailored to the instrument, or a touch screen; conducting a public health communications campaign at the local level prior to data collection; evaluating the potential for a differential incentive payment (higher or lower) for the 50+ age group based on lessons learned from planned focus groups with potential respondents in this age group; and tailoring a few brief video clips using individuals recognized by the general public and well-respected by the 50+ population that could be played on the interviewers' screening devices. While the goal is to reduce the potential differential nonresponse error, care should be taken to avoid additional measurement error in the 50+ age group through changes in the survey materials or in the interaction between the interviewer and the respondent that cause the respondents to self-report differently.

[8] There are questions about the ethics and fairness of the use of targeted incentives for certain subgroups of interest or for refusal conversion (Groves & Couper, 1998).

CONCLUSION

This paper has shown that nonresponse in the NSDUH is higher among the 50+ than among any other age group and is primarily due to a high rate of refusals, especially among sample members age 50–69, and a high rate of physical and mental incapability among those 70 and older. Taken together with evidence from interviewer focus groups, it appears that the higher rate of refusal among the 50+ may, in part, be due to fears and misperceptions about the survey and interviewers' intentions. Increased public awareness about the study may allay these fears. While an increase in the incentive amount may not automatically increase response rates among this group, other protocol changes and methodological enhancements may be effective.

The next step in this analysis is to conduct focus groups with potential NSDUH respondents age 50 and older. The thoughts and concerns of these participants will help guide the design of tailored methods that can be tested and implemented in order to assure the most accurate survey estimates possible.

REFERENCES

Chiu, P., Riddick, H., & Hardy, A. (2001). *A comparison of characteristics between late/difficult and non-late/difficult interviews in the National Health Interview Survey*. Paper presented at the annual meeting of the American Statistical Association.

Eyerman, J., Odom, D., Wu, S., & Butler, D. (2002). Nonresponse in the 1999 NHSDA. In J. Gfroerer, J. Eyerman, & J. Chromy (Eds.), *Redesigning an ongoing national household survey: Methodological issues* (DHHS Publication No. SMA 03–3768, pp. 23–51). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Gfroerer, J., Penne, M., Pemberton, M., & Folsom, R. (2002). The aging baby boom cohort and future prevalence of substance abuse. In S. P. Korper & C. L. Council (Eds.), *Substance use by older adults: Estimates of future impact on the treatment system* (DHHS Publication No. SMA 03-3763, Analytic Series A-21). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Groves, R., & Couper, M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Herzog, A. R., & Rodgers, W. (1988). Age and response rates to interview sample surveys. *Journal of Gerontology, 43*(6), S200–205.

Hoinville, G. (1983). Carrying out surveys among the elderly. *Journal of the Market Research Society, 25*, 223–237.

Jobe, J., Keller, D., & Smith, A. (1996). Cognitive techniques in interviewing older people. In N. Schwartz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative process in survey research* (pp. 197–219). San Francisco: Jossey-Bass.

Kautter, J., Khatutsky, G., Pope, G., & Chromy, J. (2003). *Impact of nonresponse on MCBS estimates final report.* Prepared by RTI for the Centers for Medicare & Medicaid Services (CMS).

Kennet, J., Gfroerer, J., Bowman, K., Martin, P., & Cunningham, D. (2003). *Effects of a $30 incentive on response rates and costs in the 2002 National Survey on Drug Use and Health.* Paper presented at the 58th Annual Meeting of the American Association for Public Opinion Research, Nashville.

Laguna, K. & Babcock, R. L. (1997). Computer anxiety in young and older adults: Implications for human-computer interactions in older populations. *Computers in Human Behavior, 13*(3), 317–326.

Moorman, P., Newman, B., Millikan, R., Tse, C., & Sandler, D. (1999). Participation rates in a case-control study: The impact of age, race, and race of interviewer. *Annals of Epidemiology, 9*(3), 188–195.

New England States Consortium (NESC). (2002). *Getting feedback from frail elders and people with disabilities: Factors to consider when selecting a method, an instrument, a vendor.* Retrieved May 28, 2004, from the University of Maryland Center on Aging Web site: http://www.hhp.umd.edu/aging/mmip/TApapers/TApaper13.pdf

Redpath, B., & Elliot, D. (1988). National Food Survey: A second study of differential response, comparing Census characteristics of NFS respondents and non-respondents; Also a comparison of NFS and FSs response bias. *Statistical News, 80*, 6–10.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2003). *Results from the 2002 National Survey on Drug Use and Health: National findings* (Office of Applied Studies, NHSDA Series H-22, DHHS Publication No. SMA 03-3836). Rockville, MD: Author.

U.S. Census Bureau. (1999). *Aging in the United States.* Retrieved May 28, 2004, from http://www.census.gov/ipc/prod/97agewc.pdf

SESSION 2 DISCUSSION PAPER: Community Participation and Community Benefit in Health Survey Research: An Alternative Perspective

Robert L. Santos, NuStats Partners, LP

Health survey research is typically—and from a scientific research perspective—appropriately attuned to achieve specific technical/statistical objectives. However, to adopt such a paradigm ignores the obvious context in which health survey research is conducted: that of a community setting (or a collection of communities). The papers presented in this session have effectively demonstrated the need for community support and participation in conducting effective health surveys. These papers also have demonstrated the range of community partnership and involvement that is possible in health surveys, with the most intensive community involvement emulating the Participatory Action Research (PAR) model as discussed in Minkler and Wallerstein (2003).

The papers in this session illustrate the trade-offs of the PAR model versus the more traditional approach of a top-down survey operation (where the researcher solicits the cooperation of the community in facilitating the conduct of the research study). It is clear that community participation via the PAR model (when implemented correctly) offers specific enhancements to the research design that the traditional approach alone cannot provide. That is, community participation in research is ***good science***:

- PAR can be used to improve measurement and increases response rates among the most at-risk populations; and
- PAR can provide qualitative, contextual insights that quantitative analysis alone cannot produce, thereby enhancing quantitative research findings by providing direction to policy makers.

The remainder of this paper addresses the following areas related to community participation and benefit:

- Research goals and their role in guiding research design,
- Research ethics,
- Threats of the PAR model, and
- Suggestions for local and national health survey research studies on how to adopt more of a community participatory approach in order to reap some of the benefits afforded by the PAR model.

RESEARCH GOALS DRIVE SURVEY DESIGN & METHODS

The PAR approach calls for the involvement of communities in the research process as a *partner* in the study and often is couched within the context of operations—how to form/operate the partnership and benefit from higher levels of subject participation. However, something fundamental to the research process itself is altered by adopting the PAR approach: The research goals change. It is important to recognize this because it is central to the success of the community participatory approaches to research.

The research design and methods used in a health survey are selected for a single purpose: to achieve the research goals of the study. This is a basic tenant of scientific research—to choose a design and methods that efficiently and effectively achieve the research goals. The papers in this session illustrate well the PAR and traditional research approaches to community research. On more than one occasion in these papers, there is mention of a tension between the researcher and the community in terms of competing goals. This tension stemmed from the community needs to gain useful action-oriented information from a research project relative to the researchers' needs to scientifically gather information that meets

specific research questions (typically but not always in the context of a quantitative, probability-based sample survey).

The PAR approach is aimed at relieving the tension that accompanies the traditional research design. The source of this tension lies in the scientific research goals of the study being conducted. Traditional approaches adopt ***quantitative*** research goals of measuring specific constructs, estimating population parameters and associations, and monitoring trends over time—these are surveillance goals. On the other hand, PAR adds a ***qualitative*** research goal of understanding the "hows" and "whys" of certain health behaviors in local communities (to acknowledging and attending to community needs) in a way that permits development of actions at the local level. The PAR model essentially evolves and amplifies the research design so that it serves two distinct albeit complementary research goals (i.e., quantitative and qualitative goals).

By thinking about participatory research from this perspective, some interesting fallacies about traditional and participatory research approaches emerge. First, an underlying fallacy of the traditional approach is that it assumes valid constructs and measures for all members of the population, including those with known, acknowledged problems of cultural relevance and linguistic equivalence across languages. On the other hand, the underlying fallacy of the PAR model is one of human nature. Unless researchers and government agencies enter into research partnerships at the concept phase of the research project, the qualitative research goals at the local community level always will be delegated to secondary status. This means that community-based goals are ultimately optional and disposable (e.g., at the first signs of a budget overrun or a schedule delay).

To illustrate these "fallacies," consider the studies discussed in this session's presentations. The National Household Survey on Drug Abuse and the National Health Examination Survey employ a traditional approach and espouse quantitative goals of estimating and monitoring population parameters. They presume robustness in their measurement of subpopulations, including language minorities and other subpopulations, such as the elderly. Yet these subpopulations are the most challenging for gaining cooperation, and their responses exhibit higher levels of measurement error when responses are elicited. These shortcomings could have been identified through a participatory research model, and actions could have been taken to address these problems. At the other extreme, the study of *colonias* in Texas employed a PAR approach and used nonprobability sampling to generate information, suggesting that the qualitative goals dominated the research agenda. While the research was useful for the community, it does not have the same ability to generalize inferences relative to that of a quantitative research design (e.g., probability sample survey). Finally, consider a participatory research approach "in the middle"—the California Health Interview Survey. We see both qualitative and quantitative research goals pursued with concerted efforts to engage community participation, but at a price in terms of time, human resources, and dollar expenditures.

Looking at the researcher-community tensions that accompany community-based health survey research and adopting the perspective that PAR fundamentally involves the adoption of dual research goals, the following conclusions can be reached:

- Any discussion of community partnership and participation in health surveys should explicitly address the research goals of the study; to the extent they can be clarified, there will be a better understanding with which to address community-researcher tensions.
- PAR is an approach that can be implemented at varying intensities along a continuum. At one extreme is the PAR model involving full community partnership in the research, while the

other extreme is the traditional top-down approach (where the research solicits the community's cooperation with no say in the research).
- Discussion of community partnerships in health surveys should explicitly articulate the important implicit assumptions underlying the quality of the data to be collected. Central among these in the health survey research arena are the cultural relevance and linguistic equivalence of health measures.

RESEARCH ETHICS AT THE COMMUNITY LEVEL

To what extent are researchers ethically bound to provide results, information, and/or feedback to the communities that participated in the research? On the one hand, rigid human subjects protection protocols exist for individual participants, as evidenced by the mandatory Institutional Review Board (IRB) reviews and Federalwide Assurances required for federal grant research (not to mention the Privacy Act).

But what about groups of participants within a community? If an unusual disease or health risk within a specific community is discovered during data collection or through analysis, what (if any) responsibility does the researcher have to communicate this to the community? The challenge is to get relevant information to the community without immolating the confidentiality of the research subjects. Community partnerships can help to fulfill such an ethical obligation without compromising subject confidentiality because community partners are members of the research team and therefore have access to privileged information (while maintaining an obligation not to disclose individual information). As such, they can act/plan for the benefit of the community in ways that nonpartners are unable to act.

DISADVANTAGES OF THE PAR MODEL

The Participatory Action Research model, when implemented well, can effect "good science." But relative to the traditional approach, the benefits of PAR are accrued at a price:

- Research goals are altered to include those related to community benefit (many see this as an advantage, but some researchers who value a high degree of individual ownership may view this as a disadvantage).
- A larger management infrastructure (committees) is needed.
- PAR approach requires more funding to manage the larger team.
- Specialized staff (to properly manage the group processes) are needed.
- A longer and more flexible schedule is needed.
- A process to collect and digest input and effect decisions is needed.
- There is a risk of political influence or local "blow-ups."

At what point, if any, do the downsides of the PAR approach to health surveys outweigh the benefits? Cost-benefit models should be developed to help frame the appropriateness of community partnerships. This should reflect the number of communities involved, the "at-risk" subpopulations that would most benefit from partnership, and the goals/uses of the research. While there are obviously no steadfast rules, the benefits of PAR in addressing the problems of measurement error of at-risk minority subpopulations should never be underestimated. Suffice it to say that some level of community participation would benefit most if not all health survey research projects.

FINAL THOUGHTS & RECOMMENDATIONS

For many local- and state-level health surveys, community partnerships make a lot of sense. But for regional and national health surveys, especially surveys that involve longstanding replication over months or years, the adoption of a PAR approach requires strategic thinking and great care. It is wrong to dismiss community partnerships outright. It is wrong to think that a full-

fledged PAR approach is the only way of effecting community participation and benefit. In fact, it is easy to think of Participatory Action Research as a *distinct design approach*. But in fact, it represents a *continuum* of community involvement that invokes the dual research study goals mentioned earlier in this discussion. For national surveys, there is much middle ground that could be usefully exploited.

Large-scale national health surveys involve conducting data collection operations in a multitude of communities simultaneously. For these projects, it is not feasible to implement the PAR approach concurrently in multiple communities (using the traditional PAR model). The periodicity of the large-scale, ongoing data collection surveys like the National Household Survey of Drug Use can be exploited by embedding selected community partnerships at any point in time. By rotating these locations over time, the number of "community partners" will grow, reaping benefits associated with PAR at each stage.

Health survey research as an industry has increasingly become sensitized to the measurement and participation issues associated with minorities and other at-risk subpopulations. Increasingly, large-scale national health surveys employ up-front qualitative research (e.g., cognitive interviews, focus groups) to validate and refine constructs. It should be straightforward to expand qualitative research to engage community participation in identifying the community-specific threats to valid data and subject participation. Such discussions would almost always identify issues of cultural relevance and linguistic equivalence of health measures among special populations. But almost always missing is a post-survey qualitative effort aimed at gathering rich contextual data to complement the survey findings. Focus groups *after the survey* can be used to give context to survey results and tie results to policy recommendations; they can be invaluable to communities, too. This could help provide the "actionable" policy-oriented information that many communities seek and need.

REFERENCE

Minkler, M., & Wallerstein, N. (Eds.). (2003). *Community-based participatory research for health.* San Francisco: Jossey-Bass.

INTRODUCTION TO SESSION 3: Cross-Cultural Challenges in Health Survey Research

Peter Ph. Mohler, ZUMA

Comparative survey research has tended to concentrate on cross-national and international comparisons. Ironically, the success stories of social surveys covering one nation make us sometimes forget the often remarkable cultural diversities and variations found within many nations. However, the accumulation of surveys over many years has resulted in long time series data that have made it difficult to neglect cultural change within some nations. Moreover and more recently, surveys in the area of health studies have identified language and cultural barriers not seen before. It is thus timely to consider concepts and issues of cross-cultural methods in the context of ongoing health survey research in the United States and elsewhere.

Cross-cultural research is a substantive and methodological challenge still. It is an extremely complex endeavor covering all the territory of substantive research (e.g., health studies) as well as basic methodology (counting vs. enumeration), questionnaire design for multiple languages and cultures, multipopulation sampling, translation, complex analyses, etc. This session's presented papers and ensuing discussion give a fair overview of that complexity, as well as of proper solutions to reduce it to doable tasks.

Culture seems to be a vague concept; thus, it might be helpful to consider which explicit definition is used for a survey:

Cultures within a nation state
Nation state as proxy for culture
Language as proxy for culture
Food as proxy for culture
Looks as proxy for culture
Money as proxy for culture
Health as proxy for culture
Culture as proxy for culture?

Before pointing out some key issues in cross-cultural research, one might consider the level of observation on which a given survey operates.

As we all know, there is no universal optical device that allows us to look into deep space as well as into the nanoworld. Like our eyes, all optical tools have a built-in observation theory, as Popper calls it, which limits its usability. Similarly, there is no survey that allows us to observe global, international, national, regional, subregional, etc. social facts simultaneously. That we use the same or similar questions or items on all levels is no argument against the necessity to specify the target level of observation for each survey.

There are a number of key issues in this field that need consideration when doing within- or across-nation cross-cultural survey research.

The first and most important one is to have clear definitions of the concepts or constructs under observation. This is quite different from dealing with "good questions." Questions are indicators of latent constructs. Before starting translation of such questions or items, evidence indicating the equivalence of constructs has to be gathered.

A good source questionnaire (formerly politically incorrectly named "master questionnaire") already takes into account cultural diversities of the population(s) under observation. Good translation, including testing, will profit from a well-designed source questionnaire. (As an aside, one should note that back translation is no longer endorsed as a translation standard. Its place is now taken by team procedures.)

Accurate sampling frames are as crucial to cross-cultural surveys as are appropriate analytical tools that can deal with cultural

bias. Furthermore, one should not forget the difference between measurement and enumeration. Strictly speaking, statistics assume measurement (i.e., the notion of systematic and random error). This is contrary to enumeration, which, in principle, is an exact procedure without error components. This distinction is crucial in the cross-cultural comparison, because all methods used to identify equivalence and bias are built on the measurement paradigm.

As said above, within the three areas identified, close cooperation between researchers from fields as different as health-related research, linguistics, anthropology, comparative statistics, cross-cultural psychology, and survey methodology is paramount. Of course, this cross-scientific cultural collaboration is a challenge on its own, but without it, neither high quality surveys nor scientific progress can be achieved.

FEATURE PAPER: Problems in Establishing Conceptually Equivalent Health Definitions Across Multiple Cultural Groups

Janet A. Harkness, ZUMA

INTRODUCTION

Most health instruments in use around the world are Western in origin and are based on Western conceptions of health, illness, and treatment. They reflect the Western frame(s) of reference of those involved in deciding the substantive content of the questionnaires and of those deciding question formulations and formats. Not surprisingly, then, instruments often fail to accommodate respondent frames of reference; as a result, the accuracy and appropriateness of assessment is threatened. Assessments of health needs in the U.S., for example, have been found to be uncertain because of (culturally) misguided sampling designs, uncertainty about how to define groups (race, ethnicity, language, socioeconomic standing, and location all being problematic but involved), and a variety of factors affecting respondent disclosure and clinician perceptions.

HOW CULTURE COMPLICATES HEALTH ASSESSMENT

Space restrictions preclude any review of how culture complicates survey research *per se*; instead, we focus on eight key aspects for health research.

Language

Thirty years ago, Sechrest, Fay, and Zaidi (1972/1988) recognised that different groups sharing a language might need adjustments to an instrument to accommodate cultural and language differences. Warnecke et al. (1997) identified numerous difficulties encountered in administrating instruments in English to African Americans not fully explained by low socioeconomic status. Guillemin, Bombadier, and Beaton (1993) suggest that the degree of acculturation of immigrants and the type of proficiency available in a given language should determine which instrument to administer. International research, on the other hand, seems to assume that if a country is basically monolingual, language adjustments will not be needed.

Disclosure & Trust

Culturally anchored disclosure norms, degrees of trust, and what is acceptable to say to whom will affect respondent interactions with interviewers and clinicians. Western medical care values the notion of keeping patients informed so as to allow them choice, while Asian populations may feel it is better not to tell patients very bad news since this might reduce their chances of recovery. Cultures also view decision making and decision makers differently. The upshot for survey questions on preferred treatments, for example, may be that respondents consider the options offered inappropriate or the answers so obvious as not to need stating.

Survey Familiarity

We cannot assume that every community will be familiar with the kind of question-and-answer exchange involved in surveys. Unfamiliarity makes it more likely that respondents will have problems using answer scales, following instructions, or providing answers (e.g., Canales, Ganz, & Coscarelli, 1995; Sechrest et al., 1972/1988; Warnecke et al., 1997). Different groups within and across countries may not be familiar with a health instrument's technical terminology, such as references to diseases, medication, or symptoms. Lack of knowledge and non-Western explanations for disorders also lead to underreporting. Respondents may underreport illnesses that do not interfere with daily life or that are very common among the community (e.g., malaria in some countries).

Acculturation

Immigrants "acculturate" to the degree that they adopt the worldviews and living patterns of a new culture. Acculturation can be vitally important in understanding symptom expressions, rates of illness, and use of services by immigrants and refugees (Guillemin et al., 1993; U.S. Department of Health & Human Services [DHHS], 2001). Canales et al. (1995) suggest that

A more comprehensive preconference paper is available on request: harkness@zuma-mannheim.de

an acculturation measure should be included in any health assessment of immigrants.

Complaints & Symptoms

Accurate prevalence data depend on the extent to which instruments capture symptoms and the degree to which analysis or diagnosis interprets these correctly. In a standardized diagnostic interview, the presence or absence of clinically significant symptoms is investigated according to diagnostic criteria, such as those in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR; American Psychiatric Association, 2000) or the *International Statistical Classification of Diseases and Related Health Problems* (World Health Organization, 1992). However, symptoms differ in prevalence across time and culture, in particular for psychological disorders. Andary, Stolk, and Klimidid (2003) and Rogler (1999) discuss misdiagnoses resulting from attempts to interpret culture-bound syndromes (see below) in terms of disorders recognized in Western allopathic nosologies. Regier et al. (1998) argue that self-reported symptoms are insufficient for case identification and that information on severity, duration, functional impairment, and comorbidity are needed.

Somatisation is an idiom of distress in which people afflicted report symptoms of physical illness not well explained in biomedical terms. In some sociocultural groups, for example, the expression of physical pain seems to be much more acceptable than the expression of mental anguish. Somatisation has been found to be a characteristic feature of depression among numerous populations (Cheng, 2001). Female Turkish immigrants in Germany tend to experience depression first and foremost as physical pain. A U.S. Surgeon General's report (DHHS, 2001) suggests that a preferred presentation of physical distress is linked to conscious or unconscious stigma avoidance strategies.

Culture-bound syndromes (CBS) are described in the DSM-IV as "recurrent locality-specific patterns of aberrant behavior and troubling experience that may or may not be linked to a particular DSM-IV diagnostic category. Many of these patterns are indigenously considered to be illnesses or at least afflictions, and most have local names" (American Psychiatric Association, 1994, p. 844). CBS have been seen as a major impediment in establishing an international classification of psychiatric disorders (e.g., Prince & Tcheng-Laroche, 1987). Cheng (2001) is among those who argue that cross-cultural differences in the prevalence of these disorders derive mainly from culture-specific illness behavior and not from differences in the basic psychopathology.

Explanatory Models of Illness or Disease

The explanations provided for illness or disability also differ across cultures. *Naturalistic* explanations see illness as due to causes that can be understood and cured through a scientific method of discovery. *Personalistic* explanatory models relate illness or injury to supernatural forces or beings or the intentional or unintentional acts or wishes of other people. Perceived etiology impacts many features relevant for health assessment, such as expectations for treatment, compliance with treatment prescribed, satisfaction with care, and disclosure of complaints. Patients in a Western setting attributing an illness or injury to black magic, the evil eye, or a lack of life balance, for example, are likely to have their explanations discounted in the process of diagnosis and treatment. Alternatively, they could be mistakenly diagnosed as suffering from a mental disorder.

Various authors have noted the need to include religion and spirituality in conceptions and definitions of health and, in doing so, to move beyond the cultural horizon of Judeo-Christian traditions. In the U.S., indigenous American Indian and Alaska Native populations provide ready examples of very different spiritual and religious traditions, as do the Maori in New Zealand.

Biophysical Differences

Differences in drug metabolization across racial, ethnic, or other groups are useful examples of a thin line to be negotiated between being aware of possible differences and engaging in misconstrued stereotyping of difference. Since the range of differences within different sociocultural groups is considerable, it would be misplaced to decide dosage on the basis of a person's skin color or assumed ethnic identity.

Cultural bias of clinicians, interviewers, & respondents

The Surgeon General (DHHS, 2001) reports that clinicians in the U.S. prescribe more and higher doses of oral and injectable antipsychotic

medications to African Americans than they do to Whites. One explanation suggested for these findings is clinician bias. Clinicians could be predisposed to judge African Americans as schizophrenic but not as suffering from an affective disorder.

Professional and lay interviewers also have been found to be sources of bias in cross-cultural measurement through unprofessional administration of instruments (experienced interviewers not always being available), culturally biased stereotyped administration and interpretation/diagnosis, or conscious or unconscious filtering of information presented.

Respondent bias often is reflected in underreporting. Various sociocultural factors can contribute to this, a frequently cited example being the strong stigma associated with mental disorders in some communities. Underreporting or overreporting also can be related to perceptions about the goals of a project, lack of trust in anything official, unfamiliarity with the general nature of surveys, and culturally anchored needs in terms of face management.

HOW CULTURE FRAMES INSTRUMENT DESIGN & PROCESSING

Cultural anchoring often is an automatic part of the conceptual framework, design, language, and worldview behind and within an instrument. The direction in which a community reads and writes, for example, would automatically (and necessarily) determine certain aspects of instrument design. Imported instruments may fail to do this in some respects (Tanzer, in press). *Cultural bias* arises when the sociocultural framework of reference appropriate for one context is imposed as the framework of reference for a different sociocultural context. *Cultural tailoring,* on the other hand, is deliberate optimizing or adaptation of design, content, and wordings of an instrument for use with a given sociocultural group. Examples include accommodating pronominal, kinship, or gender distinctions across languages and communities (see Harkness, Pennell, & Schoua-Glusberg, 2004; Harkness, 2003; Canales et al., 1995).

Respondent processing and cultural perception: Respondents, like instruments and researchers, are rooted in their individual sociocultural contexts. Braun (2003) demonstrates how this determines how respondents read and interpret questions. Questions discussed in Kortmann (1987) from the World Health Organization Self-Reporting Questionnaire (SRQ) translated into Amharic, an Ethiopian language, illustrate how frames of reference are culture-bound. Some of the questions seem obviously inappropriate for the Ethiopian context. *Is your appetite poor?,* for example, could be problematic for a population familiar with famine. Indeed, respondents made "sense" of the item by interpreting it as a question about food availability. Other questions may appear straightforward from a Western perspective, but responses to the question *Do you cry more than usual?* revealed that Ethiopian respondents thought the question related to crying—socially required weeping—at funerals.

QUESTIONNAIRE DESIGN & EQUIVALENCE ACROSS CULTURAL GROUPS

Instruments that have proved successful elsewhere have considerable appeal. At the same time, instruments successful in one context fit that context but may not fit others. One of the most frequently mentioned obstacles to cross-cultural comparison is that instrument content and/or presentation do not allow for proper comparison across sociocultural groups and countries. Even when instruments are designed with cross-cultural implementation in mind, design procedures and outcomes are not always particularly successful. There are multiple reasons for this (cf. Harkness et al., 2003). Cross-cultural design procedures lack the intensive methodological research characteristic of (assumed) monocultural survey research. Testing instruments across groups before the questionnaire is finalized is not common. Insights from cognitive research now inform monocultural instrument development; this is rarely the case in cross-national development. Moreover, in adapting and translating instruments, efforts have focused on keeping things as similar as possible. In doing so, cross-cultural research stands in contrast to more recent moves in monocultural research towards respondent tailoring and adaptive designs (see Harkness, 2003; Harkness et al., 2003).

TRANSLATION

However simple some may look, survey instruments are difficult to translate well. This is in part because questionnaires are a complex text type administered in increasingly diverse media and mode options. In addition, instruments for

cross-cultural administration may be even more generic in reference than are instruments used for one population. Varied in character as health instruments are, the basic translation problems faced are those met in translating other survey instruments.

Numerous protocols, guidelines, recommendations, and descriptions of differing quality are available of how to conduct (health) survey translations. A number of these, with references, are considered in Harkness and Schoua-Glusberg, 1998; Harkness, 2003; and Harkness et al., 2004. We focus below on "close translation"—as directly connected with equivalence—and on the vocabulary and complexity of health assessment instruments.

Close translation: Source questionnaire items often are replications, and survey research prefers to see these closely translated. Various authors point to drawbacks and misconceptions related to this (e.g., Canales et al., 1995; Guyatt, 1993; Sechrest et al., 1972/1988). Close translation can, for example, result in stilted items, increased respondent burden, and questions being understood differently than intended or not being understood at all (see examples in Harkness et al., 2004).

Because close translation often is expected, researchers discussing (or doing) translations may wrongly focus more on words than on item meaning. For example, Andary et al. (2003) suggest that "feeling blue" is difficult to translate because the color signifies different things in different cultures and languages. However, there is no need to include a color word for *blue* in translation; a British English version of this item from the SF-36 (cf. Ware & Sherbourne, 1992) has in fact "feeling *low*." Finding an appropriate corresponding level of "down-ness" across languages certainly could be a real challenge. Authors also sometimes discuss the difficulty of matching English grammatical structures in translations (e.g., Guillemin et al., 1993; McGorry, 2000). Languages do not match in grammatical structures, and it is not clear what would be gained in trying to make them do so. However, the fact that handy little features of English carry a lot of information (e.g., gerunds, the progressive *ing* form in verbs, elliptical phrases such as *if any*) can indeed be a problem for translation.

Particular problems in translating health instruments. Arguably, culture is not more central to health research instruments than it is to other survey research instruments. However, health survey translations may call for more discipline-related technical knowledge than is needed to translate opinion polls, for instance. Since health instruments often cover a wide range of topics, multiple subject fields may be involved, calling for specialized knowledge on the part of translators. It is unlikely in such contexts that one translator would have sufficient expertise to deal confidently with all of the topics, which could include finance, care provision, retirement plans, cognitive and physical impairment tests, and physical and mental health. A team approach using consistency checks would seem ideal for such instruments. Translators and field staff may need to cooperate to ensure that the final version is also one that respondents will understand and with which they will identify.

Vocabulary of health, illness, emotions, and psychological and physical states: Many instruments elicit information about physical and psychological conditions and emotions and feelings of anxiety, fear, pain, and so forth. Cultures have different taxonomies of emotions; thus, languages and cultures differ in the delineation and expression of emotions and psychological states (cf. Lu, Gilmour, & Kao, 2001; Mesquita & Frijda, 1992; Sundberg, Latkin, Farmer, & Saoud, 1991). Words such as "depressed" and "anxious," for example, find no easy match in some American Indian and Alaska Native languages, while various notions of shame seem to occupy a prominent place in Japanese culture (Mesquita & Frijda, 1992). Even equivalent identification of bodily parts can be problematic. For example, distinctions between *stomach* and *abdomen* or *foot* and *leg* in everyday (Southern) German do not match the distinctions understood in either English or (Northern) Standard German.

CULTURAL ADJUSTMENT & EQUIVALENCE

Two examples from the Medical Outcomes Trust 36-item short-form health survey (SF-36) are used to illustrate how adjustments currently are made across different versions of instruments. An SF-36 question measuring moderate activity asks if people have difficulty *walking several blocks.* Measuring distances in terms of street blocks is a North American and presumably urban convention, reflecting a certain concept of town planning. In Great Britain, with a different concept of town planning, one speaks of *streets*, not *blocks.* In the British version of the SF-36, the

blocks unit of measurement was replaced with a unit of measurement appropriate for Britain (cf. McDowell & Newell, 1996) to ask whether respondents had difficulty *walking 100 yards.* Subjective (cultural) factors doubtless determined that walking a measure of *several blocks* was an appropriate indicator for moderate activity for the U.S. Other cultural factors presumably led British researchers to decide that they knew (1) the degree of activity intended in the U.S. item and (2) that this would correspond to *100 yards* for British respondents. However, the Swedish translation of this item took *200 meters,* about double the distance in the British version. Does this imply that Swedes commonly walk more and further and that the ceiling for 100 yards would be too high? We currently lack tested procedures and guidelines for calibrating such matters. How should researchers best compare the two or the three? And if we think of other cultural and geographical contexts, what distances or points of reference might be appropriate for Mongolian herdsmen, Bajau boat-dwellers, or other nomadic communities? A further SF-36 moderate activity item asks whether respondents have difficulty *playing golf, moving a table, pushing a vacuum cleaner, or bowling.* As illustrated below for "playing golf," scenarios assumed as universal in formulating these items may not be equally salient across cultures, or the content of a given scenario may differ significantly. In each case, measurement and comparability are potentially affected.

Golf is a fairly new pastime in Germany, is expensive, and is not widely popular. Other outdoor activities would be more salient for Germany and would avoid the distracting expense issue. For respondents in Scotland, golf-playing is both salient and more affordable. However, in picturing golf-playing in the U.S. as a "moderate activity," we might well envisage a motorized vehicle for players and equipment. In Scotland, playing golf tends to be more like a hike in windy weather with a heavy backpack. Considerations of this sort are perhaps why the Swedish version of the SF-36 replaced *playing golf* with *picking berries.* At the same time, considerable local cultural information is needed to appreciate that autumn berry-picking activities in Sweden might correspond to *playing golf* in, depending on one's view, a Scottish or a U.S. setting.

DISCUSSION

Commonly used procedures for design, translation, and testing of cross-cultural instruments differ considerably from those developed for (presumed) monocultural research (cf. Harkness et al., 2004; Harkness et al., 2003). New language version development on the basis of translation, limited field pretesting of new versions, and testing based more on *ex post facto* statistical analysis are common. Guidelines and recommendations about how to design and adapt/translate, scattered across disciplines, often are, as Guillemin et al. (1993) caution, based more on common sense than on tested research. Numerous validation procedures are available, even if those developed for one discipline may be less suitable for another.

Considerable progress has been made with regard to translation and aspects of adaptation closely related to translating. It also is generally accepted in theory that multicultural development and pretesting of instruments are important keys to improving the comparability of measurement across instruments and populations. Difficulties that arise in trying to retain "tried and tested" items in studies for new populations also are not insurmountable. Technically, options such as split ballots already provide one potential solution: By retaining the original question in one ballot and using an adapted or alternative in another, results and effects of change can be monitored. The real operational obstacles to improving comparability are, we suggest, twofold. Funding constraints on pretesting draft versions across languages and on conducting basic methodological research on cross-cultural design mean that pretesting often is forgone and basic research left for another day. Simultaneously, diverse professional pressures to adopt items or procedures employed in the past perpetuate both procedures and instruments beyond their proper shelf date.

Health research is one of a few disciplines beginning to focus on the fact that past practice is not always best practice and that more information and research are needed on cross-cultural design, adaptation, and calibration. Long-term, the recognition of problems and inadequacies in health instruments may heighten awareness of the issues in other areas of survey research. In this sense, health research may get to set a flag atop an iceberg...with the flag signaling

the accomplishment but also warning of the larger territory below the waterline.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text revision). Washington, DC: Author.

Andary, L., Stolk, Y., & Klimidid, S. (2003). *Assessing mental health across cultures*. Bowen Hills: Australian Academic Press Pty. Ltd.

Braun, M. (2003). Communication and social cognition. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 57–69). Hoboken, NJ: Wiley.

Canales, P. A., Ganz, P. A., & Coscarelli, C. A. (1995). Translation and validation of a quality of life instrument for Hispanic American cancer patients: Methodological considerations. *Quality of Life Research, 4,* 3–11.

Cheng, A. T. A. (2001). Case definition and culture: Are people all the same? *The British Journal of Psychiatry*, *179,* 1–3.

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*(12), 1417–1432.

Guyatt, G. H. (1993). The philosophy of health-related quality of life translation. *Quality of Life Research*, *2,* 461–465.

Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. *ZUMA-Nachrichten Spezial No. 3. Cross-Cultural Survey Equivalence,* 87–127.

Harkness, J. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.

Harkness, J., Van de Vijver, F. J. R., & Johnson, T. P. (2003). Questionnaire design in comparative research. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 19–34). Hoboken, NJ: Wiley.

Harkness, J., Pennell, B., & Schoua-Glusberg, A. (2004). Questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires.* Hoboken, NJ: Wiley.

Kortmann, F. (1987). Problems in communication in transcultural psychiatry: The self-reporting questionnaire in Ethiopia. *Acta Psychiatrica Scandinavia*, *75,* 563–570.

Lu, L., Gilmour, R., & Kao, S.-F. (2001). Cultural values and happiness: An East-West dialogue. *The Journal of Social Psychology, 131*(4), 477–493.

McGorry, S. Y. (2000). Measurement in a cross-cultural environment: Survey translation issues. *Qualitative Market Research: An International Journal, 3*(2), 74–81.

Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: A review. *Psychological Bulletin, 112*(2), 172–204.

Prince, R., & Tcheng-Laroche, F. (1987). Culture-bound syndromes and international disease classifications. *Cultural Medical Psychiatry, 11*(1), 3–52.

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. et al. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders. *Archives of General Psychology, 55,* 109–115.

Sechrest, L., Fay, T. L., & Zaidi, S. M. (1988). Problems of translation in cross-cultural communication. In L. A. Savomar & R. E. Porter (Eds.), *Intercultural communication* (5th ed., pp. 253–262). Belmont, CA: Wadsworth. (Reprinted from *Journal of Cross-Cultural Psychology,* 3(1), 41–56, 1972)

Sundberg, N. D., Latkin, C. A., Farmer, R. F., & Saoud, J. (1991). Boredom in young adults—Gender and cultural comparisons. *Journal of Cross-Cultural Psychology*, *22*(2), 209–223.

Tanzer, N. K. (in press). Developing tests for use in multiple languages and cultures: A plea for simultaneous development. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Erlbaum.

U.S. Department of Health and Human Services. (2001). *Mental health: Culture, race, and ethnicity—A supplement to Mental Health.* A Report of the Surgeon General. Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services.

Ware, J. J., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30,* 473–483.

Warnecke, R., Johnson, T. P., Chávez, N., Sudman, S., O'Rourke, D., Lacey, L. et al. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7,* 334–342.

World Health Organization (1992). *International statistical classification of diseases and related health problems* (10th ed.). Geneva, Switzerland: Author.

FEATURE PAPER: Overview of Methods for Developing Equivalent Measures Across Multiple Cultural Groups

Gordon Willis, National Cancer Institute

Survey methodologists have increasingly sought to establish that our measures exhibit cross-cultural equivalence (Gerber, 1999; Harkness, Van de Vijver, & Mohler, 2003; Stewart & Nápoles-Springer, 2000; Warnecke et al., 1997). The general goal of establishing measurement equivalence across respondent groups is not a novel challenge but represents the blossoming of a longstanding problem that becomes somewhat more visible when we explicitly face cross-cultural issues such as language translation. In this paper, I review the empirical methods that are increasingly being used to develop survey questionnaires that purport to provide equivalent information across cultures.

DEFINING "EQUIVALENCE"

Behling and Law (2000) and Johnson (1998) cite a wide range of conceptualizations of measurement equivalence (Johnson lists 52 separate varieties), and considerable attention now is being paid particularly to *conceptual equivalence* (Harkness et al., 2003). From a cognitive point of view, this characteristic mainly concerns question comprehension—in simple terms, whether an item means the same thing across cultural groups. The general topic of comprehension is of course vital in survey pretesting. However, methods for cross-cultural questionnaire development should be (and generally are) concerned with a wider set of issues, encompassing the range of cognitive processes that may contribute to response error. Following Tourangeau (1984), Warnecke et al. (1997) have proposed that these also include respondent *retrieval*, *judgment* (sometimes referred to as *decision*), and *response editing* processes. For example, group norms may differ with respect to guessing behavior, such that the decision stage is implicitly involved as a potential source of nonequivalence.

Beyond the strictly cognitive stages of processing, instrument developers also need to focus on potential sources of response error that are not at their source cognitive in nature but have been variably classed as *Logical* or *Structural* (Conrad & Blair, 1996; Willis, Royston, & Bercini, 1991). To ensure measurement equivalence of constructs, we must not only ask questions that are successfully cognitively processed but also *ask the right questions in the first place*, in a manner that takes into account the truthfulness of underlying assumptions. For example, Ainsworth (2000) suggests that survey measures of physical activity among women are biased because they ask inappropriate questions; in particular, questions strongly oriented toward leisure-time activity do not apply well to lower-income Hispanic women—even when such respondents may have no problem comprehending the questions. As such, especially for questions asking about behavior, we must ensure that our methods assess not only the way that people *think* but also what they *do*.

LINGUISTIC VS. CULTURAL EQUIVALENCE

Cross-cultural equivalence sometimes is assumed to pertain to *linguistic equivalence* of translations (e.g., English to Spanish), and developmental methods often strive to ensure that translations are correct in the sense that meaning is retained across language versions. Beyond this, however, lies the much broader realm of *cultural equivalence*, which transcends language (Stewart & Nápoles-Springer, 2000). Some cross-cultural studies have been conducted entirely in English yet have documented potential sources of nonequivalence (e.g., Warnecke et al. 1997). For purposes of questionnaire development and pretesting, both linguistic and conceptual equivalence must be addressed.

METHODS FOR THE ESTABLISHMENT OF CROSS-CULTURAL EQUIVALENCE

Development-pretesting methods that address the above issues fall into three general categories: (1) expert review (including focus groups), (2) cognitive testing, and (3) behavior coding (Table 1).

Expert Review Methods

Expert review (i.e., appraisal) carries the connotation of either questionnaire design expertise or subject matter expertise. As applied to cross-cultural studies, this term applies more widely to any review that takes into account linguistic or cultural aspects of the survey process (I include focus groups, despite the common view that this is a unique pretesting method). The realm of expert review can be divided into two components: (1) linguistic expertise, in terms of the translation language(s) (e.g., Spanish, Korean), and (2) cultural expertise.

Table 1. Methods Used to Develop Equivalent Survey Questions

Focus of method	Category of Method		
	Expert review	Cognitive interviewing	Behavior coding
Linguistic equivalence (of translations)	Language review by expert (or team)	Multilingual cognitive interviewing	Multilingual behavior coding
Cultural equivalence: Cognitive/ Structural/ Motivational	Cultural review by expert(s), "Informant," or focus group	Cognitive or ethnographic interviewing, used cross-culturally	Behavior coding across cultural groups

Translated instruments

Again, for purposes of evaluating drafts of language-translated instruments, a major focus is cognition, and in particular, comprehension. Problems of translation failure have been recognized for many years, and although the practice of back translation has been popular as a remedy, its overuse has been criticized (de la Puente & Pan, 2003; Harkness et al., 2003; McKay et al., 1996). The U.S. Census Bureau, for instance, is developing a set of best practices for translation that relies on the use of an expert team ("Committee") approach (de la Puente & Pan, 2003). Currently, a favored practice in the U.S. consists of an *expert-based translation review* team that is sufficiently varied with respect to the linguistic variety represented by the survey (e.g. Puerto Rican, Mexican, Cuban), optimally including an individual who has familiarity with questionnaires and who may have field experience in survey administration, such as a bilingual Census Bureau Field Representative (McKay et al., 1996).

As an example of expert-based translation, Kudela et al. (2003) report that Chinese and Korean translations of an English tobacco questionnaire were too literal, resulting in questions that were wordy and even confusing. The Chinese version was too formal in places, and some questionnaire items were awkward because English rather than Chinese grammar was used. Conversely, the Korean translation was felt to be not formal enough.

Expertise in cultural issues

Questionnaire developers also encounter a variety of issues related to nonlinguistic cross-cultural equivalence that literal translation may not address. To address these, it is useful to incorporate a *cultural review*, whether integral to the translation process or a separate activity. From the earliest stages, questionnaire designers actively take into account the degree of cultural variation likely to be represented by the target population, and that involves cognitive issues related to question processing, general social norms, and question structure issues. Culturally relevant reviews can be done through focus groups involving members of the appropriate populations in order to investigate concepts that may be group-specific (e.g., conceptualization of "working for pay" among groups that may engage in alternative systems, such as barter).

To conduct a culturally oriented expert review, one can seek background knowledge concerning the manner in which members of different groups tend to react within the

somewhat contrived question-answering process presented by the standardized survey. Increasingly, this information is being included in the literature devoted to Cognition and Survey Methodology (CASM). For example, cultures that are strongly collective may eschew the expression of individual opinions, relative to individualistic cultures (Ji, Schwarz, & Nisbett, 2000), or cultural groups may differ in their use of extreme response categories (Warnecke et al., 1997). Useful examples also abound of the need for cultural review to understand social norms of conduct that might influence survey responses. Pan (2003) has addressed issues of respondent contact, rather than the equivalence of questions per se, because a key requirement for overall measurement equivalence is to obtain sufficient response rates across groups. She explains how Chinese-language interviews may require a different form of greeting than those done in English and illustrates a case study in which a Chinese interviewer de-emphasized her own name and identity relative to the survey sponsor's when approaching a Chinese respondent. Designers need to understand and adjust to such cultural differences in social interaction. To the degree that our objective is to complete interviews in a way that is not alienating or even insulting, the best means may not be to ensure literal equivalence of introductory scripts but to focus on the *equivalence of outcome* (e.g., willing respondents), as opposed to *equivalence of process* (e.g., approaching all respondents in a uniform manner).

We also can apply cultural review to address structural defects that are inherent in questions. To address the example of cross-cultural physical activity cited earlier, designers must consider these activities across as wide a range of the relevant population as possible. As such, cultural appropriateness needs to be considered earlier rather than later in the development process. A practical result of this is increased attention to *decentering*, which is usually meant to indicate that the source (e.g., English) language version is not considered inviolate but rather is open to modification itself[1] (McKay et al., 1996). This practice could be applied widely, so that not only the linguistic expression of the questions but the questions *themselves* consider the full range of respondents to whom they will be administered (e.g., by determining all the varieties of behaviors across groups that involve physical activity).

Cognitive Interviewing Methods

Cognitive interviewing is well accepted in the field of questionnaire development and pretesting (Tourangeau, Rips, & Rasinski, 2000). Because it emphasizes the entire range of issues discussed above, the extension of cognitive interviewing to cross-cultural studies seems well indicated. Especially since the landmark study by Warnecke et al. (1997), cognitive interviewing in the U.S. has increasingly been applied to assess potential sources of response error across cultural groups by actively including African-American, Hispanic, and Asian subjects in the testing process (see Miller, 2002). However, cross-cultural cognitive interviewing presents a number of challenges, especially in terms of how it should be conducted and what it is able to uniquely provide. Work done to date suggests that cognitive interviews may be particularly useful for the study of comprehension of complex terms, especially in cases where the interpretation of general, superordinate concepts may vary because these include different sets of implicit exemplars. For example, Warnecke et al. (1997) examined cross-cultural interpretation of physical activity. In their words, "Variation by race/ethnicity was found in respondent interpretation of what constitutes physical activity through probes asking whether they considered walking, household, work-related activity, and yard work to be physical activities" (p. 337). Such results may have serious ramifications when we rely only on very general and variably understood terms, as opposed to inquiring about specific activities.

[1] Alternatively, de la Puente and Pan (2003) use the term *adaptation* to refer to decentering activities.

As a potential alternative to cognitive interviewing, an active movement toward methods development is the *ethnographic interview,* an extension of the cognitive interview having an explicit cultural emphasis (Gerber, 1999). The ethnographic interviewer focuses on the background assumptions made by various questioning approaches and investigates the manner in which key concepts are subject to cultural variation. Ethnographic interviews can be conducted early in the development process, such that the researchers can fashion questions that are suitable, in cross-cultural terms, for further cognitive testing. However, Gerber points out that even within usual forms of cognitive testing, a considerable amount of "back-up ethnography" may be conducted. For example, it is easy to imagine a cognitive interview result consisting of the comment, "These questions on physical activity don't pertain to my female Hispanic subjects." Hence, attention to culture can permeate classical cognitive interviews, as it does those that are specifically ethnographic.

Logistical challenges to cross-cultural multilingual cognitive testing

The extension of cognitive techniques to studies involving translated instruments presents several additional challenges:

(1) *Recruitment issues*, as the study of multiple language groups may necessitate inclusion of individuals who are relatively unacculturated to the normative society;
(2) *Generalizability*, due to use of small samples;
(3) *Potential unreliability of results*, due to investigator "clinical judgment" concerning the nature and seriousness of observed problems; and
(4) *Appropriate staffing*, as it may be challenging to find and train appropriate bilingual individuals who also have familiarity with questionnaires and can be trained to conduct cognitive interviews.

Despite these challenges, Kudela et al. (2003) relied on cognitive testing to assess the functioning of tobacco use questions when translated to Korean, Chinese, and Vietnamese and obtained several findings relevant to cross-cultural equivalence, with respect to both translation and cultural issues. For Korean interviews:

(1) Several questions were mistranslated outright—a question asking whether the individual had smoked 100 or more cigarettes left off the phrase "in your entire life," leading the subjects to interpret this as "at one time."
(2) Other translation problems were subtler, as where a question asking about how long one waits in the morning before smoking the first cigarette of the day used a Korean phrase that required simplification.
(3) A question on switching from a stronger to lighter brand presented difficulties for subjects who had first smoked Korean cigarettes that contained no labeling concerning tar/nicotine content.

Significantly, however, many of the Kudela et al. results did not explicitly concern issues of language or culture but instead basic problems in the survey questions that also had been exhibited when testing in English. For example, a question that asked "*What is the total number of years you have smoked every day*?" and was accompanied by a follow-up phrase instructing subjects to exclude periods they had not smoked for six months or longer caused confusion for all groups. In addition, the meaning of the term "community" was vague in the question "*How easy is it for minors to buy cigarettes and other tobacco products in your community?"* This also had been found in English-language interviews of White non-Hispanics. Overall, the finding that some cross-cultural cognitive testing results were general in nature is significant and in some ways reassuring, as it suggests that this method produces reliable results.

Further, although some findings were novel and did not mirror results from earlier English (or Spanish) interviews, they did not seem particularly "cultural" in nature. For a

hypothetical question attempting to measure nicotine addiction by asking whether the respondent would go out in a heavy rain to get cigarettes, subjects objected that because they always had enough cigarettes on hand, this would be unnecessary. It is questionable that this reaction can be viewed as particularly "Asian" at core, as it is likely that a similar response would be provided by a range of smokers. Given that several of the unique problems in non-English cognitive interviews seemed generic rather than culturally specific, one can make the case that an important benefit of these interviews is that they provide an extended test of the instrument *in general*, beyond assessing culturally-specific issues. This finding again supports the use of a decentered development approach in which all interviews are viewed as equal potential contributors.

Behavior Coding

Behavior (or interaction) coding is somewhat more quantitative than cognitive interviewing, and involves the analysis of interactions between the interviewer and respondent in search of *overt* indications that difficulties may exist (Cannell, Fowler, & Marquis, 1968; Fowler, 1995). Difficulties are identified for both interviewers (e.g., failure to read questions correctly) and respondents (e.g., requests for re-reading of the question, request for clarification, uncodeable responses, or qualifying to express uncertainty). For example, if asked "*Would you say your health in general is excellent, very good, good, fair, or poor?*" a respondent may answer "*I'd say, I don't know – maybe very good….*" This would be coded as "qualified." Generally, at least 50 interviews are coded in this way on a question-by-question basis, and the codes are tabulated so that researchers can determine whether particular questions produce relatively high code frequencies and can be considered candidates for further attention.

Because survey researchers usually are not well versed in all languages of administration and have little direct means for knowing how the translated versions operate (as they certainly cannot make use of usual forms of interview monitoring), behavior coding is especially attractive as a means by which to carry out quality control within cross-cultural investigations. As a foray into the systematic and quantitative coding of behavior by multilingual coders, a collaborative project is now underway to assess cross-cultural and cross-language operation of the 2003 California Health Interview Survey in English, Spanish, and Korean.[2] In extension to the cross-cultural domain, the investigators decided to resurrect a code first applied by Cannell et al. (1968) involving extraneous conversation to determine whether this occurs to differing extents across language versions. The CHIS behavior project represents a technological advance as well, as it makes use of digital recording of interview segments suspected to produce problems, which enables quick and efficient coordination of an electronic version of the questionnaire, the interview segment being listened to and coded, and an electronic coding form.

The CHIS behavior coding study will address several questions pertaining to the effectiveness of cross-cultural behavior coding:

(1) To what extent will respondents consent to having their interviews recorded, especially across cultural groups?
(2) What logistical problems are presented in the selection, training, and monitoring of behavior coders across cultural and language groups?
(3) Will observed differences be interpretable in terms of interviewer-respondent interactions across group, as opposed to the result of idiosyncratic differences in behavior between coders?

CONCLUSION

As a means for establishing cross-cultural equivalence, each of the methods described—expert review, cognitive interviewing, and behavior coding—is unproven but promising. Overall, there is likely to be no single pretesting method that can be regarded as

[2] By staffs of Westat, Inc., The Public Health Institute, UCLA, and NCI.

"best." Rather, the various methods will have different roles in the overall questionnaire development process.

REFERENCES

Ainsworth, B. E. (2000). Issues in the assessment of physical activity in women. *Research Quarterly for Exercise and Sport, 71*, 37–42.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* London: Sage.

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

Conrad, F., & Blair, J. (1996). From impressions to data: Increasing the objectivity of cognitive interviews. In *Proceedings of the Section on Survey Research Methods* (pp. 1–10). Alexandria, VA: American Statistical Association.

de la Puente, M., & Pan, Y. (2003). *An overview of proposed Census Bureau Guidelines for the translation of data collection instruments and supporting materials*. Paper presented at the meeting of the Federal Committee on Statistical Methodology, Arlington, VA.

Fowler, F. J. (1995). *Improving survey questions*. Thousand Oaks, CA: Sage.

Gerber, E. (1999). The view from anthropology: Ethnography and the cognitive interview. In M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, & R. Tourangeau (Eds.), *Cognition and survey research*. New York: Wiley.

Harkness, J. A., Van de Vijver, F. J. R., & Mohler, P. Ph. (Eds.) (2003). *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Ji, L. J., Schwarz, N., & Nisbett, R. E. (2000). Culture, autobiographical memory, and behavioral frequency reports: Measurement issues in cross-cultural studies. *Personality and Social Psychology Bulletin, 26,* 586–594.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. *ZUMA— Nachrichten Spezial, 3*, 1–39.

Kudela, M. S., Kevin, K., Tseng, M., Hum, M., Lee, S., Wong, C. et al. (2003). *Tobacco Use Cessation Supplement to the Current Population Survey Chinese, Korean, and Vietnamese translations: Results of cognitive testing*. Final report submitted to the National Cancer Institute, Rockville, MD.

McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M. et al. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation, 70,* 93–105.

Miller, K. (2002). *The role of social location in question response: Rural poor experience answering general health questions.* Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Pete Beach, FL.

Pan, Y. (2003). *The role of sociolinguistics in the development and conduct of Federal surveys.* Paper presented at the meeting of the Federal Committee on Statistical Methodology, Arlington, VA.

Stewart, A. L., & Nápoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care, 38*(9 Suppl. II), II-102–II-124.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J. M. Tanur, & R. Tourangeau (Eds.), *Cognitive aspects of survey design: Building a bridge between disciplines* (pp. 73–100). Washington, DC: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Warnecke, R. B., Johnson, T. P., Chávez, N., Sudman, S., O'Rourke, D. P., Lacey, L. et al. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology, 7*, 334–342.

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology, 5*, 251–267.

FEATURE PAPER: Cross-National Comparisons of Disease Prevalence: Asthma in America and Europe

John M. Boyle, Schulman, Ronca, & Bucuvalas, Inc.

OVERVIEW

Health surveys continue to provide the most widely accepted estimates of disease prevalence and characteristics of the patient population within politically defined geographic areas. In the United States, the National Health Interview Survey (NHIS) was developed in the 1950s to provide policy makers with population statistics on health. A number of other countries have since adopted annual or periodic national health surveys for the same purpose. Nonetheless, cross-national comparisons of rates of disease, characteristics of the patient population, treatment, and health outcomes are sparse and frequently problematic. In the absence of equivalent measures and procedures across nations, it is difficult to evaluate whether apparent cross-national differences reflect differences between populations or methods.

Since 1998, national surveys of asthma have been conducted in North America, South America, Europe, and Asia as part of the Asthma Insights and Realities (AIR) project sponsored by GlaxoSmithKline (GSK). The original study, Asthma in America, was conducted by Schulman, Ronca, & Bucuvalas, Inc. (SRBI) for GSK as a national telephone survey of 41,000 households to identify and interview 2,500 persons with current asthma. The following year, the survey was replicated in seven countries in Western Europe as Asthma Insights and Realities in Europe (AIRE) by SRBI for GSK. In each nation, the screening, selection, and interviewing schedules remained the same, while the sampling and data collection procedures were kept as equivalent as possible.

The AIR surveys were designed to provide national probability samples of persons with asthma, not national estimates of the incidence or prevalence of the disease. Nonetheless, since detailed records were kept of all screening interviews, which included the number of persons in all households screened and the number of persons who met disease criteria, these surveys do provide an almost unparalleled measure of the comparative prevalence of this medical condition across a number of different countries based upon equivalent samples and equivalent measures. The similarity and differences in these comparative rates of asthma may help us to better understand the issues of cross-national comparability of self-reported medical indicators of disease and wellness.

DEFINING ASTHMA: ASTHMA IN AMERICA

The Survey of Asthma in America, conducted between May and July 1998, included a national sample of the American public (*N*=1,000), a national sample of physicians (*N*=500), a national sample of nurses (*N*=100), and a national sample of pharmacists (*N*=100). The core of the project, however, was a national sample of approximately 2,500 persons (or parents of children) with asthma.

The national sample of persons with asthma was identified by random-digit-dial screening of a national sample of approximately 41,000 households. There were three key elements in the operational definition of asthma that could affect eligibility rates (i.e., prevalence rates) for the survey. First, a household informant was used for household enumeration and disease screening. Second, the condition was defined by diagnosis rather than symptoms.

The household informant initially was asked whether anyone in the household had ever been diagnosed with asthma. A follow-up probe for all negative responses identified only 57 additional households with cases of diagnosed asthma. However, the secondary screen identified another 637 households with someone who had asthma but had never been diagnosed with it. Hence, the requirement for

"diagnosed" asthma excluded nearly 10% of households reporting any persons with asthma in the household.

Third, the Asthma in America survey limited "asthma" to current or past year asthma. The household informant was asked whether any of the persons diagnosed with asthma either (1) had asthma symptoms in the past 12 months or (2) were currently taking medication for asthma. Among the 5,786 households with persons with diagnosed asthma, the household informant most commonly reported that an asthma sufferer had both current medication and asthma symptoms in the past 12 months (47%). Only 7% reported asthma symptoms in the past 12 months but no current medication. In contrast, 16% reported current asthma medication but no symptoms in the past 12 months. The remainder reported neither current asthma medication nor past-year asthma symptoms (27%) or failed to answer these two screening questions (2%).

Based on screening 41,235 households with a total of 114,247 persons living in those households, the Asthma in America survey yielded a population prevalence rate of 4.7% for current diagnosed asthma. This prevalence rate is somewhat lower than the 5.7% rate for asthma found in the NHIS. Differences between the two surveys that could contribute to this difference include

- The mode of interview and sampling frame (telephone versus in-person);
- Household screening versus embedded question about asthma;
- Inclusion of undiagnosed asthma in the NHIS; and
- Inclusion of medication-using but nonsymptomatic patients with diagnosed asthma in Asthma in America.

Indeed, given these differences in methodology, the similarity in estimated rates of current asthma between the two surveys is remarkable. The main point, however, is that while Asthma in America does not purport to yield the true prevalence of asthma in the United States, the methodology used in that survey yielded an estimate that was very close to the accepted rate from the NHIS.

DIAGNOSED MEDICAL CONDITION: LIMITATIONS OF THE HEALTH SURVEY CRITERION

A parallel national survey of adults was conducted as part of the Asthma in America project. Respondents to this national random-digit-dial survey of 1,000 persons age 18 and older were asked whether they had certain respiratory symptoms in the past 4 weeks or 12 months. If a respondent had experienced any of these symptoms in the past year, they were asked whether they had ever seen a doctor about those symptoms. If they had, they were asked what they were told by the doctor. All respondents who did not report an asthma diagnosis, including those who did not report asthma symptoms in the past 12 months, were asked whether they had ever been diagnosed as having asthma. A total of 12.5% of the national public sample reported having been told by a doctor that they had asthma in response to one of these two questions. Using prescription and over-the-counter medication use among those diagnosed with asthma to classify persons with current asthma, 7.4% of adults in the public survey were classified as having current diagnosed asthma.

Since the methods and measures for assessing current and lifetime asthma differed between the public and patient surveys, the differences in the prevalence rates are not very enlightening. What the public survey does illuminate is the problem of symptomatic but undiagnosed asthma in the population. Prior to any discussion of asthma, all survey respondents were asked whether they had four types of asthma symptoms in the past 4 weeks, and another four types of asthma symptoms in the past 12 months. Virtually all persons identified as having current asthma (96%) reported one or more of these conditions, as did 65% of those classified as having past asthma and 51% of those never diagnosed with asthma.

Although there was a correlation between diagnosis and number of symptoms in the

past year, the results suggested a considerable potential for symptomatic but undiagnosed asthma in the general population. The average number of asthma symptoms in the sample of persons with current diagnosed asthma was 4.4. Nearly seven out of ten (69%) persons with current diagnosed asthma reported four or more of these asthma symptoms. By contrast, 19% of those classified as past asthma and 10% of the sample who had never been diagnosed with asthma reported four or more of these symptoms in the past year. Hence, the population segment with four or more asthma symptoms in the past year that had never been diagnosed with asthma was actually larger than the segment of diagnosed patients who were currently being treated for asthma.

Persons who reported one or more types of asthma symptoms in the past 12 months were asked whether they had ever seen a doctor about these symptoms. The majority (53%) reported that they had never done so. As noted earlier, most estimates of disease prevalence from health surveys are based on the respondent's report that he or she has a specific medical condition. If respondents have not sought treatment for symptoms, they are unlikely to know the diagnosis for the medical condition responsible for the symptoms. There is no real solution to the problem of underdiagnosis of medical conditions because few clinicians would accept an exclusively symptomatic definition of the disease. Nonetheless, there appears to be a relatively large population of adults in the United States with asthma symptoms who have never sought medical advice about these symptoms. To the extent that treatment seeking varies among population subsets, either within or between countries, prevalence rates based upon reported diagnosis are likely to vary.

ASTHMA INSIGHTS & REALITIES IN EUROPE

The methodology for Asthma in America was replicated in a subsequent cross-national survey of Asthma Insights and Realities in Europe (AIRE). The AIRE survey was conducted in March and April 1999 in seven Western European countries: the United Kingdom, France, Germany, the Netherlands, Sweden, Spain, and Italy. The household telephone penetration rate was equivalent to that in the U.S. in five of these countries and over 80% in Spain and Italy, so the same sampling and interviewing protocol could be used.

A total of 73,880 households were screened across the seven countries in order to identify and interview a sample of 400 persons with current asthma in each of these countries. Using identical measures and procedures, the survey found significant variation in the comparative prevalence of asthma in these countries. The rate of diagnosed asthma in households, which had been 14% in the U.S., was 20% in the United Kingdom, 12% in Sweden, 9% in France, 7% in the Netherlands, 6% in Italy and Spain, and only 4% in Germany.

The household prevalence of current asthma (medication or symptoms in the past year) varied by country, as did the household rate of lifetime asthma. What is notable, however, is the similarity in the rates of current asthma as a proportion of lifetime asthma. In the United States (71%), Germany (69%), Sweden (69%), Italy (70%), and Spain (69%), the proportion of households with lifetime asthma that have current asthma were virtually identical. The rate of current asthma as a proportion of lifetime asthma was a little higher in the United Kingdom (76%) and the Netherlands (77%), and a little lower in France (67%).

The comparative population prevalence of diagnosed, current asthma in the United Kingdom was 5.7%. This was about twenty percent higher than the U.S. rate of 4.7% in the Asthma in America survey. The comparative population prevalence of current asthma in Sweden (3.3%) was nearly fifty percent less than in the United States. The population rates of current asthma were less than half of the U.S. rates in France (2.1%) and the Netherlands (2.0%). The population rates of current asthma in Italy (1.3%) and Spain (1.3%) were approximately a quarter of the

U.S. rates. The population rate of current asthma in Germany (0.9%) was less than a fifth of the U.S. rate.

A sample of 400 households that screened out in Germany was rescreened, but no false negatives were found. An independent sample of 500 households in the U.K. was screened by a separate field organization, which found identical rates of diagnosed and current asthma. Hence, the screening procedures appear to be reliable, and the cross-national differences in reported asthma prevalence are real.

COMPARISON TO OTHER SOURCES

The International Study of Asthma and Allergies in Childhood (ISSAC) provided an external source of comparison for prevalence rates from AIRE. The ISSAC study was limited to children age 6–7 and 13–14, while the sampling was limited to schools within selected communities within each country. The samples of 6- and 7-year-olds and 13- and 14-year-olds in the AIRE country samples are too small to permit direct comparison to the lifetime prevalence of asthma from the ISAAC survey. Nonetheless, the ISAAC data appears to provide support for and confirmation of the AIRE data. First, the ISAAC data confirms that the prevalence rate for diagnosed asthma varies considerably in Western Europe. The lifetime prevalence of asthma among 6- and 7-year-olds was six times higher in the United Kingdom (22.9%) than in Germany (3.6%), while the lifetime prevalence of asthma for 13- and 14-year-olds was nearly four times higher in the United Kingdom (20.7%) than in Germany (5.7%).

Second, the order of prevalence among countries in the ISAAC study was virtually the same as found in the AIRE survey. The United Kingdom had the highest prevalence. The rate in France was about half the rate in the United Kingdom but higher than the rates in Italy and Spain. Also, the rate in Germany was about half the rate found in Italy or Spain.

The actual lifetime and current rates of asthma in the AIRE and ISAAC were, of course, different. This should not be surprising, given the difference in study populations, sampling frames, mode of administration, and survey measures. What is notable is that two independent sources of population-based asthma prevalence data agree on the high degree of variability of these prevalence rates across these countries, the relative ranking of these countries on asthma prevalence, and even the general order of magnitude of difference in asthma prevalence rates between countries.

Together, these cross-national studies confirm that there are real and substantial differences in the population rates of diagnosed asthma among highly developed North Atlantic nations. They also confirm that cross-national studies that use comparable measures and study protocols are able to yield estimates of the comparable prevalence of diagnosed diseases. However, the outstanding issue is whether the differences reflect true differences in the underlying rate of disease in these populations or something else.

SOURCES OF CROSS-NATIONAL VARIABILITY IN PREVALENCE

The findings from the public survey in Asthma in America suggest that there might be a substantial amount of undiagnosed asthma in the United States. If treatment-seeking practices for asthma symptoms vary among the populations of these eight countries, this could yield substantial differences in prevalence rates. Some differences in the characteristics of asthma patients in the eight countries are suggestive of diagnosis-related issues. Using a common measure of symptom severity, the survey suggests that Germany had a higher proportion of asthma patients with severe persistent asthma (26%) than the U.S. (19%) and the U.K (18%), and a substantially smaller proportion of patients with mild intermittent asthma (33%) than the U.S. (44%) or the U.K. (46%). The hospitalization rates for asthma were highest in Europe for the three countries with the lowest disease prevalence—Germany (10%), Spain (10%) and Italy (9%). Germany also had a larger proportion of asthma patients who were adults and one of the oldest average age at diagnosis. This suggests

that it may take longer or require more severe symptoms for a diagnosis of asthma to be made in some countries than in others.

Underdiagnosis related to symptom severity may be a result of differences in treatment seeking by patient or of disease recognition by physicians. There are substantial differences among countries in the medical care systems that could affect access to care and/or access to specialized care. The survey data suggests an unexpected correlation between the differences in prevalence rates and the pattern of medical care for asthma in these countries. The country with the highest proportion of primary care for asthma, the United Kingdom (88%), also had the highest rate of diagnosed current asthma. The countries with the lowest proportion of primary care for asthma—Germany (54%), Spain (51%), and Italy (49%)—also had the lowest rates of asthma.

A second cross-national survey of disease in many of these same countries, Confronting COPD [Chronic Obstructive Pulmonary Disease] in North America and Europe, provides some further clues to the reason for cross-national differences in disease prevalence reported in health surveys. The COPD survey was conducted by telephone among national RDD samples of all of the AIRE countries except Sweden. Similar procedures were used in both surveys to identify persons with these respiratory conditions, except that the COPD survey included undiagnosed cases that met the symptomatic definition of COPD.

Unlike asthma, the comparative prevalence of COPD among households with persons age 45 and older was almost identical across six of the eight study countries. The comparative prevalence of COPD (including symptomatic but undiagnosed COPD) was 5.7% in the United States, 5.7% in Canada, 6.3% in the United Kingdom, 5.9% in France, 6.1% in Italy and 5.7% in Spain. The household prevalence of COPD actually was somewhat higher in Germany (7.6%) and the Netherlands (8.6%).

The similarity in prevalence rates across countries for COPD, however, masked substantial differences in the diagnostic components by country. Among those who met the operational definition of COPD, nearly a third (32%) had been diagnosed with COPD in the United States, compared to 21% in Canada and the United Kingdom and 15%–16% in other countries. The proportion of the total COPD population, as operationally defined in the survey, with undiagnosed symptomatic COPD ranged from 9% in the United States to 30% in the United Kingdom. In other words, if we were to compare the prevalence of any of the component diagnoses of COPD across the eight countries, we would see a variability of prevalence rates similar to that seen for asthma. Only when the four component diagnoses are pooled and an undiagnosed symptomatic component added, do we find similar prevalence rates across these eight countries in North America and Western Europe.

These differences suggest that medical labeling of diagnosis may contribute to cross-national differences in reported disease prevalence in health surveys. Survey reports of medical conditions are premised on patient knowledge of the nature of their conditions. In the COPD study, national samples of approximately 100 physicians in each country were asked whether they use the term COPD when discussing the diagnosis with their patients. The proportion of doctors who said that they used the term COPD with their patients with the condition ranged from 81% in the U.S., 76% in Canada, and 77% in the Netherlands; to 61% in the United Kingdom and Germany; to 36% in France and 33% in Italy. Hence, survey estimates of the population prevalence of COPD based on patient reports of that diagnosis will vary significantly between these countries, even when underlying rates of the medical condition do not.

DISCUSSION

The findings from the Asthma Insights and Realities in Europe (AIRE) survey continue to support the longstanding conclusion that surveys yield reliable population estimates of medical conditions.

Moreover, a comparison of the AIRE findings with those from parallel surveys of children and young adults in many of the same countries suggests that surveys with equivalent designs, procedures, and measures yield reliable estimates of comparative prevalence in a cross-national design.

Nonetheless, the findings of the cross-national surveys of asthma and COPD re-emphasize some important conceptual issues in the design and analysis of cross-national surveys of disease prevalence. Prevalence surveys usually are based on self-reported medical diagnosis. In most cases, patient reports of the nature, pattern, and frequency of symptoms do not yield clinically acceptable measures of diagnostic category. By contrast, it is assumed that patient-reported physician diagnosis of a medical condition implies clinical testing and verification of the condition.

The problem of differential rates of diagnosis for underlying medical conditions is well known in studies conducted in the United States and elsewhere. While disease rates may be higher in low-income and minority populations as a result of differences in access to care and quality of care, the corresponding rate of diagnosed conditions may be lower than in the general population. Age, gender, education, and other factors that influence treatment seeking also may influence the rate of diagnosis relative to the rate of disease in particular population segments. The medical specialization and age of the physician also may impact the likelihood of diagnosis and disclosure of diagnosis to the patient that the physician ensures that the patient understands the diagnosis of his/her condition.

Differential diagnosis may introduce even more variance in international studies. A national emphasis on a particular condition may produce greater disease awareness among health care providers, greater treatment seeking among patients with symptoms, and higher rates of diagnosis in the population. On the other hand, national differences in professional standards for the diagnosis, labeling, and disclosure of medical conditions to patients may produce national differences in the prevalence rates for diseases whose prognosis, complexity, or stigma may encourage the substitution of euphemisms by either health professionals or patients. Finally, national differences in the medical specialization of the usual source of medical care and the degree of specialization required for a specific diagnosis also may produce national differences in the comparative rates of disease prevalence in health surveys. These potential sources of cross-national differences in disease prevalence rates in health surveys should be considered in the future design and analysis of such research.

FEATURE PAPER: Methodological Issues in Quantitative Research with Minority Ethnic Populations

Melanie Doyle and Margaret Blake, National Centre for Social Research, London

This paper will examine methodological issues in quantitative research with minority ethnic populations, drawing on the experience of the Health Survey for England (HSE) in 1999 and 2004 (Erens, Primatesta, & Prior, 2000a, 2000b). The HSE is a series of annual surveys commissioned by the Department of Health. It was first carried out in 1991 and since 1994, has been carried out by the National Centre for Social Research and the Department of Epidemiology and Public Health at University College London.[1] The survey is designed to provide annual data for nationally representative samples to estimate the prevalence in England of specified health conditions and associated risk factors, to monitor trends in the nation's health and progress towards selected health targets, and to examine differences between subgroups in the population, such as children, young people, and minority ethnic groups. Following a brief overview of the key results from HSE 1999, the paper will focus on methodological challenges in 2004, particularly sample design and questionnaire adaptation.

THE HEALTH OF MINORITY ETHNIC GROUPS

In 1999 and 2004, the HSE focused on key ethnic differences in health, particularly cardiovascular disease in adults. Previous research has highlighted ethnic differences in the prevalence of cardiovascular disease and associated risk factors. While research with European populations has shown the impact of diet, elevated serum cholesterol, and elevated blood pressure, research among South Asian (Indian, Pakistani, and Bangladeshi) populations has shown increased levels of glucose intolerance, central obesity, and different changes to blood biochemistry (Anand & Yusuf, 2001; McKeigue, Shah, & Marmot, 1991). More recently, research within different ethnic groups suggests that environment has a significant impact on risk factors, particularly movement from rural to urban environments and migration (Anand & Yusuf, 2001). The 1999 HSE compared the pattern of health and health-related behaviours among minority ethnic groups to those reported by the general population. Adults of South Asian origin reported higher prevalence of cardiovascular disease, particularly diabetes, and showed a risk factor profile similar to that reported in earlier studies.

SAMPLING THE MINORITY ETHNIC POPULATION

According to the 2001 U.K. Census, non-White minority ethnic groups make up 7.9% of the U.K. population, with half of this group being Asian or Asian British and a quarter being Black or Black British. A further 1% of the U.K. population was classed as Irish. The minority ethnic population in England is greater (9% of population) than in other U.K. countries (less than 2% in Scotland, Wales, and Northern Ireland). As a significant proportion of the population are of minority ethnic origin, it is important to obtain better information about the health of minority ethnic groups living in England. The 2004 HSE focuses on the larger minority ethnic groups, selecting respondents of Indian, Pakistani, Bangladeshi, Black Caribbean, Black African, Chinese, and Irish origin. It is designed to yield a representative sample of approximately 1,000 adults and 500 children in each ethnic group. Those of Black African origin are boosted for the first time in 2004, following the results of a feasibility study

[1] For information, see the following Web sites: the Department of Health (www.doh.gov.uk), the National Centre for Social Research (www.natcen.ac.uk), University College London (www.ucl.ac.uk/epidemiology).

designed to examine the demography and geography, diversity, and health experiences of Black Africans in England (Elam & Chinouya, 2000; Elam, McMunn, Nazroo, Apwonyoke, Brookes et al., 2001; Elam, McMunn, Nazroo, Apwonyoke, Dektor et al., 2001; McMunn, Brooks, & Nazroo, 2001).

HSE respondents define their own ethnicity within a fixed ethnic classification adopted from the U.K. Census.[2] (HSE does not use the Census definition for Irish respondents but defines a person as Irish if he or she, or either parent, were born in Ireland.) Correspondence between Census and survey categories makes it easier to use Census data in sample design and to compare HSE data with that of other sources, such as large-scale surveys and hospital data (Majeed, Cook, Poloniecki, & Martin, 1995). The U.K. Census ethnic classification is based on ethnic and national group data but does not take into account the full range of biological, cultural, and social factors that define ethnicity and may influence health and behaviours (Erens et al., 2000a; Macbeth, 2001). A limited number of fixed categories are used, and there is considerable variation within groups. This is particularly marked for smaller, more diverse populations, such as Black Africans (Elam et al., 2001). Variation within ethnic groups may increase over time, due to generational differences or interethnic family formation, for example. The 2001 Census indicated that 1.2% of the population of England is of mixed ethnic origin (National Statistics, 2003). Those of mixed ethnic origin will be included in HSE 2004, provided that their origin is partly one of the selected ethnic groups.

To generate a cost-effective sample, the geographical distribution of the minority ethnic population must be taken into account. In sampling the minority ethnic population, it is necessary to select a large number of addresses and screen to identify eligible households. The efficiency of the sample design can be improved by limiting the sampling frame to areas with a high proportion of minority ethnic residents and by oversampling within these areas. In HSE 2004, both strategies were adopted within limits. Wards, each of which has around 2,300 addresses, were selected as the primary sampling unit. These were sorted into strata based on the proportion of residents in selected ethnic groups. Strata in which less than 2% of residents were of South Asian, Black Caribbean, or Black African origin were excluded from the boost sample, as were those with less than 0.8% of Irish residents. The sampling frame ensures over 90% coverage for Black African, Black Caribbean, Indian, Bangladeshi, and Pakistani groups. Areas with a particularly high proportion of minority ethnic residents were oversampled so more wards were selected in these strata, and within each stratum, wards were selected with a probability proportional to size (the number of addresses). The minority ethnic population is concentrated in urban areas, with 45% of the non-White population living in London. This strategy has amplified the geographical distribution of this population, with an increased proportion of the sample in London and other urban areas.

Some ethnic groups, such as the Irish and Chinese, are more dispersed so it is not possible to identify areas with a high proportion of residents. Coverage is lower for these than for other groups, and for Chinese, an additional boost sample was needed as the population is relatively small and dispersed. HSE 2004 adopted a sampling method devised by the Office of National Statistics specifically for sampling Chinese residents and successfully used in a number of surveys (e.g., Sproston, Pitson, Whitfield, & Walker, 1999). Census data was used to stratify wards based on the proportion of Chinese residents and additional wards selected from those with 15 or more adults and children of Chinese origin. Within selected wards, a sampling frame was devised using information from the Electoral Register to identify households where a resident had one of the 1,300 most common Chinese surnames. This method is a cost-effective means of targeting Chinese

[2] See National Statistics Web site for details of questions on National Identity and Ethnic group: www.statistics.gov.uk/Harmony/Primary/national.asp

respondents in selected areas but has the disadvantage of excluding certain people, such as Chinese residents with a non-Chinese surname and those not registered to vote.

SCREENING TECHNIQUES

Addresses were sampled using the small user Postcode Address File (PAF), a file containing all addresses in England. This provides address details but no information about the householders, so the interviewer must make contact at the household to establish eligibility. The majority of addresses are assumed to be residential because of the low levels of mail delivered, and in practice, only a small proportion of addresses are lost because they are ineligible (e.g., business addresses, institutions, empty properties). In boost samples, selection criteria, such as age or ethnic group, further reduce the number of issued addresses that are eligible. A large number of addresses must be screened to obtain a reasonable sample of respondents from minority ethnic groups. In HSE, usual practice is to assign an interviewer to a single point with around 20 addresses. In 2004, the number of addresses per point in the boost sample ranges from 35 to 115. This variation is primarily driven by sample requirements but also influenced by workload at the interview stage and field work deadlines. At the interview stage, each point is covered by a single interviewer, but at the screening stage, addresses can be assigned to one interviewer or to a team of interviewers.

Screening is used to establish both eligibility and translation requirements. In areas with a high proportion of residents in the selected ethnic groups, screening is carried out at the sampled address. Up to three households can be included at each address and, within each household, up to four adults and three children can be interviewed. However, the sampling frame used in HSE 2004 includes some areas with a low proportion of minority ethnic residents where a reasonable sample size could not be obtained by screening at sampled households alone. Rather than increase the number of addresses issued in these points, HSE employs a screening technique specifically designed for sampling minority ethnic residents in areas with few eligible residents. This technique, know as "focussed enumeration," was jointly devised by the National Centre and Policy Studies Institute. In essence, focussed enumeration seeks to increase the number of addresses screened by allowing the interviewer to screen up to five addresses at each sampled address. At each sampled address, the interviewer establishes whether anyone at that address is eligible to take part and then asks about the neighbouring addresses, specifically whether anyone in the two previous or two next addresses are of Asian, Black Caribbean, Black African, or Chinese origin. Full screening is carried out at the adjacent address if the interviewer is told that someone may be eligible to take part. Focussed enumeration cannot be used to sample Irish respondents as the definition of Irish used in HSE incorporates parental place of birth.

CROSS-CULTURAL ISSUES IN QUESTIONNAIRE DESIGN

The Health Survey consists of an interview and nurse visit. Information about health and health-related behaviours is obtained from questionnaires, physical measurements, and analysis of biological samples. Computer-assisted personal interviewing (CAPI) is used to control the interview and nurse visit. The interviewer completes household and individual interviews in CAPI format, administers self-completions, and collects height and weight measurements. Core question modules, such as general health, smoking, and alcohol consumption, are included each year along with additional modules on selected topics. In 2004, these include physical activity, cardiovascular disease, and child breathing. The nurse conducts a CAPI interview, carries out a range of measurements, and collects biological samples, such as blood, saliva, and urine. In 2004, physical measurements include blood pressure, waist and hip circumference, and lung function. Analytes include haemoglobin,

ferritin, total and HDL cholesterol in blood, and sodium levels in urine.

In studies focusing on minority ethnic groups, it is particularly important to consider cultural and language differences that may influence response to the survey as a whole and the quality of response to specific questions. According to a recent report, around a quarter of Chinese and South Asian immigrants have "no functional English skill" (Hunt & Bhopal, 2003). The use of translations in the Health Survey bears out this observation as, in 1999, interviews were carried out wholly in translation for over a fifth of Chinese and Pakistani adults, over half of Bangladeshi men, and around two-thirds of Bangladeshi women (Erens et al., 2000b). Translation requirements may be higher among certain subgroups, such as recent or older immigrants, who have very different patterns of health. Translations are important but costly, and it may be difficult to cover all languages spoken by potential respondents. In practice, the need for a particular language and the cost of translated materials and interpreters are taken into account when selecting languages for translation. In 2004, the Health Survey will be translated into Punjabi, Gujarati, Hindi, Urdu, Bengali, Mandarin, and Cantonese. Translations will not be provided for Black African respondents, as the experience of other surveys (e.g., The National Survey of Sexual Attitude and Lifestyles) indicates that translations would not be required by the majority of Black Africans living in England. The diversity of the group (Elam & Chinouya, 2000; McMunn et al., 2001) makes it difficult to select a language for translation, and where a translation need is identified (e.g., Black Africans from Somalia), the numbers included in a random sample will be so small that the translations are uneconomical.

Translations were carried out and checked by an external agency using a forward translation method. Translated and source questions then were compared by bilingual HSE interviewers to ensure that the correct meaning was conveyed and that the appropriate language (e.g., in terms of formality) was used. Based on the interviewers' recommendations, a "best-fit" was negotiated with the translators. Additional guidelines suggested for checking the adequacy of translations, such as translating between two translated languages (Hunt & Bhopal, 2003), were not feasible in this context. Screening cards were produced in seven languages to allow interviewers to obtain information about household eligibility and translation needs from respondents who do not speak English. These cards are designed for use by English-speaking interviewers with householders who are literate in their spoken language. Although interpreters are not employed in screening, they may have a positive effect on recruitment through their role in introducing the survey at interview (Oakley, Wiggins, Turner, Rajan, & Barker, 2003). Standardised translations of the CAPI questionnaires, self-completions, and information documents are provided in paper format for interviews and nurse visits. Although it is more cost effective to use interviewers who speak the translated languages, in practice it is difficult to recruit enough bilingual interviewers to cover field work, so interpreters often are used. The interviewer or nurse controls the pace of the interview, ensures that the correct routing is followed, and records answers in CAPI: Each question is read in English from the CAPI questionnaire and in the translated language by the interpreter who then feeds back the response to the interviewer or nurse. The use of interpreters increases interview length and fieldwork costs and may have a negative impact on response. Using interpreters also may lead to field work delays, as the interviewer, interpreter, and respondent must arrange a time that suits all involved.

Language differences between ethnic groups may be partly addressed by providing translations; however, there may be conceptual and cultural differences as well. Although changes to question wording may enhance the common meaning of questions in different ethnic groups, in established questionnaires, such as HSE, this may be precluded by the need to examine time series

data. Furthermore, tailoring questions to each group may make it difficult to compare data with that of other ethnic groups or with that of the general population. It may be possible to tailor questions so that they have the same meaning in different cultural contexts, but the questions may not "fit" all respondents within a group if a relatively broad ethnic classification is used. Similarly, differences within the White population, such as age and education, may affect the understanding and conceptual equivalence of survey questions. If questions were tailored for different languages and ethnic groups, questions would need to be revised on a regular basis due to ongoing change in the population as a whole and within different ethnic groups.

Although the Health Survey does not alter question wording, additional questions may be included in certain years to adapt the questionnaire for different ethnic groups (e.g., questions about forms of tobacco use, such as paan, bidi, and hookah). Such questions may be particularly effective in highlighting ethnic differences and in uncovering potentially hidden data. For example, in the 1999 Health Survey, only 1% of Bengali women reported smoking cigarettes but 26% reported alternative forms of tobacco use. Ethnic differences also may be considered in the design and piloting of new modules, such as the module on Complementary and Alternative Medicine in HSE 2004.

DATA QUALITY & ANALYSIS

Achieved sample size, the inclusion of respondents of mixed origin, response variation across ethnic groups, and factors affecting data quality must be taken into account in analysis. Data from HSE 2004 will be used to examine health within ethnic groups and to make comparisons with the general population. Respondents of mixed ethnic origin are included and are classified into existing ethnic groups based on maternal ethnic origin. Such classification is more meaningful than a "mixed" ethnic category including all respondents of mixed ethnicity, as respondents of a particular origin and respondents partly of that origin may share characteristics that have some impact on health. It also ensures that sample size is maintained in ethnic groups, particularly those with a high proportion of people of "mixed" ethnic origin, such as the Black Caribbean group, but may increase diversity within the group.

Ethnic variations in response may occur for a number of reasons and at a number of levels. Individuals may not take part in the survey for a number of reasons, but availability of translations and interview length may influence nonresponse in some ethnic groups. In 1999, response rates were lowest among Chinese and Bangladeshi respondents for whom translation needs were highest. Item nonresponse, or the quality of response on certain items, may be influenced by cultural and religious differences. Certain questions may be regarded as sensitive—e.g., Muslim respondents may regard questions about alcohol consumption as sensitive—and respondents may feel inhibited from answering certain questions if interviewed in the presence of other family members. The way in which the interview is carried out can be adapted to take such factors into account. For those age 18–24, interviewers can administer questions about smoking and drinking in self-completion format and can interview adults individually rather than conduct concurrent interviews with several adults. Data quality is a further consideration, particularly where information may not fit pre-existing data structures, e.g., naming conventions for Muslim men or where accurate data is not available, such as birth data.

HSE data is published annually in a main report and trend tables and is available to academic and other researchers.[3] Secondary analysis and follow-up studies may help to clarify some of the methodological issues highlighted in this paper (e.g., Bhopal et al., 2004) and add further value to the data

[3] HSE data is archived annually at the UK Data Archive (www.data-archive.ac.uk/) and support for users provided by the Economic and Social Data Service (www.esds.ac.uk/).

through additional research on related health issues (e.g., Sproston & Nazroo, 2002).

REFERENCES

Anand, S., & Yusuf, S. (2001). Ethnic variations in cardiovascular disease. In H. MacBeth & P. Shetty (Eds.), *Health and ethnicity.* London, New York: Taylor and Francis.

Bhopal, R., Vettini, A., Hunt, S., Wiebe, S., Hanna, L., & Amos, A. (2004). Review of prevalence data in, and evaluation of methods for cross-cultural adaptation of, UK surveys on tobacco and alcohol in ethnic minority groups. *British Medical Journal, 328,* 76–80.

Elam, G., & Chinouya, M. (2000). *Feasibility study for health surveys among Black African populations living in the UK: Stage 2 – Diversity among Black African communities.* Joint Health Surveys Unit (JHSU).

Elam, G., McMunn, A., Nazroo, J., Apwonyoke, M., Brookes, M., Chinouya, M., Decktor, G., Ibrahim, S., & Lutaaya, G. (2001). *Feasibility study for health surveys among Black Africans living in England: Final report – Implications for the Health Survey for England 2003.* JHSU. (Note: ethnic boost was delayed until 2004 so that data from the 2001 Census could be used.)

Elam, G., McMunn, A., Nazroo, J., Apwonyoke, M., Decktor, G., Ibrahim, S., & Lutaaya, G. (2001). *Feasibility study for health surveys among Black Africans living in England: Stage 3 – The health experiences of people of African origin living in London.* JHSU.

Erens, B., Primatesta P., & Prior, G. (2000a). *Health Survey for England 1999. The health of minority ethnic groups, volume 1: Findings.* London: The Stationery Office.

Erens, B., Primatesta, P., & Prior G. (2000b). *Health Survey for England 1999. The health of minority ethnic groups, volume 2: Methodology and documentation.* London: The Stationery Office.

Hunt, S., & Bhopal, R. (2003). Self reports in research with non-English speakers. *British Medical Journal, 327,* 352–353.

MacBeth, H. (2001). Defining the ethnic group: Important and impossible. In H. MacBeth & P. Shetty (Eds.), *Health and ethnicity.* London, New York: Taylor and Francis.

Majeed, F. A., Cook, D. G., Poloniecki, J., & Martin, D. (1995). Using data from the 1991 Census. *British Medical Journal*, *310*, 1511–1514.

McKeigue, P. M., Shah, B., & Marmot, M. G. (1991). Relation of central obesity and insulin resistance with high diabetes prevalence and cardiovascular risk in South Asians. *Lancet, 337*, 382–386.

McMunn, A., Brookes, M., & Nazroo, J. (2001). *Feasibility study for health surveys in Black African populations living in Britain: Stage 1 – The demography and geography of Black Africans in Britain.* JHSU.

National Statistics. (2003). *Census 2001–Ethnicity and religion in England and Wales.* Retrieved May 29, 2004, from http://www.statistics.gov.uk/pdfdir/ethnicity0203.pdf

Oakley, A., Wiggins, M., Turner, H., Rajan, L., & Barker, M. (2003). Including culturally diverse samples in health research: A case study of an urban trial of social support. *Ethnicity & Health, 8*(1), 29–39.

Sproston, K., & Nazroo, J. (2002). *Ethnic minority psychiatric illness rates in the community.* London: The Stationery Office.

Sproston, K., Pitson, L., Whitfield, G., & Walker, E. (1999). *Health and lifestyles of the Chinese population in England.* London: Health Education Authority.

FEATURE PAPER: Enhancing Data Collection from "Other Language" Households

Mary Cay Murray, Mike Battaglia, and Jessica Cardoni, Abt Associates, Inc.

INTRODUCTION

A recent report from the U.S. Census Bureau notes that the foreign-born population increased by 57% between 1990 and 2000, from 19.8 million to 31.1 million. In 2000, over 16 million (52%) of the foreign-born were from Latin America, 8.2 million (26%) were from Asia, and 4.9 million (16%) were from Europe. More than half of this population lived in three states: California, New York, and Texas. The foreign born account for more than a quarter of the population of California (26%), and they exceed the national average of 11% in New York (20%), New Jersey and Hawaii (18%), Florida (17%), Nevada (16%), Texas (14%), the District of Columbia and Arizona (13%), and Illinois and Massachusetts (12%). The four largest cities in the U.S. have the largest foreign-born populations: 2.9 million in New York, 1.5 million in Los Angeles, 0.6 million in Chicago, and 0.5 million in Houston. Other large cities with substantial numbers of foreign-born residents include Philadelphia, Phoenix, San Diego, Dallas, San Antonio, San Jose, San Francisco, and Miami (Malone, Baluja, Costanzo, & Davis, 2003).

This increase in the foreign born and its concentration in certain states and large cities has implications for large telephone surveys conducted in the U.S. At the very least, the screener and questionnaire must be translated into Spanish, and the interviewing staff must be able to work with the large Spanish-speaking population. But to reach the broader constellation of non-English speakers, other accommodations must be made.

BACKGROUND

Since 1995, Abt Associates has been using Language Line Services (formerly part of AT&T) to screen households and conduct interviews with families that would otherwise not be able to participate in the National Immunization Survey (NIS). This service has become an integral part of the household screening and interviewing process and is used for all interviews in languages other than English and Spanish. The NIS screener and questionnaire are translated into Spanish, and Spanish-speaking interviewers are available to work all shifts of data collection. However, we have not translated the NIS into other languages, relying instead on the Language Line Service to include "other language" households in the survey. Most households (over 96%) screen out of the NIS with just one or two questions, so there is very little to translate. The NIS questionnaire requests relatively straightforward factual information, such as dates of immunizations, basic demographics, and permission to contact providers. These questions are not difficult to translate. (The more complex questionnaire for the Children with Special Health Care Needs component of the State and Local Area Integrated Telephone Survey [SLAITS], which was coordinated with the NIS, was translated into ten languages and used fluent interviewers, often native speakers, to conduct telephone interviews. The SLAITS components use the NIS screening sample to identify eligible households and conduct interviews after the NIS screening and interviewing are completed [Brady, Osborn, Blumberg, & Olson, 2003].)

The NIS is conducted by Abt Associates for the National Immunization Program and the National Center for Health Statistics of the Centers for Disease Control and Prevention. The NIS's target population is children age 19 to 35 months living in households in the U.S. at the time of the interview. The NIS uses a random-digit-dial (RDD) telephone survey to identify households containing children in the target age range and interview an adult most knowledgeable about the child's vaccinations. With the consent of the child's parent or guardian, the NIS also contacts the child's

health care providers by mail to request information on vaccinations from the child's medical records.

Samples of telephone numbers are drawn independently for each calendar quarter within 78 Immunization Action Plan (IAP) areas. Of the 78 IAP areas, 28 (including the District of Columbia) are urban areas. The remaining 50 are either an entire state or a "rest of state" IAP area (where the state contains one or more urban IAP areas). This design makes it possible to produce annualized estimates of vaccination coverage levels within each of the 78 IAP areas with a specified degree of precision. Further, by using the same data collection methodology and survey instruments in all 78 areas, the NIS produces vaccination coverage levels that are comparable among IAP areas and over time.

Over the course of a year, approximately 3.4 million telephone numbers yield household interviews for approximately 32,000 children. In 2002, the sample contained 21,410 children with adequate provider data. Estimates of vaccination coverage are based mainly on the data from children's immunization providers.

METHOD

This paper examines the contribution of the Language Line Service (LLS) to the NIS in calendar years 1996, 1998, 2000, and 2002. We examine the languages most frequently used, the IAP areas with the most LLS interviews, and the impact of the LLS on vaccination coverage estimates for the U.S. as a whole and for IAP areas. Next, we look at demographic characteristics of LLS cases in 2002. Finally, we examine the effect of the LLS on interviewing the Asian subpopulation, a growing minority group in the U.S.

RESULTS

As might be expected, some languages are used more frequently than others, and some IAP areas benefit disproportionately from the service. However, every IAP area had at least one LLS interview in the calendar years examined. The number of distinct languages increased by year: 27 in 1996, 34 in 1998, 38 in 2000, and 41 in 2002. In the tables below, we show which languages (other than English and Spanish) were most commonly used and how the IAP areas were affected by this service. The most common languages, with at least 10 interviews in one of the four years, are shown in Table 1. Although as a group Asian languages predominate (Vietnamese in particular), there also are substantial numbers of Portuguese and Arabic speakers.

Table 2 shows the percentage of completed cases interviewed using the LLS for those IAP areas that had at least ten household interviews completed with the LLS in one or more of the four years. Generally, these IAP areas are urban rather than Rest-of-State (ROS) areas, with the exception of Massachusetts ROS, New Jersey ROS, and Hawaii. (Most of the population in Massachusetts ROS, New Jersey ROS, and Hawaii is located in Metropolitan Statistical Areas.) In 2002, Language Line usage increased to over 4% of completed interviews in Boston, Newark, New York City, and King County (Seattle), Washington. Detroit showed a marked increase in LLS cases over the four years, with 3% in 2002. Santa Clara County (San Jose), California, also benefited considerably in 2002, obtaining 3.3% of its completed interviews through the LLS. On average, an IAP area had around 435

Table 1. Percent Distribution of Most Common Foreign Languages (Other than Spanish) in the NIS Interviews, by Year

Language	1996	1998	2000	2002
Arabic	5.4	8.5	5.9	13.4
Cantonese	4.9	8.1	6.8	5.0
Haitian Creole	4.3	2.0	6.3	3.7
Japanese	4.3.	4.0	3.8	4.4
Korean	10.3	7.7	5.9	7.7
Mandarin	4.9	4.0	5.1	7.4
Portuguese	6.5	6.1	9.7	11.4
Russian	6.5	5.7	8.0	4.4
Somali	0.0	3.2	3.0	3.7
Vietnamese	27.6	24.6	12.7	16.1
Other languages	25.3	26.1	32.8	22.8
Total Number of LLS Cases	**185**	**248**	**237**	**298**

Table 2. Percentage of Completed Cases that Used the Language Line Service Among IAP Areas with at Least 10 Cases Completed Through the LLS, by Year

IAP Area	1996	1998	2000	2002
MA ROS	0.5	2.0	2.7	2.5
MA Boston	3.1	2.6	2.9	4.8
NJ ROS	1.5	2.3	1.4	3.2
NJ Newark	0.2	2.9	3.1	4.7
NY New York City	3.2	3.0	3.6	4.2
PA Philadelphia	0.7	1.2	1.3	2.2
IL Chicago	2.2	2.5	2.1	0.8
MI Detroit	0.5	1.4	1.7	3.0
CA Los Angeles Co.	2.2	1.3	0.9	0.8
CA Santa Clara Co.	1.3	3.2	2.0	3.3
HI	2.0	2.0	2.0	1.6
WA King Co.	1.6	2.4	2.0	4.1

completed interviews in a calendar year between 1996 and 2002.

Since the NIS estimates rely heavily on provider data, we looked at the proportions of LLS cases with provider data. As Table 3 shows, in calendar year 1996, 63.6% of the LLS cases had provider data, comparing favorably with those cases not using the LLS, which had 63.4% with provider data. In calendar year 1998, the LLS interviews were as likely as the rest of the sample to have provider data, with both groups at 67.1%. In calendar year 2000, a higher proportion of the LLS interviews (76.9%) had provider data, compared to 67.3% of cases not using the LLS. The LLS cases also did better on provider data in 2002, with 76.2% having provider data, as opposed to only 67.2% of the rest of the sample.

Next we examined the impact of the use of the Language Line Service on the estimates of up-to-date status for the 4:3:1:3:3[1] vaccination series, the most comprehensive measure of vaccination status. We did this by calculating weighted estimates of vaccination coverage for each IAP area with and without the LLS-completed interviews among children with adequate provider data. Overall, the impact is minimal at the national level (no table provided) but more noticeable at the IAP-area level. For 1996, there was no national impact. With or without the LLS cases, the 4:3:1:3:3 coverage was the same at 67.7%. For 1998, including or excluding the LLS cases also had no impact on the national estimate (72.6%). For both 2000 and 2002, including the LLS cases in the estimates decreased the national estimate by 0.1%. Nevertheless, the impact of using the service varied at some IAP areas in all four years, as shown in Tables 4 and 5.

Table 4 shows the number of IAP areas that changed by year and the direction of the yearly changes. On balance, the impact of the LLS was neutral to slightly positive.

Across the four years, the IAP areas most affected in one direction or the other by the use of the Language Line Service are not the same, nor are they necessarily the IAP areas with the most LLS cases. Table 5 shows the IAP areas that had an absolute change in value of at least 0.5 percentage point (ppt) in one or more of the four years. No IAP area's differences crossed the 0.5 ppt threshold in all four years, and only two IAP areas (MA Rest of State and Hawaii) did so in three years. In the later years, more IAP areas have large absolute changes. Among the 32 IAP areas listed in this table, in 1996, there were 9 that had no difference and 6 that had an increase

Table 3. Percentage of Cases with Provider Data: LLS Cases and Those Not Using the LLS

Year	LLS Cases	Non-LLS Cases
1996	63.6	63.4
1998	67.1	67.1
2000	76.9	67.3
2002	76.2	67.2

Table 4. Impact of LLS on IAP Areas, by Year

Year	Number Lower	Number the Same	Number Higher
1996	19	36	23
1998	25	26	27
2000	28	23	27
2002	20	30	28

[1] 4:3:1:3:3: 4+ DTP (Diphtheria and Tetanus toxoids, and Pertussis vaccine), 3+ Polio, 1+ MCV (Measles-Containing Vaccine), 3+Hib (*Haemophilus influenzae* Type B), and 3+ Hepatitis B.

or decrease of at least 0.5 ppt. The largest change was –0.6 ppt in Kansas; San Diego County, California; and Hawaii. In 1998, there were 11 IAP areas with absolute differences of at least 0.5 ppt with changes greater than -1 ppt in New Jersey ROS (-1.5 ppt); Davidson County (Nashville), Tennessee (-1.4 ppt); and Oklahoma (-1.2 ppt). In 2000, there were 14 IAP areas whose absolute change was at least 0.5 ppt, but only one was greater than 1 ppt: Santa Clara County (San Jose), California (-1.7 ppt). In 2002, there were again 14 IAP areas whose absolute change was at least 0.5 ppt and 4 with absolute values greater than 1 ppt. These included Maine (-1.2 ppt); Newark, New Jersey (+1.6 ppt); Miami-Dade County, Florida (-1.5 ppt); and North Carolina (-1.2 ppt).

Table 5. IAP Areas in Which Use of LLS Changes the Estimate of 4:3:1:3:3 Coverage by at Least 0.5 Percentage Point (+ or -) in One or More of the Four Years: 1996, 1998, 2000 and 2002

	(Difference percentage points)			
IAP Area	**1996**	**1998**	**2000**	**2002**
MA Rest of State	-0.2	**-0.7**	**-0.7**	**0.6**
MA City of Boston	0.3	**-0.8**	0.1	**-0.5**
Maine	0.1	0.1	0.2	**-1.2**
New Hampshire	0.0	**-0.7**	0.1	0.1
NJ Rest of State	-0.2	**-1.5**	**0.9**	-0.4
NJ City of Newark	0.1	0.0	**-0.8**	**1.6**
NY Rest of State	0.1	0.2	**-0.5**	**-0.5**
NY 5 Counties	0.2	-0.2	**-0.6**	0.1
Dist of Columbia	0.2	-0.2	0.0	**-0.6**
PA Philadelphia	0.4	0.4	**-0.7**	0.0
FL Rest of State	0.1	-0.3	0.4	**-0.7**
FL Miami/Dade Co.	0.1	0.1	**-0.5**	**-1.5**
Kentucky	0.0	**-0.5**	-0.3	0.0
North Carolina	-0.3	0.0	0.1	**-1.2**
TN Rest of State	0.0	**-0.5**	0.0	0.0
TN Davidson Co.	0.0	**-1.4**	0.1	0.2
IL Chicago	0.1	0.2	-0.2	**0.7**
IN Marion County	0.0	**-0.9**	0.2	0.1
MI Rest of State	0.0	0.0	-0.4	**-0.8**
Minnesota	0.0	-0.2	0.2	**-0.8**
Oklahoma	0.0	**-1.2**	0.2	0.2
TX Houston	0.3	-0.1	**-0.7**	0.3
Kansas	**-0.6**	0.2	0.2	**-0.5**
Missouri	**-0.5**	-0.2	0.0	0.2
Colorado	0.0	**-0.5**	0.2	0.2
CA Los Angeles	**0.5**	0.1	**-0.7**	0.3
CA Santa Clara	-0.4	0.4	**-1.7**	-0.3
CA San Diego Co.	**-0.6**	0.0	**0.5**	0.0
Hawaii	**-0.6**	0.1	**0.5**	**-0.7**
Oregon	**-0.5**	0.4	-0.4	0.0
WA Rest of State	-0.1	**-0.5**	**-0.7**	0.2
WA King County	-0.4	0.1	**0.9**	**0.6**

Next we reviewed the demographic characteristics of the users of the Language Line Service for the most recent year, 2002, and compared them to households not using the LLS. (These comparisons are based on raw rather than weighted data.) In that year, the respondents using the LLS were approximately 1% of the completed interviews (298 out of 31,693).

As shown in Table 6, the LLS cases differed from the non-LLS cases on a number of demographic dimensions:

- LLS children's mothers tended to have less education, with close to 66% having only 12 years of schooling, while in the non-LLS group, slightly over 41% had only 12 years of schooling.
- LLS children were poorer, with over 30% below the poverty level, compared to almost 18% in the non-LLS group. LLS children also were more likely to come from households that did not report income, with 25.5% unknown, while income was unknown for only 12.5% of the non-LLS group.
- Almost 15% of the LLS children were foreign born, compared to 1.2% of the non-LLS group.
- Members of the LLS group were less likely to be firstborn, with 33.6% the first born in their families, while 38.6% were first born in the non-LLS group.
- About 89% of LLS mothers were currently married, compared to 72.2% of the non-LLS group.
- LLS mothers were older, with no mothers under age 20, and 66.8% were 30 or older. Among the non-LLS group, 2.8% were under age 20 and 54.9% were 30 and older.
- More than half of the LLS group reported the child's ethnicity as Asian. (This also is evident in the languages reported in Table 1.) In the non-LLS group, only 4.3% of children were reported as having Asian ethnicity.

Over 10% of the households with Asian children in the 2002 NIS sample were interviewed using the LLS. Altogether, in 2002, there were 1,489 Asian children, or 4.7% of the interviews.

Because such a large proportion of the LLS children are of Asian ethnicity, we did a comparison, using weighted data, of their up-to-date status to that of Asian children not using the LLS. Though the LLS children are less likely than the non-LLS children to be 4:3:1:3:3 up-to-date (70% vs. 78%, respectively), the difference was not statistically significant.

CONCLUSION

In summary, the Language Line Service makes a valuable contribution to the NIS. It expands the survey to a potentially underserved segment of the population—those who are linguistically isolated—and helps reduce potential bias in estimates in areas of the country where the NIS encounters such respondents. This service can yield almost 5% of the household interviews in

Table 6. Demographic Characteristics of LLS Cases Compared to Non-LSS Cases

Characteristic	% of LLS Cases	% of Non-LSS Cases
Education: only 12 years of schooling	65.8	41.4
Below poverty level	30.5	17.9
Household did not report income	25.5	12.5
NIS child was foreign-born	14.8	1.2
NIS child was firstborn in family	33.6	38.6
Mother currently married	88.6	72.2
Mother under age of 20	0.0	2.8
Mother age 30 or older	66.8	54.9
Child's ethnicity is Asian	51.0	4.3

some IAP areas in a given calendar year and can have an impact on vaccination estimates in some IAP areas, even though its effect on the national estimate is minimal. Also, as we can see from Tables 4 and 5 above, LLS children may have different vaccination coverage rates than children in English- and Spanish-speaking households. Finally, use of this service strengthens the sample of Asians included in the NIS, allowing for a more reliable examination of this growing segment of the U.S. population.

REFERENCES

Brady, S., Osborn, L., Blumberg, S. J., & Olson, L. (2003, May). *Collecting data in multiple languages: Development of a methodology.* Paper presented at the annual meeting of the American Association for Public Opinion Research Conference, Nashville.

Malone, N., Baluja, K. F., Costanzo, J. M., & Davis, C. J. (2003, December). *The foreign-born population: 2000* (Census 2000 Brief C2KBR-34). Washington, DC: U.S. Census Bureau.

SESSION 3 DISCUSSION PAPER: Advancing Measurement Equivalence of Health Outcome Measures

Colleen A. McHorney, Regenstrief Institute, Inc.

INTRODUCTION

Great progress has been made in measuring health status and quality of life (QOL) outcomes in the past 50 years. Close to two dozen generic QOL instruments have been developed (McHorney, 1997), and literally hundreds of disease-specific QOL exist (Bowling, 2001). There are over 85 tools that measure basic and instrumental activities of daily living (McHorney, 2002), and close to as many depression measures exist as well (Task Force for the Handbook of Psychiatric Measures, 2000). Such measures traditionally have been developed with a keen eye toward documenting the psychometric properties of reliability and validity. However, one limitation of our armamentarium of measures is that they rarely have been fully psychometrically vetted in diverse population groups (Stewart & Nápoles-Springer, 2000). That is, most measures have been developed and validated in mainstream population groups. Thus, questions remain largely unanswered about the equivalence of older and newer QOL and health status measures in diverse population groups.

This state of affairs, which I call psychometric ethnocentrism, is regrettable because the U.S. is becoming a more diverse population. Our population is aging, and the proportion of the population that is elderly will double by 2030 when 20% of the population will be age 65 or older (Day, 1996). The U.S. is also becoming more ethnically diverse. In 1990, 76% of the population was White, non-Hispanic; this percentage will decrease to 64% in 2020 and to 53% in 2050. The greatest increase in ethnic pluralism will be for persons of Hispanic origin and Asian origin (Day, 1996). Because of our growing cultural pluralism, there is more so than ever before a need for evidence that health status and QOL tools exhibit measurement equivalence across diverse population groups (Stewart & Nápoles-Springer, 2000).

One of the national health goals in Healthy People 2010 is to eliminate health disparities (U. S. Department of Health and Human Services [DHHS], n.d.). Currently, we do not know whether group differences in self-reported function and well being reflect "true" pathology, artifactual differences due to measurement bias, or a combination of the two. Our bolus of health outcome measures needs to be reliable, valid, and relevant across diverse groups. Further, although many health status and QOL measures have been developed in the U.S., they are increasingly used in cross-national and multinational clinical trials (Berzon, Hays, & Shumaker, 1993). Such applications cry out for rigorous and thorough documentation of cross-cultural measurement equivalence before measures are fielded and data are interpreted. The Harkness and Willis papers in this volume spoke elegantly about theory and methods for establishing cross-cultural measurement equivalence.

EMERGENT PARADIGM SHIFT IN MEASUREMENT PLATFORMS

We are on the cusp of a paradigm shift in our measurement and testing platforms. Our history in health outcomes assessment has been characterized by a group-testing paradigm (McHorney, 1997). The defining feature of group testing is the use of a fixed set of items regardless of the appropriateness

Supported in part by R01 AG022067, Department of Veterans Affairs RR&D C-2488-R and Department of Veterans Affairs RCS 02-066-1 to Dr. McHorney.

Address requests for reprints to Colleen McHorney, Regenstrief Institute, Inc., RHC 6th Floor, 1050 Wishard Blvd, Indianapolis, IN 46202.
cmchorney@regenstrief.org

of any given item for any given respondent. Item selection for group tests tends to be geared toward the middle of the continuum in terms of item difficulty. Because an era of psychometric efficiency has dominated health assessment in the last decade (McHorney, 1997), items tend to be selected that are near alternate forms of one another, thus maximizing reliability at the expense of breadth and depth of measurement.

These measurement practices have two consequences: poorly targeted items and imprecise measurement. As to the former, *poorly targeted items*, respondents become frustrated by redundant items and items that are of low salience to them (McHorney & Bricker, 2002). Fixed-length health surveys can bore healthier respondents (because they have to respond to multiple items that are very easy for them to do) and frustrate more impaired respondents (because they have to trudge through multiple items that are very difficult for them to do). As to the latter, *imprecise measurement*, because item selection is geared toward the middle of the road in content coverage and difficulty, the end points of the health continuum tend to be poorly defined. Such score imprecision has two principal consequences. First, it is impossible to distinguish among persons at the ceiling or floor, even though they vary in the underlying construct (McHorney, 1997). Thus, ceiling and floor effects paint a more favorable image of population health than is true, produce Type II errors for group-level hypothesis testing, and yield false-negative outcomes for the individual-patient assessment. The second consequence is that it is impossible to measure decline in health over time for persons at the floor and improvement in health over time for persons at the ceiling. Thus, score distributions that are skewed at baseline will underestimate or miss the effects of treatment or natural history on health outcomes.

Like our colleagues in educational and psychological testing, health status assessment is moving away from fixed-length tests toward a new paradigm of computerized-adaptive testing (CAT) (McHorney, 1997, 2003). A recent NIH RFA on CAT (DHHS, 2003) has solicited research applications to develop a CAT system for patient-reported outcomes, including health status and QOL. CAT uses a computer to administer items to respondents. Items are housed in large unidimensional repositories called item banks. The item banks are precalibrated using item response theory to obtain theoretically sample invariant estimates of item difficulty and item discrimination. CAT is adaptive in a literal sense because each "test" is tailored to the unique ability level of each respondent. Each person taking a CAT is taking a different version of the test because items are administered on the basis of the respondent's previous answers. For example, if a respondent cannot walk one block, the computer knows not to ask if they can walk a mile. Instead, the computer asks if they can walk across the room. Item response theory is the glue that allows all of the different forms of a test to connect to each other on the same yardstick.

The advantages of CAT for large-scale assessment are numerous (Sands, Waters, & McBride, 1997):

(1) Reduced testing time by one quarter to two-thirds;
(2) Reduced human capital in survey administration and scoring;
(3) Superior capability to use graphics, audio, video, animation, and text in item presentation;
(4) Enhanced potential to assess low-literacy patients through the use of graphical and pictorial display;
(5) Sophisticated method to appropriately challenge survey respondents instead of boring or discouraging them;
(6) Real-time scoring and feedback of results;
(7) Improved precision of obtained test scores; and
(8) Capacity to add new items to the item bank and retire items that become outdated.

However, CAT places great demands on assumptions of measurement equivalence. Because fewer items are needed to assess

ability in CAT, it is crucial that each banked item be free of measurement bias or differential item functioning (DIF). Thus, identification and eradication of items that exhibit DIF is an essential cornerstone of item bank development and CAT operations.

DIFFERENTIAL ITEM FUNCTIONING

An item functions differentially if two individuals with equal ability (e.g., the same amount of the measured state or trait) do not have the same probability of item endorsement. For example, a mental health item on crying would function differentially if men and women had the same underlying level of depression yet differentially endorsed it. Self-report measures of functioning and well being can fall prey to DIF because human beings interpret such items within the context of culturally and socially determined mindsets. Because of subgroup variations in socialization, perceptions, norms, customs, and values about physical and mental health (and their expressions), it is possible that some health status and QOL items lay outside the cognitive, behavioral, and experiential realm of some subgroups. It is equally possible that the content of some health status and QOL tools do not adequately assess, in the right degree or distinction, parameters of daily living as interpreted, articulated, and experienced by some population subgroups. It also is important to note that even if items are socially, culturally, or existentially irrelevant, people will still answer them to "save face" or "satisfice" (Warnecke et al., 1996). Deutscher (1973) refers to this tendency as the "courtesy" bias.

To date, DIF has been identified in a large number of health assessment tools, including functional status (Fleishman, Spector, & Altman, 2002; McHorney, 2002), cognitive status (Teresi, Holmes, Ramirez, Gurland, & Lantigua, 2001; Teresi, Kleinman, & Ocepek, 2000), and mental health (Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Stommel et al., 1993). DIF has been identified by age, race, gender, ethnicity, socioeconomic status, language, and nationality. Including DIF-biased items in a scale can exaggerate or attenuate "true" differences, thereby measurably affecting substantive study conclusions or case rates.

THEORETICAL EXPLANATIONS FOR DIF

The social-cognitive framework of survey response (Sudman, Bradburn, & Schwarz, 1996) holds that a survey (whether it is self-administered or administered by an interviewer, telephone, or computer) is a social phenomenon that involves elaborate cognitive work by respondents. In essence, the process of completing a survey is "fundamentally a social encounter" (Sudman et al., 1996) and, as such, is governed by social rules and norms. The request to complete a survey may conflict with other norms held by respondents. One such norm, which would serve to produce DIF, is social desirability or presenting oneself in a good light.

The social-cognitive model asserts that response to survey items assumes a sequential process and that the question-answering process involves four cognitive tasks: (1) question interpretation; (2) retrieval of information from memory to answer the question; (3) judgment formation and response formation; and (4) response evaluation and response editing. Measurement error, including DIF, can be introduced at any of these four stages (e.g., a word might be misinterpreted, respondents may not be able to access information from their memory, and the formed response may be mismatched to the response options). However, survey researchers believe that error introduced at stages one through three are mainly unintentional, whereas error introduced at stage four is believed to be intentional and due largely to social desirability (Johnson & van de Vijver, 2003). Stage four of the question-answering process essentially involves organizing and articulating symptoms, behaviors, and feelings in light of what respondents feel others will expect is appropriate for someone like them (whether "them" is defined by gender, age, race, ethnicity, etc). Response editing (deciding what exactly to report once the response has been formulated) may or

may not occur depending on the item in question, its threat value, and social and cultural norms held by the respondent about the item under review.

Survey questions on sensitive topics, such as mental health or physical disability, are likely to activate cultural perceptions of desirable or undesirable responses by different population subgroups depending on their expression norms (Angel & Thoits, 1987), which are norms about the disclosure and display of behaviors and feelings to others. Further, different types of items will activate different expression norms. For example, somatic symptom items will generate less response editing than will affective items because they are less threatening. It must be remembered that self-reports do not directly tap symptoms of physical or psychological distress. Self-reports only assess such symptom states through the cognitive processes involved in perceiving, categorizing, filtering, and interpreting said symptoms (Chang, 1985). As Angel and Thoits argue, although physical or emotional experiences may be privately regarded as significant and problematic, willingness to report or express them publicly (to an interviewer, clinician, or anyone else) may be low. Thus, we hypothesize DIF to be a result of adhering to cultural norms (with culture defined broadly) about expression and disclosure of desirable or undesirable behaviors and feelings.

QUALITATIVE DISCOVERY METHODS FOR DIF RESEARCH

Stricker and Emmerich aptly describe the state of current research on DIF when they refer to DIF as "a phenomenon that has thus far remained an enigma despite extensive research efforts" (1999, p. 363). Mere inspection of item content rarely provides obvious and comprehensive answers as to why certain items contain DIF and others do not, or why an item exhibiting DIF favored a particular group over another. Thus, qualitative discovery research represents a novel step in attempting to solve these enigmas. Using both focus groups and cognitive interviews, one can generate qualitative data on the ways in which response styles (including social desirability) and other cognitive, psychological, and motivational processes of respondents influence the four stages of the question-answering process. Such qualitative data can be used to identify specific terms that raised problems in item interpretation (stage one of the question-answering process) and identify problems in memory retrieval that respondents noted (stage two). Such qualitative research will enable researchers to contextualize issues related to judgment formation (stage three) within and across different measures and different subpopulations. Such discovery research will allow researchers to identify all of the different response editing strategies (stage four) noted by respondents and why they were motivated. Traditional psychometric analyses utilized in quantitative DIF work could never reveal these cognitive processes that generate DIF and compromise measurement equivalence across diverse groups.

CONCLUSIONS

Measurement inequivalence generally and DIF specifically are serious potential threats to validity. Tools containing DIF items may be invalid for between-group comparisons because their scores are indicative of attributes other than that which the test is intended to measure. Culturally-fair health status assessment is crucial when individual decisions are in balance, such as with mental or physical health screening and diagnostic decisions. If items in health status instruments are biased, detection rates can be biased, leading to over- and underdetection and over- and undertreatment. Due to DIF, population forecasts for need for services or resource allocations could be flawed, and research on health disparities could be misguided.

We need to continue to use sophisticated multimethod research to advance our knowledge base about characteristics of items and population groups that cause items to have parameters that are invariant.

Importantly, DIF needs to be assessed *a priori* when measures initially are being developed instead of after they have been in long use. Furthering this line of research will bring about necessary change in how scientists and clinicians conceptualize and develop patient self-report measures of physical and mental health. This much-needed attention to DIF will result in measurement tools that are relevant and fair to members of a multicultural society and that result in equitable treatment for individuals and groups.

REFERENCES

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

Berzon, R., Hays, R. D., & Shumaker, S. A. (1993). International use, application and performance of health-related quality of life instruments. *Quality of Life Research, 2,* 367–368.

Bowling, A. (2001). *Measuring disease: A review of disease-specific quality of life measurement scales* (2nd ed.). Buckingham, UK: Open University Press.

Chang, W. (1985). A cross-cultural study of depressive symptomology. *Culture, Medicine, and Psychiatry, 9,* 295–317.

Day, J. (1996). *Population projections of the United States by age, sex, race, and Hispanic origin: 1995 to 2050.* Washington DC: U.S. Government Printing Office.

Deutscher, I. (1973). Asking questions cross-culturally: Some problems of linguistic comparability. In D. P. Warwick & S. Osherson (Eds.), *Comparative research methods* (pp. 163–186). Englewood Cliffs, NJ: Prentice-Hall.

Fleishman, J., Spector, W., & Altman, B. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology, 57B*(5), S275–S284.

Johnson, T., & van de Vijver, F. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. van de Vijver, & P. Ph. Mohler, (Eds.), *Cross-cultural survey methods* (pp. 193–202). Hoboken, NJ: Wiley.

Kim, Y., Pilkonis, P., Frank, E., Thase, M., & Reynolds, C. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging, 17*(3), 379–91.

McHorney, C. (1997). Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine, 127*(8, Suppl.), 743–750.

McHorney, C. (2002). Use of item response theory to link three modules of functional status items from the Asset and Health Dynamics Among the Oldest Old Study. *Archives of Physical Medicine and Rehabilitation, 83*(3), 383–394.

McHorney, C. (2003). Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Annals of Internal Medicine, 139*, 403–409.

McHorney, C., & Bricker, D. (2002). A qualitative study of patients' and physicians' views about practice-based functional health assessment. *Medical Care, 40*(11), 1113–1125.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Stewart, A., & Nápoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care, 38*(9, Suppl. II), II-102–II-24.

Stommel, M., Given, B., Given, C., Kalaian, H., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Research, 49*, 239–250.

Stricker, L., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement, 36*(4), 347–366.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Task Force for the Handbook of Psychiatric Measures. (2000). *Handbook of psychiatric measures.* Washington, DC: American Psychiatric Association.

Teresi, J., Holmes, D., Ramirez, M., Gurland, B., & Lantigua, R. (2001). Performance of cognitive tests among different racial/ethnic and education groups: Findings of differential item functioning and possible item bias. *Journal of Mental Health and Aging, 7*(1), 79–89.

Teresi, J., Kleinman, M., & Ocepek, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651–1683.

U.S. Department of Health and Human Services. (n.d.) *What are its goals?* Retrieved May 30, 2004, from the Healthy People 2010 Web site

http://www.healthypeople.gov/About/goals.htm

U.S. Department of Health and Human Services. (2003). *Dynamic assessment of patient-reported chronic disease outcomes.* RFA-RM-04-011. Retrieved May 30, 2004, from http://grants.nih.gov/ grants/guide/rfa-files/RFA-RM-04-011.html

Warnecke, R., Ferrans, C., Johnson, T., Chapa-Resendez, G., O'Rourke, D., Chávez, N. et al. (1996). Measuring quality of life in culturally diverse populations. *Journal of the National Cancer Institute Monographs, 20*, 29–38.

SESSION 3 DISCUSSION PAPER: Cross-Cultural Challenges in Health Survey Research

Richard B. Warnecke, University of Illinois at Chicago

These five papers present two sides of the problem of designing interviews for conducting cross-cultural health surveys. Clearly, it is increasingly difficult to assert that there is a single cultural perspective among the populations of the U.S. or in most of Europe. As trends in population migration and the resulting population diversity grow, we can expect that collecting and interpreting data obtained from population surveys will become increasingly complex, and finding ways to generalize from survey results either within any country or transnationally will be more challenging. Approaches that in the past have enabled us to track trends in health and illness, establish policy, and assess the need for health services or even interact with patients at the clinical level are likely to be increasingly problematic.

The papers by Harkness and Willis provide an excellent overview of the issues related to establishing equivalent cultural constructs and developing appropriate survey questions that incorporate them. These two papers address the basic issue of validity—"How do we know that the respondent is answering the question that we are asking?" The papers by Doyle and Blake, Boyle and Wilkinson, and Murray et al. illustrate the reality of how various surveys or survey organizations have dealt with the problems that arise from the need to obtain comparable data from multicultural populations in a single country or across a variety of countries. These papers highlight the issues that may constrain how we translate what we know about cultural equivalence and its likely effect on validity of survey data to the large surveys and the statistical offices that currently provide the information used to establish policy.

The need for translating what we know about the effects of changing patterns of culture on questionnaire design is increasingly recognized at national levels, and efforts are being made to address the issues, as noted in the papers presented in this section. However, as Doyle and Blake clearly and directly describe, there are tensions between the bureaucratic commitments to existing large data sets and costs associated with change and movement to address the impact of increasing cultural diversity in ways recommended by Harkness and Willis and in McHorney's very interesting discussion.

The solution to translation adopted by Abt in the National Immunization Survey (NIS) addresses the problem created by the need to administer that interview in many languages, but at least in the paper, there is no indication that the challenges of arriving at culturally appropriate translations were addressed in the translation process. The costs of applying the recommended translation strategies would probably be prohibitive in a survey of that magnitude, and it is not clear that the data collected by the NIS require extensive efforts to assess conceptual equivalence. On the other hand, some evidence of preliminary testing to address that issue would have been reassuring. Moreover, the paper refers to a second survey of Children with Special Needs, which is more complex. For that survey, they apparently plan to rely on fluent interviewers and direct translation even to the point of including translators as part of the data collection team, which is the same strategy employed in the survey described by Doyle and Blake. These efforts will ensure that the respondents understand the questions, but it is unclear whether these approaches will address the underlying issues related to cognitive equivalence. Even if one can get the words of the question in a format that can be understood by the respondents, the translated words may not mean the same thing to respondents from other cultures as it does to the investigators. Unless the respondent is

asked—or an expert is asked—what question is being answered, the resulting information may not be meaningful to the issue around which the survey is designed. As such data are added to existing data files, the resulting trends that are used for policy may reflect response errors due to differing understanding of the question content rather than differences relevant to the purpose of the survey. Moreover, the addition of a translator to the team may change the dynamic of the interviewing process in unanticipated ways that may further affect the relevance of the responses obtained.

Willis presents a schema that incorporates expert review, cognitive testing, and behavioral coding to assess linguistic and cognitive equivalence. It provides an excellent framework for ascertaining what respondents may be thinking in response to a particular question and whether they are experiencing problems answering the questions as asked. There need to be ways to incorporate that type of evaluation into the large-scale data collection efforts. Even if change does not result, at least there will be indications of where problems are likely to arise. The efforts described by Doyle and Blake seem to be moving in that direction. With sufficient evidence, the "sacred questions" eventually may be evaluated.

Two concrete examples from our own experience with assessing whether a quality-of-life index could be used with African-American and Latino respondents illustrate some of the kinds of problems that may arise when a scale or index is adopted in the absence of cognitive evaluation (Warnecke et al., 1996). The index in question had a strong pedigree and excellent psychometric properties and had been used with Latinas with breast cancer and African-American cancer patients.

Respondents who were treated for cancer were first asked to rate how satisfied they were with various aspects of their lives; they then were asked to rate the importance of each aspect regardless of level of satisfaction. The rating scales presented to the respondents each contained seven verbal descriptors ranging from "very satisfied" to "very dissatisfied" and from "very important" to "very unimportant," respectively.

As we administered the items, we noticed respondents were having difficulty responding. We tested the discriminant and convergent validity of these scales using as a visual analogue a thermometer in which the score of "0" represented the negative end, "50" represented the midpoint (neither), and "100" represented the most positive point. We then asked the respondents to indicate where on the thermometer they would place a number of written descriptors ranging from "very satisfied" to "very dissatisfied" and from "very important" to "very unimportant." The descriptors included the terms "neither satisfied nor dissatisfied" and "neither important nor unimportant." When we evaluated these results, we discovered that the respondents did not discriminate in the way intended, and there was considerable overlap in the values assigned to each descriptor. Moreover, the midpoint descriptors ("neither satisfied nor dissatisfied" and "neither important nor unimportant") were usually scored as zero. Upon further probing, we found that respondents decomposed each scale and treated each as ranging from the neutral point through "very satisfied" or "very important" and from the neutral point to "very dissatisfied" or "very unimportant."

A second illustration also resulted from our attempts to evaluate this response scale. We did the initial cognitive testing in Chicago where most Latino respondents were to some degree bilingual. The Chicago respondents did not seem to have difficulty with the descriptors, although they still had problems using the overall scale. When we tested the same descriptors with Latino respondents in Houston who were not bilingual, we discovered that the terms "dissatisfied" and "unimportant" had no conceptual equivalence to any terms in Mexican Spanish. This finding indicated that Latino respondents probably did not understand the negative response categories.

We solved both problems by creating simple five-item scales ranging from "very

satisfied" to "not at all satisfied" and from "very important" to "not at all important." We expressed the choices in the format of a simple bar graph in which five bars were ordered by height and the end points of the graph were labeled with the extreme values. This strategy worked very well, and subsequent cognitive interviews confirmed that respondents understood the scales and used them appropriately.

In summary, the initial scale contained constructs ("dissatisfied" and "unimportant") that were not cognitively relevant to Mexicans. In addition, the scale itself was overly complex and understood by neither the African-American nor the Mexican respondents. Despite the fact that the original scale had been used in other research that included African-American and Latino respondents, it did not work with the subjects in this research. Without cognitive testing, these problems would not have been identified. It is also important to note that the revised scale addressed the issues of cultural equivalence for the respondents who had difficulty and also improved the response of those who did not have these problems.

As individual researchers, we are able to take these issues into account as we develop or revise interview schedules for new research. However, there is a tension between being able to report trends over time while still accounting for changing cultures in the population that in the future will be providing data for these trends. As the diversity of populations in many countries that collect national statistics grows, it is likely that it eventually will be very difficult to ignore the questions about validity trend data, given increased likelihood that the underlying concepts may not be equivalent across cultural groups. In some cases where the concepts are not complex, as in the study by Murray et al., it may be sufficient to simply translate. The method of translation described by Murray et al. sounds reasonable if the underlying concepts are consistent cross-culturally. Both Doyle and Blake and Murray et al. describe plan to use translators with the interviewers. The use of onsite translators may affect the interviewer-respondent relationship, which is one of the advantages of face-to-face interviewing. This strategy clearly needs to be evaluated for potential mode effects. Boyle and Wilkinson's approach of using multiple data sets and multiple questions to ensure that all respondents in each country use the same criteria for reporting incidence of asthma and COPD is a creative, interesting, and apparently effective approach to the problems created by variations in criteria for disclosure across cultures, but it may be limited by the availability of alternative data sources and issues of cost and efficiency.

It may be useful to approach these problems as they arise. It would seem wise to initiate studies that assess the quality of data being collected in ways that account for varying cultural constructs and the resulting impact on the actual responses. Special studies of ethnic populations would be one approach. One effective strategy described in the literature has been the use of confirmatory factor analysis in a covariance structure analysis to test whether respondents from different cultures respond to the same questions from shared conceptual frameworks (Riordan & Vandenberg, 1994).

The results of these analyses could provide guidance for deciding when there are real differences in the conceptual frameworks of the respondents by comparing variance-covariance matrices across cultural groups. Where the matrices are clearly different, the groups are not using the same latent cognitive framework to address the questions. In such cases, the resulting data cannot be combined meaningfully. In cases where there are some differences in the resulting variance-covariance matrices, the modification indices that are supplied by the analysis may help select items that will produce a common conceptual framework.

Another strategy might be to periodically monitor interviews using Willis's schema and then further examine questions that appear to be causing respondents difficulty or that cultural experts question. These questions

could then be further tested using cognitive interviews.

These are steps that can be taken without changing an entire dataset. In the end, as these population changes continue, the trend data are likely to reflect response error rather than meaningful patterns. It might be worthwhile to consider introducing strategies that allow for modest change in the face of clear evidence of cultural effects.

REFERENCES

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20,* 643–671.

Warnecke, R., Ferrans, C., Johnson, T., Chapa-Resendez, G., O'Rourke, D., Chávez, N. et al. (1996). Measuring quality of life in culturally diverse populations. *Journal of the National Cancer Institute Monographs, 20*, 29–38.

SESSION 3 SUMMARY

Timothy P. Johnson, University of Illinois at Chicago
Tenbroeck Smith, American Cancer Society

The discussion for this session lasted approximately 50 minutes. The following general themes were considered during that discussion:

- Identifying and measuring health constructs that are appropriate for multiple groups.
- Distinctions among subgroups when studying minority populations.
- Evaluating the cross-cultural transferability of our research methods.

In addition, questions were asked about the roles of physician culture and translation in cross-cultural research.

CROSS-CULTURAL HEALTH CONSTRUCTS

Ideally, the field of health survey research would identify a concept and measure of health that is equivalent across cultures, but whether this can be achieved remains in question.[1] Given that health is multidimensional, the first step toward a truly culturally equivalent measure of health would be to define a parsimonious yet inclusive set of constructs of health. The greatest difficulty lies in the fact that definitions of health vary among cultural groups. For example, cognitive testing has revealed that respondents from some cultures base their response to the common global health-rating question[2] solely on physical health, whereas those from other cultures also may consider their mental and/or spiritual condition. Qualitative methods were identified as perhaps the most useful approach to this challenge.

In contrast to a parsimonious set of health constructs, it may be impossible to achieve cross-cultural equivalence for certain domains where cultures do not share underlying structures. For example, questions about health services and insurance may be difficult because systems vary from country to country. Similarly, if a group does not have knowledge on a subject, it doesn't matter how the question is asked (e.g., asking recent United States immigrants about the U.S. health care system). For those domains for which different cultures do not share underlying concepts, there will be no single answer. In addition, some subgroups will require special methods in order to achieve acceptable response rates. For these domains and subgroups, measurement will move more towards understanding individual groups, and measures will not be totally equivalent. As new subgroups are identified, special studies—including cognitive testing and focus groups—will be required to understand their needs.

In the context of special studies, the role of community partnership in interpreting survey data was discussed. It was observed that one of the reasons for community partnerships in

[1] This issue was raised in the context of Colleen McHorney's discussion of the development of item banks for Computer Adaptive Testing (CAT) to assess self-reported health status. Ideally, these item banks would be developed from the beginning to be culturally equivalent, as otherwise it can be difficult to export measures to new cultures effectively. McHorney indicated she was concerned about repeating past mistakes and would like to see a "monocentric" item bank that could readily be exported to other countries. Unfortunately, the efforts she described are not headed in that direction because other countries are not involved. These efforts are based on an NIH RFA focused on clinical trials, and not population assessment. It was noted that consideration of U.S. subpopulations (e.g., African Americans and Hispanics) would not only make these item banks more broadly applicable within the U.S. but also readily portable to other countries. The NIH RFA calls for the development of item banks for the following constructs: physical health, pain, mental health, and fatigue. Finally, the issue of the appropriateness of using CAT in population-based research was raised.

[2] In general would you say your health is: Excellent… Poor.

research is that they provide the opportunity to associate context with the individual responses. One weakness of past survey research has been an inability to take context into account.

Finally, a cautionary note was struck with regard to differences between groups. Differential Item Functioning (DIF)[3] implies that cross-cultural challenges lead to measurement error. However, cultural differences on measures are not always due to measurement error but sometimes are simply part of the differences between groups. Some population differences are part of the robust measurement, not error. This concern was initially expressed in terms of measurement error: can a point could ever be reached in population research in which all cross-cultural differences are no longer regarded as measurement error but rather become an important part of the construct being measured and therefore do not require correction? In all likelihood, the tension between universal constructs—which allow comparisons between cultures—and special studies—which investigate the unique issues of any given culture—always will exist in health survey research.

DISTINCTIONS AMONG SUBGROUPS

A related concern was raised regarding the seemingly endless diversity of population subgroups.

The definition of cultural groups need not be restricted to race and ethnicity. Unique cultural groups can be defined along different dimensions, such as age (the elderly), sexual preference (the gay community), or political party (Democrat/Republican). One discussant likened considering this multitude of cultures in study design to opening Pandora's box.

One presenter commented that a practical way of handling this potential "Pandora's box" is to pay attention to the measurement objectives for the current investigation and recruit relevant population subgroups accordingly. For example, the research must consider whether differences exist between Democrats and Republicans on the outcomes of interest, and if so, if these differences are relevant to a given research question. Another discussant commented that this is an issue that is relevant for all questionnaire design, not just cross-cultural work. At minimum, researchers should invest effort in finding out if questions are understood within the subgroups of interest and if those questions are perceived by respondents in a way that is consistent with the underlying constructs being measured.

Related to this topic is the issue of identifying invisible minorities. For example, Caucasians and African Americans are generally treated as homogeneous strata in the U.S. Within each racial category, however, there are invisible minorities (e.g., Russian immigrants, Caribbean Africans). As new groups are identified, they need to be better understood though special studies similar to those noted in the previous section.

CROSS-CULTURAL TRANSFERABILITY OF RESEARCH METHODS

A third challenge that was not mentioned directly during the presentations concerned the transferability of qualitative research methods across groups. Many of the issues of comprehension in survey research raised during the presentations also apply to the methods of qualitative and cognitive research now being utilized to improve the cross-cultural validity of measures. A series of examples of potential limitations or confounds to conducting this sort of research then were given. Focus groups are based on an individualistic paradigm in which individuals feel comfortable sharing their opinions in a group; this may not work with people from collectivist cultures. Some elderly immigrant populations might be intimidated by sterile laboratory settings, and some groups will not come to a laboratory. (It was suggested that a solution would be to conduct these studies in homes or community centers.) Respondents from various subgroups also may respond differentially to interviewers. In the United

[3] Gary King's work using vignettes to anchor items in a scale was mentioned as a potential solution to the problem of DIF.

States, it is assumed that respondents will tell the truth in the absence of some item-specific reason for not doing so. For persons coming from totalitarian regimes, that assumption may be incorrect. There may be some groups for which everything is sensitive. Finally, the notion that opinions are valuable, no matter how ill-informed, is not the norm that all respondents bring to an interview. Consequently, methods may need to be tailored to specific populations. This led to a call for methods research on the methodologies used to evaluate and achieve cross-cultural research.

Issues with sample selection for focus groups and cognitive research also were addressed. Typically, this research cannot rely on random samples to provide adequate cultural variance because sample size is generally small. Therefore, it is incumbent on the researcher to consider study design, identify the relevant cultural groups, and purposefully select cases to insure inclusion of these groups. For example, a study of Latinos would be misdirected if only Puerto Ricans were included in focus groups and cognitive research, since it would miss issues that affect Mexicans and Dominicans.

OTHER TOPICS

In the context of a discussion about adaptation and centering, the problematic nature of our assumptions about what we are measuring was observed. For example, in one study, researchers interpreted the meaning of parents yelling at their children to be the same for both Caucasian and African-American parents; however, the norms of good parenting vary by culture. Another example provided was that optimal Body Mass Index (BMI) and Hip-Waist Ratio (HWR) may vary by race. Consequently, stepping back from the questionnaire is necessary to insure that the constructs being represented are appropriate—that is, considering whether the answer means what we think it means. One presenter agreed, commenting that "questions are really just indicators of things we want to get at." The semantic content of a question is not necessarily synonymous with its pragmatic meaning or with the underlying construct that the researcher is measuring.

A question was directed to Janet Harkness concerning the current role of back translation techniques in cross-cultural research. She commented that she would like to see translation included as part of the questionnaire design process. Where this is not possible, it should at least be a part of the adaptation process. And where this cannot be done, the translation must at a minimum be tested in an appropriate fashion. She observed that back translation was a "rough and ready way of looking for errors," and as such, the first approach to come into wide use. Researchers today have more sophisticated approaches at their disposal. For example, relying upon a panel of experts allows for the evaluation of both the semantic and pragmatic interpretation of a question, whereas back translation focuses primarily on the semantic meaning (as noted in the previous paragraph, semantic and pragmatic meaning do not always coincide). However, the field is in transition from viewing back translation as a gold standard to using more sophisticated approaches. One indication of this transitional state is that more sophisticated approaches are sometimes referred to as "back translation." Another indication is that many review bodies (e.g., IRBs) still view back translation as a gold standard, accepting it without question, but they often require those using a more sophisticated, up-to-date translation methods to provide proof that the method in question is effective. Several people added that including bilingual interviewers in the translation process was very helpful because it led to the development of questions that were easier to use in the field. Alternatively, in situations where an interpreter is required, it is important to consider the space the interpreter occupies: he or she can be closer to the interviewer or closer to the respondent.

Conference participants also considered the potential role of the cultures of physicians and other health care providers. It was observed that leading physicians frequently disagree on the parameters of disease and benchmarks for standards of care. Even when

consensus conferences adopt diagnostic and treatment standards, there is at best a slow dissemination and adoption of new standards among the physician community. Awareness and adoption usually occur differentially among the subcultures of tertiary care physicians, other specialists, and primary care doctors.

INTRODUCTION TO SESSION 4: How to Conduct General Population Surveys in the 21st Century

Floyd J. Fowler, Jr., University of Massachusetts Boston

One of the most important changes in the health survey landscape today is the declining feasibility of random-digit-dial (RDD) telephone surveys for producing credible data. Some of our most important federal surveys, as well as countless local surveys, rely on this methodology. Response rates often are very low, and the potential threat of the deterioration of the sample frame looms on the horizon. Meanwhile, Web-based surveys and mail surveys may offer potential substitutes or perhaps complementary components for dual-mode protocols. In-person interviews may continue to be the gold standard; they also may be most feasible when used in some cost-saving combination with some other mode.

In this session, we want to take a big-picture view of where we are with respect to how to do general population health surveys in 2003. From a total survey design perspective, the discussion has to consider costs, the quality of sample frames, the rates and biases associated with nonresponse, and the issues of data quality and data comparability associated with alternative modes of collection data.

The presentations in this session were chosen to set the stage for an informed and thoughtful discussion. Martin Frankel provides his view of the present and future of surveys done solely based on random-digit dialing (RDD). Blumberg and associates present data from the National Health Interview Survey on one of the threats to RDD surveys: households that have substituted individual cell phones for household telephone service. Papers by Baker and Zahs, Link and Mokdad, and Gallagher and me explore the potential of Internet, mail, and in-person interviewing, respectively, to complement or substitute for RDD-based telephone surveys. Finally, our discussants were asked to think broadly about two aspects of the challenges we face. Couper addresses issues related to nonobservation due to limitations in the sample frames that are used and to nonresponse. Dillman discusses the challenges related to measurement when data are collected using more than one mode. All of our presenters were asked to address the state of our current knowledge and what research is most needed to prepare us to do high-quality surveys in the next five or ten years.

FEATURE PAPER: RDD Surveys: Past and Future

Martin R. Frankel, Baruch College, CUNY

INTRODUCTION

Practitioners of survey research often lose sight of the impact that surveys in general and telephone surveys in particular have had on American and worldwide culture. Surveys provide "facts" and other statistical information for academic researchers, government policy makers, the news media, marketers of goods and services, charitable organizations, political office holders and seekers, and the legal profession.

The phenomenal success of RDD surveys in the U.S. and elsewhere is an example of the "better mousetrap" story. Several factors paved the way for this development: the telephone itself, implementation of direct distance dialing, the digital computer, and the availability of computer readable directory listings.

Perhaps the most important and obvious factor that permitted the development and use of RDD surveys was the development of the telephone and associated technological infrastructure. Most students are taught in grade school that Alexander Graham Bell invented the telephone. Actually, Bell's 1876 patent application was fought by Antonio Meucci, who claimed to have invented the telephone in 1849, and by better-known inventor Elisha Gray. Bell also had to deal with competitors who simply ignored his patent and built their own telephone systems. It is clear from a fascinating book, *The History of the Telephone* by Herbert N. Casson, published in 1910, that the telephone had an impact on society comparable to the impact of the personal computer and the Internet. By 1898, a long-distance line linked New York and Chicago. By 1903, more than three million telephones were in use. By 1910, the Bell Company was contemplating long-distance lines with overseas countries. Over time, telephone infrastructure grew and improved, but not until the 1960s was it possible to direct dial (using a ten-digit number) almost any other telephone subscriber in the United States.[1]

EARLY DEVELOPMENT OF RDD

The collection of survey information by telephone probably dates to the 1930s or before. Documentation indicates that local telephone surveys, based on the selection of random numbers within specified exchanges, were in use by the early 1960s (Frankel & Frankel, 1977). Once direct long-distance dialing was possible, survey practitioners were able to select national samples (but not strict probability samples) of telephone households. Sampling methods typically involved selecting primary sampling units (PSUs), obtaining published directories for these PSUs, and selecting lines from the telephone directory to produce telephone numbers. Some organizations applied various randomizations to these numbers in order to include numbers that were "unpublished." For example, one strategy was to add 1 to the last digit.

The development of a methodology for valid probability sampling of all telephone households did not occur until the 1970s. In their 1972 paper, Glasser and Metzger laid out the methodology for generating ten-digit random numbers within the confines of assigned six-place area code prefix combinations that were in use in the U.S. The specification of all valid area-code prefix combinations was available on a computer tape that could be obtained from AT&T Long Lines Division at a modest charge. Glasser and Metzger reported that with this methodology, it was possible to generate a valid national probability sample of telephone households. The limitation was that only about 20% of the generated numbers produced working households.

[1] The North American Numbering Plan (NANP) was first published in 1947, but the adoption of the Numbering Plan Area code (NPA) was not complete until the early 1960s.

The availability of the digital computer and the production of a computer-readable list of names, addresses, and phone numbers derived from paper directories led to development of a more cost-efficient method of generating probability samples of all telephone households. The development of the methodology that is now known as list-assisted RDD sampling was carried out independently by two groups. One of these groups was led by Thomas Danbury, who subsequently founded Survey Sampling, Inc. (T. Danbury, personal communication, 1977). The second group was a team of statisticians and computer programmers at the A. C. Nielsen Company (A. C Nielson Company, 1976). Both of these groups used the computerized lists of residential numbers to produce a frequency distribution of the number of "directory listed" numbers in all possible banks of 100 numbers. By restricting the generation of ten-place random numbers to those banks that contained two or more listed numbers, they were able to achieve working-number rates near 50% while missing under 5% of all telephone households. In the same time frame, Joseph Waksberg and Warren Mitofsky, working for CBS news, developed a two-stage sampling method that carries their names (Waksberg, 1978).

THE GOLDEN AGE OF RDD SAMPLING

By 1980, the development of the two basic RDD methods made it possible for survey researchers to generate or purchase efficient and valid national and sub-national probability samples of both listed and unlisted telephone households. Two additional factors contributed to the widespread success of RDD telephone sampling. The first factor was that by 1980 the percentage of U.S. households that had at least one telephone line was approaching 90%. The second factor that served to make the use of telephone sampling and data collection financially and methodologically appealing was the development of computer programs to carry out computer-assisted telephone interviewing (CATI).[2]

There is general agreement that telephone surveys of all qualities reached their zenith in the period from 1980 to 2000.

THE PRESENT

Since the 1990s, we have witnessed the emergence and growth of substantial impediments to the continued health of RDD surveys. The increasing challenges to RDD and other telephone surveys have been documented in a number of papers, both by practitioners of high-quality surveys and those practitioners who are willing to accept relatively low response rates (Council for Market and Opinion Research, 2002). In certain segments of the research community, RDD surveys and, in fact, all probability samples are considered dead (Spaeth, 2002). Many former users of telephone surveys are undertaking data collection using mail, mall intercept studies, or Internet surveys using pop-ups or e-mail invitations. In other segments of the research industry, telephone surveys with one or two callbacks are used, and response rates of 10%–20% are considered acceptable for practical use.

Several papers and presentations have discussed the factors that have lead to difficulties in conducting RDD surveys (de Leeuw, Lepkowski, & Kim, 2002). My list of factors includes:

- Telemarketing,
- The volume of telephone surveys,
- The complexities and challenges of daily life, and
- Enhancements to telephone technology.

Telemarketing

Prior to the advent of state and federal Do-Not-Call lists, telemarketing grew to a multibillion-dollar industry. Most unsolicited telephone calls received by U.S. households were not for surveys but rather for the

[2] An additional factor was the availability of automatic dialing programs for the elimination of a portion of the "nonresidential" numbers.

purpose of selling something or soliciting a donation. The American Teleservices Association Web site indicates that prior to the DNC Registry, telemarketing generated annual sales of more than $500 billion. It is not clear how much DNC will hurt this industry, but it will certainly survive.

The Volume of Telephone Surveys

The success of RDD surveys led to their widespread acceptance and use. Even though the size of the telemarketing industry far exceeded the survey industry by at least one and possibly two orders of magnitude, households receive a relatively large number of requests to participate in telephone surveys. In the early days of RDD surveys, most people were flattered to be called as part of a nationally representative sample. Trends in response rates, and in particular refusal rates, show that this situation has certainly changed over time.

The Complexities & Challenges of Daily Life

In the U.S., it is generally accepted that the complexities and challenges of daily life have increased over time (Peers, 2004). It appears that people spend more time away from home. Even when they are home, most individuals have more demands for non-leisure activities and more options for leisure activities. From trends in cooperation and response rates, it appears that an increasing number of persons would prefer to spend their available time on something other than participating in a survey.

Enhancements of Telephone Technology

The first enhancement of telephone technology that had a negative impact on the RDD survey was the answering machine. By using a telephone answering machine, it is possible not only to receive information from wanted calls when you are unavailable to answer the phone but also to apply real-time call screening. It is possible to hear callers identify themselves prior to actually accepting a call with a "hello." The first answering machines were electro-mechanical devices that used recording tape. Now they are digital and often are part of the telephone unit itself. Call answering is also an option offered by many local telephone companies.

The second major enhancement that has had a negative impact on telephone surveys was the decision of the telecommunication industry to market the electronic signature of the originating number that accompanies the ring signal sequence. This caller ID allows people to "see" the originating number (or the fact that this tag is blocked) before picking up the telephone. For an additional charge, this ID is accompanied by a "listing" associated with the originating number. A number of local telephone companies offer features such as "anonymous call rejection," which intercepts calls from originating numbers with blocked caller IDs prior to the actual telephone ring. Other services include call blocking (preventing an incoming ring) from numbers with caller IDs that do not match a list of acceptable caller IDs. Other stand-alone devices are available that deliver a message to the caller that the number should be placed on their "do-not-call list."

The third technological enhancement that may have a substantial impact on telephone surveys is the cellular telephone. In current practice, most RDD surveys eliminate banks of telephone numbers that are designated as cell-phone exchanges. If the number of persons who do not have landlines but rely exclusively on cell phones increases, current practice will exclude this portion of the population from RDD surveys. In addition to the potential challenges associated with cellular telephones, there is the recent development of "number portability." It is now possible for cell-phone subscribers to take their number with them when they switch cell-phone providers. This practice also may be implemented for landline service. Survey organizations that use RDD at the subnational level may face increased noncoverage because of number-portability across geographic areas.

THE FUTURE OF RDD SURVEYS

RDD surveys are at a crossroads. Many researchers have predicted that telephone surveys will probably disappear from use. I do not share this belief. Many telephone surveys will disappear, but the method will remain viable. This viability is driven by the lack of any lower-cost but valid probability sampling alternative. Even when Internet use approaches the 90% levels of telephone coverage, a valid sampling methodology is still lacking. Unless some lower-cost method is found for door-to-door in-person interviewing, the RDD survey will remain our only alternative.[3]

Telephone surveys based on RDD methods will certainly change. Here are some of the factors that I predict will influence the nature of telephone surveys.

Federal Do-Not-Call Registry

The Federal Trade Commission DNC Registry will result in a decrease in the number of unwanted telephone calls received by individuals. People who put themselves on the list actually may be more receptive to certain types of RDD survey requests.

A Decline in the Number of Telephone Surveys

There will be a further decline in telephone surveys conducted by market, opinion, political and survey researchers. The increased cost and difficulties associated with telephone surveys will drive certain types of survey research to mail, Web, and intercept surveys. The net result will be a decrease in the number of solicitations for telephone surveys in the general population. This may help the remaining high quality RDD surveys.

More Research Will Be Carried Out on the Differences Between Responders & Nonresponders

Further research will result in the acceptability of lower response rates or in the development of adjustments. Some evidence indicates that, if adjustments are necessary, they will be developed. The recent development of "telephone interruption" adjustments for noncoverage represents such a development.

Respondents Will Be Paid for Their Participation

In earlier days, a large majority of the population was willing to participate in surveys without receiving any monetary benefits. Various social exchange theories postulated that respondents received sufficient "nonmonetary" benefits. I believe this situation has changed. Potential respondents who do not find the "non-monetary" benefits of survey participation sufficient will not be persuaded by offers of $1 or even $5. It will be necessary to provide more substantial compensation. It is not unreasonable to expect that it will be necessary to offer potential respondents $15–$25 for participation in a short survey and $50 or more for participation in a longer survey. This compensation will be effective.

Better Methods Will Be Developed To Invite & Achieve Survey Participation

In the future, RDD telephone surveys will pay more attention to the survey recruitment process. The positive impact of the "advance letter" is well documented, but the simple advance letter may not be enough. The growth and success of the direct mail business has led to modification in mail-opening behavior. Junk mail is as common as e-mail spam and pre-DNC telemarketing. It may be necessary to introduce a survey and request participation via some type of premium mail or delivery (e.g., Federal Express or Express Mail). This type of advance contact is more expensive than the typical advance letter, but it may be necessary to demonstrate the sincerity of the survey sponsor and to deliver a pre-participation payment. More attention will be paid to the "packaging" of advance materials.

[3] I have excluded mail surveys from this discussion because, though structure listings are available for a large proportion of the U.S., the lack of up-to-date name information results in substantial non-coverage.

Multiple Modes of Participation Will Be Offered to Respondents

In order to maximize respondent cooperation, it will be necessary to offer potential respondents multiple modes by which they may respond to survey questions. It will be necessary to provide callbacks at predetermined times. It will be necessary to allow a respondent to respond over the Web, by mail, and even by calling a toll-free number at the respondent's convenience. It will be necessary to more carefully study the potential differentials associated with mixed-mode data collection.

Interviewer Qualifications & Training Will Become Even More Important

Given the increased costs associated with telephone surveys, increased expenditures on interviewer training will be seen as more cost effective. Most of us who have monitored telephone interviewing recognize the large quality variation that exists among interviewers. These differentials are also evident in any analyses of "interviewer effects" as part of a total survey error analysis. A decrease in the number of telephone surveys may result in a relative increase in the pool of excellent interviewers. However, compensation levels for these desirable interviewers will increase. Further, in order to attract more qualified and successful interviewers, it may be desirable to decentralize the interviewing process. New technology should make this possible.

CONCLUSION

Over the next decade or so, telephone surveys will become more expensive. Even with this higher cost, RDD sampling will remain in use until an equally valid but lower cost, probability sampling approach is developed. With all of their limitations, telephone surveys offer the only scientifically sound alternative to in-person surveys based on area or list-based probability sampling. Given this lack of options, RDD surveys may decline in use and volume, but they will certainly survive.

REFERENCES

A. C. Nielson Company. (1976). *Total telephone frame.* [Brochure]. Chicago: Author.

Casson, H. N. (1910). *The history of the telephone.* Chicago: A. C. McClurg & Co. Retrieved June 1, 2004, from the Web site of the Electronic Text Center, University of Virginia: http://etext.lib.virginia.edu/toc/modeng/public/CasTele.html

Council for Market and Opinion Research. (2002). *2002 cooperation tracking analysis.* Cincinnati: Author.

de Leeuw, E. D., Lepkowski, J., & Kim, S.-W. (2002, August). *Have telephone surveys a future in the 21st century?* Paper presented at the International Conference on Improving Surveys, Copenhagen, Denmark.

Frankel, M. R., & Frankel, L. R. (1977). Some recent developments in sample survey design. *Journal of Marketing Research, 3,* 280–293.

Glasser, G. J., & Metzger, G. D. (1972). Random-digit dialing as a method of telephone sampling. *Journal of Marketing Research, 9*, 59–64.

Peers, M. (2004, January 26). Buddy, can you spare some time? *The Wall Street Journal,* p. B1.

Spaeth, J. (2002, August). Life after random sampling. *Informed, 5*(4), 1, 4.

Waksberg, J. (1978). Sampling methods for random-digit dialing. *Journal of the American Statistical Association, 78,* 40–46.

FEATURE PAPER: Has Cord-Cutting Cut into Random-Digit-Dialed Health Surveys? The Prevalence and Impact of Wireless Substitution

Stephen J. Blumberg, Julian V. Luke, and Marcie L. Cynamon
National Center for Health Statistics

The U.S. had 140 million wireless telephone users in 2002. That is nearly one wireless phone for every two persons in the U.S., and 42.6% of all U.S. phones are now wireless (International Telecommunications Union, 2003). Moreover, wireless subscribers average 490 minutes of use per month, which is now greater than the average number of minutes per person per month on residential landline phones (Yankee Group, 2003a). It is perhaps not surprising that many wireless users have considered "cutting the cord"—substituting their residential landline telephone with a wireless telephone.

The sampling frames for most current random-digit-dialed (RDD) household surveys are limited to landline (i.e., wired, fixed) phones. Therefore, the substitution of wireless (i.e., cellular, mobile) phones for residential landline phones may affect the representativeness of these surveys. To better understand the implications of wireless substitution for RDD health surveys, one needs to determine the current size of the wireless-only population and determine if their demographic, socioeconomic, and health characteristics differ from those of the population with landline phones. For example, some studies report that younger adults are more likely to have wireless phone service only (Yankee Group, 2003b). Given that age is related to health status and health care service use, RDD health surveys that exclude wireless-only persons and fail to account for this noncoverage bias (e.g., in the creation of the sampling weights) will produce biased estimates.

What proportion of U.S. adults has only wireless phones? Who are they? Do their health characteristics differ from the population with landline phones? Can these differences be explained by demographic and socioeconomic differences between wireless-only and landline persons? Answering these questions is the focus of this paper.

DATA SOURCE

The National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention added a series of questions to the National Health Interview Survey (NHIS) to assess the prevalence and characteristics of households that have substituted wireless service for their residential landline telephones. The NHIS is an annual multistage probability survey that collects comprehensive health-related information from a large sample of households representing the civilian noninstitutionalized population of the U.S. The face-to-face interview is administered by trained field representatives from the U.S. Census Bureau.

For many years, the NHIS has included questions on residential telephone numbers to permit recontact of participants. In 2003, additional questions confirmed that the number provided was a landline phone. All respondents also were asked whether "anyone in your family has a working cellular telephone." Families were identified as wireless families if anyone in the family had a working cell phone. Households were identified as wireless-only if they included at least one wireless family and if there were no working noncellular phones in the home. Persons were identified as wireless-only if they lived in a wireless-only household. The ownership of the wireless telephones and their primary use (personal or business) were not obtained nor considered in making this assignment.

The analyses were based on preliminary weighted data from the first six months of 2003. Demographic and socioeconomic characteristics, household telephone status, and some health-related information were obtained for 30,991 adults from 16,677 households. From each family, one adult was randomly selected for a more detailed interview about health and health care service use (n = 14,353). Response rates for 2003 are not yet available. In 2002, the household response rate was 89.6%.

ANALYTIC PLAN

Sampling weights were provided by the NHIS Early Release (ER) Program, which produces and releases estimates for key health and health care access measures six months after data collection has been completed for each quarter (Ni et al., 2003). These weights were used to determine the percent of households with only wireless service and the percent of persons living in wireless-only households.

Prevalence estimates also were calculated by race/ethnicity, age, sex, education, household size and composition, household income, employment status, geographic region, Metropolitan Statistical Area status, and home ownership. For households with more than one family, family size, composition, and income were aggregated from the family level to the household level.

For 14 key measures of health and health care service use, prevalence estimates for adults with landline phones were compared with estimates for adults with only wireless service and with estimates for adults without any phone service. The estimates for the key measures were derived using the ER Program's specifications (Ni et al., 2003). Next, to determine if any observed differences in health and health care service use by phone status could be explained by other demographic or socioeconomic differences (e.g., whether wireless-only adults report better health status because they tend to be younger), weighted logistic regression analyses were used to predict health and health care service use from telephone status. All demographic and socioeconomic variables used in the prevalence analyses were included in the regression analyses as concomitant variables. Statistically significant adjusted odds ratios would indicate that wireless-only status still accounts for variance in health and health care service use after controlling for the other demographic characteristics.

All variance estimation and statistical tests were conducted using SUDAAN.

KEY RESULTS

During the first six months of 2003, about 3.2% of U.S. civilian households did not have a landline telephone but had a wireless phone. Among civilian noninstitutionalized adults, 2.8% had wireless phones only. Only 1.6% of adults did not have any phone service (landline or wireless). The percentage of adults with wireless service only is greater for certain demographic subgroups (Table 1). This percentage exceeded 6% for several subgroups, including those age 24 or younger, those with household income below $10,000, and those renting their homes. When compared with adults with landline phones, wireless-only adults also were more likely to be male and be living alone or with unrelated roommates (Table 2). Few

Table 1. Percent of Adults in Households with Only a Wireless Phone by Selected Characteristics: U.S., January–June 2003

Demographic Characteristic	%	Standard Error
ALL ADULTS	2.8	0.17
Race/ethnicity		
Hispanic	3.3	0.38
White, non-Hispanic	2.7	0.19
Black, non-Hispanic	2.9	0.39
Other single race, non-Hispanic	2.7	0.63
Multiple race, non-Hispanic	5.5	1.73
Age		
18–24 years	6.1	0.54
25–44 years	3.7	0.26
45–64 years	1.4	0.17
65 years or more	0.5	0.13
Sex		
Male	3.2	0.20
Female	2.4	0.16
Education		
8th grade or less	2.0	0.38
Some high school	3.4	0.39
High school graduate/GED	2.9	0.28
Some post-high school	3.6	0.30
4-year college degree or more	1.9	0.21
Employment status last week		
Working at a job or business	3.2	0.19
Keeping house	2.4	0.34
Going to school	4.3	1.03
Other (including unemployed)	1.8	0.20
Household size		
1	5.3	0.43
2	2.8	0.29
3	1.8	0.24
4	2.2	0.34
5 or more	2.2	0.47
Household structure		
Living alone	5.3	0.43
Living w/roommate(s)	10.8	2.55
Living w/spouse &/or related adults	2.0	0.23
Adult w/children households	2.4	0.24
Household income		
Up to $9,999	6.3	0.89
$10,000–$19,999	5.8	0.72
$20,000–$39,999	5.0	0.48
$40,000–$59,999	3.2	0.38
$60,000 and over	1.5	0.24
Geographic region		
Northeast	1.3	0.27
Midwest	3.2	0.42
South	3.2	0.27
West	2.9	0.35
MSA status		
Metropolitan	3.1	0.20
Not metropolitan	1.8	0.29
Home ownership status		
Owned or being bought	1.4	0.13
Renting	6.7	0.46
Other arrangement	4.0	1.04

differences in the prevalence of wireless-only service were observed based on race/ethnicity or education.

Relative to adults with landline phones, adults with only wireless service were more likely to report being in excellent or very good health but also were more likely to have had five or more alcoholic drinks on one occasion, to smoke, and to report feelings from the past 30 days that indicate psychological distress. Adults with only wireless service were more likely to be uninsured and have experienced financial barriers to needed

Table 2. Percent of Adults with Selected Characteristics by Telephone Status: U.S., January–June 2003

Demographic characteristic	All (*n*=30,991)[1]	With any type of service (*n*=30,125)	With landline phone (*n*=29,274)[2]	With landline & wireless (*n*=13,479)	With wireless only (*n*=851)	No phone service (*n*=554)
Race/ethnicity						
Hispanic	12.2	12.1	12.0	9.3	14.4	20.5
White, non-Hispanic	71.8	72.3	72.4	77.1	68.5	51.0
Black, non-Hispanic	11.1	10.8	10.7	8.8	11.4	24.0
Other single race, non-Hispanic	4.1	4.1	4.1	3.9	4.0	3.3
Multiple race, non-Hispanic	0.9	0.9	0.8	0.9	1.7	1.3
Age						
18–24 years	13.0	12.8	12.3	13.3	28.3	21.2
25–44 years	39.0	38.8	38.4	42.5	52.3	45.2
45–64 years	31.9	32.1	32.5	34.0	16.5	26.2
65 years or more	16.1	16.3	16.7	10.2	2.9	7.4
Sex						
Male	48.0	47.9	47.6	48.8	55.5	55.7
Female	52.0	52.1	52.4	51.2	44.5	44.3
Education						
8th grade or less	6.1	5.9	6.0	3.1	4.3	14.5
Some high school	10.7	10.3	10.3	6.8	12.6	29.1
High school graduate or GED	30.2	30.1	30.1	27.1	30.6	34.9
Some post-high school	28.7	29.0	28.7	31.6	36.2	16.7
4-year college degree or higher	24.3	24.7	24.9	31.5	16.3	4.7
Employment status last week						
Working at a job or business	65.0	65.2	64.9	73.4	73.2	55.3
Keeping house	7.4	7.3	7.4	7.0	6.3	11.9
Going to school	2.9	3.0	2.9	3.1	4.5	1.8
Other (including unemployed)	24.6	24.5	24.8	16.6	16.1	31.0
Household size						
1	15.2	14.7	14.3	9.2	28.7	33.9
2	33.8	34.0	34.0	33.2	33.8	27.5
3	19.6	19.7	19.9	21.8	12.6	16.6
4	17.0	17.1	17.2	19.8	13.7	11.7
5 or more	14.4	14.5	14.6	16.1	11.2	10.3
Household structure						
Living alone	15.2	14.7	14.3	9.2	28.7	33.9
Living w/roommate(s)	1.4	1.3	1.2	1.7	5.2	2.4
Living w/spouse &/or related adults	44.1	44.5	44.9	44.9	31.9	28.7
Adult w/children households	39.4	39.4	39.6	44.2	34.1	35.0
Household income						
Up to $9,999	6.4	5.9	5.7	2.3	11.6	32.0
$10,000–$19,999	10.9	10.6	10.3	5.1	18.3	29.9
$20,000–$39,999	23.9	23.9	23.5	17.7	34.9	25.3
$40,000–$59,999	19.4	19.6	19.6	19.8	18.3	6.2
$60,000 and over	39.4	40.2	41.0	55.0	17.0	6.7
Geographic region						
Northeast	18.8	19.0	19.3	19.2	8.8	13.1
Midwest	23.4	23.4	23.3	24.3	27.0	19.4
South	37.8	37.5	37.3	37.8	43.3	54.3
West	20.0	20.1	20.1	18.7	20.9	13.2
MSA status						
Metropolitan	73.4	73.4	73.1	72.0	82.7	71.0
Not metropolitan	26.6	26.6	26.9	28.0	17.3	29.0
Home ownership status						
Owned or being bought	71.7	72.8	73.9	80.4	35.3	25.2
Renting	26.4	25.3	24.2	18.1	61.9	70.5
Other arrangement	2.0	1.9	1.9	1.6	2.8	4.2

[1]Includes 312 adults with insufficient information to classify telephone status.
[2]Includes 13,479 adults who also have wireless telephone service.

Table 3. Prevalence Rates for Various Health and Health Care Service Use Measures by Telephone Status: U.S., January – June 2003

Health Measure	Prevalence for Adults			Adjusted Odds Ratios[1] & Confidence Intervals			
	Landline[2]	Wireless only	No phone	Wireless only		No phone	
	%	%	%	OR	95% CI	OR	95% CI
Health-related behaviors							
5+ alcoholic drinks in 1 day at least once in past year	19.1	36.1	25.9	1.48	[3](1.16, 1.87)	1.32	(0.93, 1.86)
Current smoking	20.7	34.0	46.3	1.30	[3](1.03, 1.65)	1.77	[3](1.32, 2.38)
Regular leisure-time physical activity	33.3	35.0	25.1	0.96	(0.76, 1.22)	1.00	(0.72, 1.38)
Health status							
Health excellent/very good	62.2	67.2	48.7	1.08	(0.87, 1.33)	0.94	(0.74, 1.19)
Experienced serious psychological distress in past 30 days	3.0	5.3	6.0	1.43	(0.90, 2.26)	0.88	(0.54, 1.42)
Obesity among adults age 20+	23.7	24.0	29.1	1.09	(0.84, 1.41)	1.17	(0.85, 1.63)
Asthma episode in past year	3.5	4.6	2.2	1.22	(0.71, 2.07)	0.52	(0.27, 1.02)
Diagnosed diabetes	6.3	3.7	4.6	1.05	(0.52, 2.14)	0.66	(0.36, 1.20)
Health care service use							
Has a usual place to go for medical care	86.6	69.9	60.8	0.66	[3](0.51, 0.86)	0.50	[3](0.37, 0.68)
Influenza vaccine in past year	30.2	14.3	17.2	0.69	[3](0.52, 0.92)	0.85	(0.59, 1.23)
Ever received pneumococcal vaccination	16.3	7.5	11.2	0.88	(0.59, 1.32)	1.02	(0.66, 1.56)
Ever been tested for HIV	35.4	45.1	42.2	1.03	(0.81, 1.30)	0.98	(0.77, 1.26)
Did not obtain needed medical care in past year due to financial barriers	5.7	12.2	17.5	1.23	(0.95, 1.59)	1.50	(1.09, 2.07)
Uninsured	15.5	32.1	45.8	1.39	[3](1.11, 1.75)	1.69	[3](1.28, 2.25)

[1]The logistic regression analyses predicted health and health care service use from telephone status. Landline service was the referent. Adjusted odds ratios are based on regression models that included race/ethnicity, age, sex, education, employment status, household size and composition, household income, geographic region, Metropolitan Statistical Area status, and home ownership as concomitant variables.
[2]Includes adults who also have wireless telephone service.
[3]This confidence interval does not include 1.00 and indicates a statistically significant odds ratio, $p < .05$.

medical care; they were less likely to have a usual place for medical care and to have received influenza or pneumococcal vaccinations. Also, HIV testing was more common for wireless-only adults. These differences mirror those observed between adults with no phone service and adults with landline phones, with one notable exception: wireless-only adults were more likely to report excellent or very good health than were those with landline phones, while adults without any service were less likely to report excellent or very good health.

When adjusted to account for demographic and socioeconomic differences between adults with landline phones and wireless-only adults, significant differences still were observed in the odds of several health-related behaviors and indicators of health care access (Table 3). Compared to adults with landline phones, wireless-only adults were still more likely to have had five or more alcoholic drinks on one occasion and to smoke, were more likely to be uninsured, were less likely to have a usual place for medical care, and were less likely to have received an influenza vaccination in the past 12 months.

IMPLICATIONS

Because the population without a landline phone is small—4.4% of adults have no phone or only a wireless phone—estimates for health and health care service use measures derived for adults with landline phones showed little bias from estimates for all adults (Table 4). As of 2003, with a survey sample of approximately 14,000 adults and a significance level set at .05, we estimated that the noncoverage of adults without landline phones in RDD surveys would result in only one significantly biased estimate (current uninsurance) out of 14 key health measures. It is tempting to conclude that the implications of wireless substitution for estimates from health-related RDD surveys are negligible. But RDD surveyors cannot assume that the state of affairs in early 2003 remains. Wireless substitution continues to grow. A survey of purchasers of wireless phones in the first quarter of 2003 found that 7% were replacing an existing landline phone or purchasing the wireless phone as their only phone (Schiela, 2003). A RoperASW study found that 9% of wireless subscribers were "almost certain" or "very likely" to use their wireless

phone for all calls in the next year (Tuckel & O'Neill, 2003), and a study by Ernst & Young and PriMetrica (2003) found that nearly half of all U.S. households would be willing to substitute wireless phones if the price and features were right. New FCC rules permitting landline customers to transfer their home number to a wireless phone are expected to increase these estimates by at least 10% (Standard & Poor's, 2003).

As wireless substitution continues to grow, so does the size of the population without landline telephones. Assuming that the wireless-only population continues to differ from the population with landline phones, RDD surveys soon will find it necessary to include wireless phones in their sampling frames. Adding wireless phones to RDD sampling frames generally has been considered inappropriate for household surveys because these phones most often are linked with individuals rather than households. Wireless users also may incur costs to receive calls, which is the reason for certain restrictions on cold calls to wireless phones contained in the Telephone Consumer Protection Act (47 U.S.C. 227). Still, surveyors should recognize that nearly two-thirds of adults without landline phones have wireless service. Including wireless-only adults in a dual-frame RDD sampling design (landline and wireless) would substantially reduce the noncoverage bias in RDD surveys. For example, when estimates for health and health care service use measures derived for adults with any phone were compared with those for all adults, differences for the 14 key health measures were all very small and not statistically significant (Table 3). Several researchers already are exploring the feasibility of conducting RDD surveys of wireless subscribers, and their promising results include evidence that response rates are consistent with landline RDD surveys (Steeh, 2003) and may be higher for wireless-only adults (Arbitron, 2003).

If inclusion of wireless-only adults in RDD surveys proves untenable, improved statistical adjustments to RDD survey sampling weights will be necessary to account for adults without landline phones. Current efforts to adjust the sampling weights for completed landline interviews to account for nontelephone households typically rely on one of three methods: (1) ratio adjustments to match U.S. Census Bureau estimates for the demographic distribution of the overall population, (2) adjustments to the weights of households with interruptions in landline service during the previous year to account for households without service at the time of the interview (Frankel et al., 2003), or (3) adjustments to the weights of households with landline phones based on logistic regression estimates of the propensity for each household to have been without service(Ferraro & Brick, 2001). All three methods incorrectly assume that the population without

Table 4. Prevalence Rates for Various Health Measures by Telephone Status: U.S., January–June 2003

Health Measure	Prevalence			95% Confidence Interval		
	All adults	Adults w/a landline	Adults w/any phone	All adults	Adults w/a landline	Adults w/any phone
Health-related behaviors						
5+ alcoholic drinks in 1 day at least once in past year	19.8	19.1	19.6	(18.8, 20.8)	(18.1, 20.1)	(18.6, 20.6)
Current smoking	21.6	20.7	21.1	(20.8, 22.5)	(19.8, 21.5)	(20.2, 22.0)
Regular leisure-time physical activity	33.2	33.3	33.3	(32.1, 34.3)	(32.2, 34.4)	(32.2, 34.4)
Health status						
Health excellent/very good	62.1	62.2	62.3	(61.2, 63.0)	(61.2, 63.1)	(61.4, 63.2)
Experienced serious psychological distress in past 30 days	3.2	3.0	3.1	(2.8, 3.5)	(2.7, 3.4)	(2.8, 3.4)
Asthma episode in the past year	3.5	3.5	3.5	(3.2, 3.9)	(3.1, 3.8)	(3.2, 3.9)
Diagnosed diabetes	6.2	6.3	6.2	(5.7, 6.6)	(5.8, 6.8)	(5.8, 6.7)
Obesity among adults age 20+	23.8	23.7	23.7	(22.9, 24.7)	(22.8, 24.7)	(22.8, 24.7)
Health care service use						
Has a usual place to go for medical care	85.5	86.6	86.1	(84.7, 86.4)	(85.9, 87.4)	(85.3, 86.9)
Influenza vaccine in past year	29.4	30.2	29.7	(28.5, 30.3)	(29.3, 31.1)	(28.8, 30.6)
Ever received pneumococcal vaccination	15.9	16.3	16.0	(15.2, 16.7)	(15.6, 17.1)	(15.3, 16.8)
Ever been tested for HIV	35.9	35.4	35.7	(34.8, 36.9)	(34.4, 36.5)	(34.7, 36.8)
Did not obtain needed medical care in past year due to financial barriers	6.1	5.7	5.9	(5.8, 6.5)	(5.4, 6.0)	(5.6, 6.2)
Uninsured	16.7	15.5	15.9	(16.0, 17.4)	(14.8, 16.1)	(15.3, 16.6)

landline telephones is relatively homogenous. This population, however, consists of both wireless-only persons and persons with no phone service; household incomes and less education and are more likely to be older and unemployed (Table 2).

Future efforts to use statistical adjustments to account for nontelephone households must account for these two populations separately, and the present data suggest that such adjustments will not be simple. Ratio adjustments that account for only key demographic characteristics will not be sufficient. Even after accounting statistically for demographic differences, wireless-only adults differed from adults with landline telephones in their likelihood to smoke, consume alcoholic beverages, be uninsured, and have no usual place for medical care (Table 3). It also is unlikely that data from adults with both landline phones and wireless service can be used to account for adults with only wireless service. Adults with both types of service tend to have higher household incomes and more education, and they are more likely to be older, have families, and own their homes (Table 2).

The ability to include wireless-only adults in RDD sampling frames or to sufficiently adjust sampling weights to account for their exclusion will not cure all of the ills of RDD surveys. RDD surveyors will continue to have difficulties achieving high response rates and navigating the new technologies used by potential respondents to avoid unwanted calls. It could be said, however, that the future of RDD surveys rests on the ability to identify and account for persons in the population who cannot be reached. For example, we found that the exclusion of adults who could not be reached because they have "cut the cord" did not have a demonstrable effect on estimates of health and health care service use. Yet the size and characteristics of this population are likely to change over time, and it will continue to be necessary to monitor these changes using high-quality household-based face-to-face surveys for the foreseeable future.

REFERENCES

Arbitron. (2003, November/December). Multiple research initiatives inform continuous improvement. *Arbitron Outlook*, 6.

Ernst & Young, & PriMetrica. (2003). *Wireline/wireless substitution study [Executive report].* Retrieved June 1, 2004, from the PriMetrica Web site: http://www.primetrica.com/products/wireline_wireless/pdf/wwss_exec_sum.pdf

Ferraro, D. & Brick, J. M. (2001). Weighting for nontelephone households in RDD surveys. *Proceedings of the Section on Survey Research Methods* [CD-ROM]. Alexandria, VA: American Statistical Association.

Frankel, M. R., Srinath, K. P., Hoaglin, D. C., Battaglia, M. P., Smith, P. J., Wright, R. A. et al. (2003). Adjustments for non-telephone bias in random-digit-dialling surveys. *Statistics in Medicine*, *22*, 1611–1626.

International Telecommunications Union. (2003). *Mobile subscribers, subscribers per 100 people: 2002.* Retrieved June 2, 2004, from http://www.itu.int/ ITU-D/ict/statistics/at_glance/cellular02.pdf

Ni, H., Coriaty-Nelson, Z., Schiller, J., Hao, C., Cohen, R.A., & Barnes, P. (2003, December). *Early release of selected estimates based on data from the January–June 2003 National Health Interview Survey.* Retrieved June 2, 2004, from the National Center for Health Statistics Web site: http://www.cdc.gov/nchs/about/major/nhis/released200312.htm

Schiela, J. (2003, June). *The right stuff: Putting together the best mix of products and services for your customers.* Paper presented at the Competitive Telecommunications Association Summer Showcase, San Francisco.

Standard & Poor's Equity Research Services. (2003, December). *More Americans plan to drop their home phone* [news release]. Retrieved June 1, 2004, from http://www.cellular-news.com/story/10302.shtml

Steeh, C. (2003, May). *Surveys using cellular telephones: A feasibility study.* Paper presented at the annual conference of the American Association for Public Opinion Research, Nashville.

Tuckel, P., & O'Neill, H. (2003, May). *Ownership and usage patterns of cell phones: 2000–2003.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Nashville.

Yankee Group. (2003a, April). *Yankee Group reports wireless subscribers use cellphones more than home phones* [news release]. Retrieved June 1, 2004, from the Reuters Web site: http://about.reuters.com/investors/events/pressreleases/index.asp?pressid=1195

Yankee Group, (2003b, August). *Twelve percent of U.S. young adults are totally wireless, according to the Yankee Group* [news release]. Retrieved June 1, 2004, from http://www.yankeegroup.com

FEATURE PAPER: Health Surveys in the 21st Century: Telephone vs. Web

Reg Baker and Dan Zahs, Market Strategies
George Popa, Medstat

INTRODUCTION

As the barriers to telephone survey research grow and multiply, researchers are looking more aggressively toward new methodologies. The twin forces of falling response rates and rising costs are making it increasingly difficult to produce high quality survey data at a reasonable cost. These trends probably are irreversible, and alternative data collection strategies will be needed. It is against this backdrop that we have seen Web surveys emerge as a serious alternative to telephone and other traditional methodologies.

Over the last five years, commercial research companies have enthusiastically embraced Web surveys across a wide spectrum of study types. U.S. market research firms expected that almost 25% of their total revenues would be tied to online research in 2003 ("Strong '02 Internet MR growth slows in '03," 2003). Researchers in the public sector, both academic and government, have been less enthusiastic. They see a variety of unanswered questions, especially in the areas of sample bias and potential mode effects (Couper, 2001).

THE WEB OPPORTUNITY

The World Wide Web is one of the great technical achievements of our time. Barely known outside of esoteric government and academic circles just ten years ago, it now plays a fundamental role in global communications, the conduct of commerce, and social interaction.

Survey researchers were quick to spot the Web's potential as a medium for communicating with respondents and collecting data. Early technical hurdles were quickly overcome, and it now is possible to administer even the most complex questionnaires over the Web. Online sample sources, once difficult to come by, have proliferated and now are readily available, although issues of bias remain. Perhaps best of all, Web-based data collection often is faster and less expensive than other methodologies, including telephone.

But Web-based data collection is not without its problems. Internet adoption by U.S. households progressed rapidly in the late 1990s but has slowed dramatically in the early years of this century. While there is quibbling over exact numbers, most agree that less than 70% of U.S. households have Internet access (Milla, 2003). Among that 70%, bias continues with persistent underrepresentation of the elderly and low SES groups, a phenomenon called "The Digital Divide" (Lenhart et al., 2003).

Limited penetration and use bias are significant problems for the survey researcher. They are made worse by problems in sampling and poorly understood mode effects. For example, there still is no good way to build a probability-based Web sample, no analogue to the national frames from which RDD samples are drawn. There are legitimate concerns about mode effects originating either from self-administration or impacts in visual presentation, many of which have been documented in the literature (e.g., Tourangeau, Couper, & Conrad, in press). On the other hand, there also are instances in which Web results have been shown to match those obtained by telephone, especially in the domain of political opinion polling (e.g., Chang & Krosnick, 2003).

OVERVIEW OF THE PULSE SURVEY

PULSE is the largest ongoing private health care survey in the United States.

The authors wish to thank Medstat for both funding this research and agreeing to let us share the results. We also acknowledge the assistance of Chris Montaglione, Roma Locke, and Heather McKnight, all from Market Strategies.

Funded by Medstat and conducted by telephone every year since 1988, PULSE interviews 100,000 respondents annually in a series of ten consecutive waves. Over the course of its 16-year history, PULSE has examined over 80 health care topics. Relying on an RDD sample, PULSE is designed to represent U.S. households by Designated Market Area (DMA).

PULSE has suffered from the same forces afflicting RDD telephone research generally. Refusal rates and sample bias have increased. Contact rates and response rates have declined. Costs have risen. Over the last three to four years, these impacts have been especially severe, and response rates have fallen precipitously.

In 2002, Medstat and Market Strategies, Inc. (MSI) began an evaluation of the feasibility of migrating a substantial number of PULSE interviews (20%–50%) from telephone to Web. MSI has extensive Web experience, and we were confident in our ability to execute a study of this scale. We believed that moving such a large number of cases to the Web would reduce overall costs by as much as 25% to 30%. However, due to concerns about the potential for significant response bias, we designed two tests to detect it. If we found that Web administration introduced bias, we planned to investigate methods for correcting it.

THE TESTS

In 2002 and early 2003, we conducted two field tests. The goal of these tests was to address Medstat's concerns in five primary areas:

(1) Demographic bias, which might produce a sample with unacceptable demographic characteristics.
(2) Response bias, which might produce significantly different measures of health-related behaviors.
(3) Geographic bias, which might disturb the desired distribution of sample by DMA.
(4) Methodological issues, which might raise concerns in the marketplace where Medstat sells data products based on the PULSE data.
(5) The availability of geocodes for Web respondents so that the normal practice of enhancing the data with ecological information (Prizm clusters) could be maintained.

The 2002 Field Test

We conducted our first test in October of 2002. We adapted the Wave 8 telephone questionnaire to the Web and interviewed approximately 2,000 respondents concurrently with the Wave 8 telephone survey. Survey Sampling, Inc. (SSI) provided the sample from their Survey Spot Panel, an online opt-in panel with over one million members. Results were mixed:

- Compared to telephone, the Web produced a younger and more affluent sample.
- Web respondents reported being less healthy, despite being younger and of higher SES.
- Web respondents reported smoking more, had a higher incidence of some chronic conditions, and were more often without health insurance.
- There was a modest geographic bias, with the Web sample producing more completed interviews in the southwestern and southeastern U.S. than did the telephone sample.
- We were able to obtain sufficient address information from Web respondents to successfully geocode at levels comparable to telephone.

Some of the problems uncovered were manageable. We felt we could correct the demographic bias by either drawing a more balanced sample from the panel or with poststratification. We also could correct the geographic imbalance in the sample pull. However, the differences in some key health behavior items—the heart of the survey—were deeply troubling. We developed three hypotheses to explain what we found:

(1) Web users are behaviorally different from nonusers.
(2) Web panel members are different from nonusers and from Web users generally.
(3) A mode effect causes Web respondents to report differently than telephone respondents. This might be due to question presentation issues (i.e., seen on the screen rather than read over the telephone) or social desirability.

The 2003 Test

We designed the 2003 test into the first wave of data collection for that year. The design called for us to interview 5,000 respondents by telephone and 5,000 by Web. Because we were concerned that the telephone/Web differences in the first test might be due to some peculiar features of the SSI sample, we added a second Web sample source by drawing on the membership of AOL, which we thought might produce a broader representation of Internet users. SSI randomly drew sufficient sample from the Survey Spot panel to produce 2,500 completed interviews. AOL does not maintain a panel but rather uses "river sampling" techniques. Survey invitations are posted at their Opinion Place site and other sites on the service. Respondents who click through to participate in a survey are screened and, based on screener outcomes, passed through to a waiting questionnaire. We asked AOL to deliver 2,500 respondents using this technique. We made no attempt to control the sample selections by demographic characteristics from either of the Web sample sources.

We also added questions on Internet use, some social isolation measures, and a question about educational attainment. We hoped that these additional questions might help explain the differences observed in the first test and perhaps serve as the basis for postsurvey adjustment.

We fielded the study in late January and early February of 2003. We completed a total of 10,932 interviews, distributed as follows:

- 2,874 RDD respondents
- 2,126 PULSE panel respondents[1]
- 2,270 SSI respondents
- 2,562 AOL respondents

As in the earlier test, there were substantial differences between the demographic distributions for the Web completes and the telephone completes, as shown in Table 1. The Web sample reported higher levels of education and higher income and had fewer respondents age 65 or older. Within the RDD sample, we saw differences between Web users and nonusers that in many ways mirrored the differences between RDD and Web generally. There were few meaningful demographic differences between the two Web samples.

Table 1. Sample Demographics

	African American	Some College or Above	Household Income >$25,000	Age 65+
RDD Total	**7%**	**66%**	**70%**	**23%**
Web users	4%	80%	84%	11%
Nonusers	9%	49%	52%	38%
Web Total	**5%**	**76%**	**78%**	**7%**
SSI	4%	78%	77%	6%
AOL	7%	74%	78%	7%

Given the different distributions between RDD and Web respondents and because health behaviors often are influenced by age and socioeconomic status, we elected to poststratify all three samples (RDD, SSI, and AOL) by weighting them to match the U.S. population on age, gender, and income. This effectively eliminated the age and income bias, although some educational bias remained.

[1] A portion of each year's PULSE sample is comprised of respondents interviewed in previous years. The results reported here are for RDD only.

RESULTS

Our analytical plan focused on the key behavioral measures collected in the Wave 1 survey:

- Self-assessed health status
- Smoking behavior
- Drinking behavior
- Weight loss
- Health insurance coverage

One of the most disturbing findings in the 2002 test was the lower proportion of Web respondents reporting their health as excellent. This finding was replicated in the 2003 test, with only about half as many Web as RDD respondents reporting excellent health. There were no mode differences for those reporting their health as "fair" or "poor," but there were significant differences for all other categories (Table 2). While there were some differences between the two Web sample sources, they were nowhere near as dramatic as those between modes of interview.

Smoking behavior differences observed in the first test also were replicated. While 40% of Web respondents reported that someone in their household had smoked cigarettes regularly during the past 12 months, only 33% of telephone respondents did so. As with health status, the two Web panels were virtually identical on reported levels of smoking.

There also were significant differences in weight loss attempts: 71% of Web respondents reported having tried to lose weight in the last year versus 59% for RDD respondents. Further, the reasons respondents gave for trying to lose weight differed. RDD respondents were more likely to cite health-related reasons, while Web respondents chose reasons related to appearance and family pressures.

We found differences in health insurance coverage as well. Just 12% of RDD respondents reported that at least one person in their household was not covered by health insurance. Twenty percent of Web respondents reported an uninsured person.

The only important behavioral variable for which we did not find significant differences across modes was drinking behavior. The number of occasions on which respondents reported drinking and the number of drinks consumed were statistically identical.

DISCUSSION

We note at the outset of this discussion that the results reported above have been replicated in other settings. As part of its evaluation of a Web-based strategy, Medstat arranged for a parallel test by another research company using a different Web sample panel. Those results were essentially the same as those obtained by MSI. A Medstat competitor also designed and tested Web-based data collection, with similar results.

In the survey literature, Schonlau and his colleagues (2004) reported similar findings in a study of health issues in California. They compared the results of an RDD survey to those obtained from a parallel Web survey conducted with Harris Interactive. In an attempt to adjust for differences between RDD and Web collection, Harris Interactive performed a poststratification adjustment using propensity scores, a proprietary

Table 2. Health Behavior Comparisons, by Sample

	RDD	Web Total	Web SSI	Web AOL
Health Status				
Excellent	27%	15%	12%	17%
Very good	31%	37%	38%	35%
Good	28%	32%	32%	32%
Fair	11%	12%	13%	12%
Poor	4%	4%	5%	4%
Smoking				
Cigarette smoking	33%	40%	39%	40%
Attempted to lose weight				
Yes	59%	71%	69%	74%
Health insurance				
Not covered	12%	20%	20%	19%
Drinking occasions in past month	4.3	4.1		

Note: Based on weighted data.

Table 3. Health Status

Health Status	Schonlau RDD	Schonlau Web	PULSE RDD	PULSE Web
Excellent	23%	13%	27%	15%
Very Good	33%	40%	31%	37%
Good	27%	33%	28%	32%
Fair	14%	12%	11%	12%
Poor	3%	3%	4%	4%

Source: Schonlau et al., 2004.

technique the company frequently uses in its Web survey research. Even after this adjustment, 29 of 37 items were statistically different.

A comparison of Schonlau's results with those of PULSE (Table 3) is especially compelling. Bearing in mind that the former is for California only and the PULSE results are for the U.S. as a whole, it is nonetheless intriguing that both studies have Web respondents reporting excellent health at about half the rate of parallel RDD studies.

One potential explanation for these differences is the earlier hypothesis that Web users are behaviorally different from nonusers. From previous telephone-only PULSE data collections, we know that Web users typically report better health and fewer risk behaviors than nonusers. For example, Web users in the RDD sample reported their health status as excellent about twice as often as nonusers. Twenty-nine percent of Web users in the RDD sample reported that someone in their household smoked on a regular basis during the previous year, compared to 38% of nonusers.

The easy explanation for these differences would appear to be the demographic bias in Web use. The generally younger and higher SES Web users are likely to be healthier and smoke less than the older, lower SES non-Web users. This is largely borne out in the data. However, respondents to the Web survey, while similar demographically to Web users in the telephone sample, are not at all similar in terms of health status and behaviors. In fact, they more closely resemble the non-Web users in the telephone sample. This is a mystery we have yet to solve.

Our first attempt at explanation entailed a search for some underlying behaviors that might explain the different reporting of Web users across modes. Two avenues seemed promising: time spent online and social isolation.

We found significant differences in time spent online across modes, with Web users in the RDD sample generally being more casual users than our Web respondents. Almost half of the RDD Web users reported spending fewer than five hours per week online, compared to just 17% of Web respondents. At the other end of the spectrum, almost 60% of Web respondents reported 11 or more hours of Web use per week, compared to 30% of Web users in the RDD sample. Despite these aggregate differences, time spent online was a poor predictor of health status and health behavior.

We also included a battery of eight questions designed to measure social isolation. These questions asked respondents whether they found themselves spending more time or less time in various social activities—shopping, spending time with family and friends, attending entertainment events, etc.—compared to five years ago. There were almost no significant differences between RDD Web users and Web respondents on these questions.

In the end, we found ourselves with two possible explanations for the differences observed: mode effects and/or bias in the Web panel.

Unfortunately, the study as currently designed makes it difficult to detect mode effects. We note, however, that the reports by telephone respondents on health status and smoking compare much more favorably with other surveys (such as the National Health Interview Survey) than do those by Web respondents. We also note the strong comparability of our results to those of the studies referenced at the beginning of this section. Our suspicion, and it is no more than that at this stage, is that the differences observed are primarily due to the use of volunteer samples rather than probability samples for the Web portion of the study. We

think it likely that the self-selection bias inherent in Web panels is the primary cause of the observed differences, but that is still speculation.

CONCLUSION

Despite the problems encountered in this study, we continue to believe that Web-based surveys hold substantial promise for high quality data collections, especially given the problems increasingly besetting more traditional methods. It seems clear to us that additional research is required if we are to leverage the considerable benefits of the Web in speed, cost, and survey complexity for health surveys such as PULSE.

Further research into techniques for using nonprobability samples is arguably the most pressing need. It is difficult to imagine a time in the near future when we will have a reliable frame from which to draw true probability Web samples. Even if we could, the enduring behavioral differences we believe exist between Web users and nonusers make it difficult to use such samples in ways that are representative of the general population. More sophisticated propensity models and calibration techniques are needed.

Mode effects is the second area where research is needed, and fortunately, there is a great deal of promising work already underway there. As more of that research begins to find its way into publication, we can expect to gain considerable clarity about the best designs to control and account for bias in Web surveys.

In the meantime, the Web remains something of a niche survey methodology. Surveys of online communities (such as college students), mixed-mode surveys—especially when longitudinal, surveys of businesses, and methodological research are all areas where the Web already plays a significant role. It may also be that for some domains, Web panels can work, although it seems clear from our experience that health behavior is not one of them. Overall, there remains much work to do.

REFERENCES

Chang, L., & Krosnick, J. A. (2003). *National surveys via RDD telephone interviewing vs. the Internet: Comparing sample representativeness and response quality.* Columbus, OH: The Ohio State University.

Couper, M. P. (2001). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64,* 464–494.

Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M. et al. (2003). *The ever-shifting Internet population: A new look at Internet access and the Digital Divide.* Washington, DC: Pew Internet and American Life Project. Retrieved June 3, 2004, from http://207.21.232.103/pdfs/PIP_Shifting_Net_Pop_Report.pdf

Milla, P. (2003). *Overview and introductions.* Paper presented at The CASRO Technology Conference, New York.

Schonlau, M., Zapert, K., Payne Simon, L., Sanstad, K., Marcus, S., Adams, J. et al. (2004). A comparison between a propensity weighted Web survey and an identical RDD survey. *Social Science Computer Review*, *22,* 128–138.

Strong '02 Internet MR growth slows in '03. (2003). *Inside Research, 14*(1), 1–4.

Tourangeau, R., Couper, M. P., & Conrad, F. (in press). The impact of the visible: Images, spacing, and other visual cues in Web surveys. *Public Opinion Quarterly*.

FEATURE PAPER: Are Web and Mail Modes Feasible Options for the Behavioral Risk Factor Surveillance System?

Michael W. Link, RTI International
Ali Mokdad, Centers for Disease Control and Prevention

Response rates in random-digit-dial (RDD) surveys have been declining for at least the past decade (de Leeuw & de Heer, 2002). When combined with differences in attitudes, behaviors, and beliefs between respondents and nonrespondents, nonresponse threatens the validity and reliability of data reported in probability sample surveys (Babbie, 1990; Dillman, Eltinge, Groves, & Little, 2002). The use of multiple modes of questionnaire administration is one possible means of addressing this problem, particularly utilizing Web and mail surveys as complements to telephone data collection (Dillman, 2000).

Mixing survey modes provides the potential for extending the reach of a survey, encouraging participation across a broader mix of the population. Research has shown that some sample members prefer and respond more readily to different modes (Groves & Kahn, 1979). In this respect, mixed-mode approaches have the potential for increasing response rates and, presumably, the validity and reliability of the data collected.

There are potential drawbacks, however. Different modes have been shown to produce different results even when questions are asked of the same sample members (de Leeuw, 1992; Dillman, 2000; Dillman, Sangster, Tanari, & Rockwood, 1996). Therefore, use of alternative modes may increase response rates but also may also increase measurement differences. Moreover, while there is a relatively large body of literature examining combinations of telephone, mail, and face-to-face surveys, our understanding of how Web surveys fit into this mix is quite limited.

As one of the largest ongoing state-based RDD telephone surveys, the Behavioral Risk Factor Surveillance System (BRFSS) is confronted with declining response rates and questions regarding data reliability and validity. Use of multiple modes is one possible means of addressing this problem. Conducted in all 50 states, as well as the District of Columbia, Puerto Rico, Guam, and the Virgin Islands (for simplicity, hereafter referred to collectively as "states"), 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. For BRFSS and surveys of similar design to continue to meet the public health data needs of local, state, and national researchers and policy makers, alternative data collection means need to be examined.

To this end, a set of experiments was conducted in four states to test the effectiveness of Web and mail surveys when used in conjunction with telephone follow-up of nonrespondents as a means of increasing BRFSS survey participation. The research addresses several key questions: can alternative modes help to increase BRFSS response rates? How does use of multiple modes impact participation among different subgroups of the population? What effect, if any, does combining modes have on the resulting survey estimates?

METHODS

This study involved two sets of experiments in four states (Arkansas, Indiana, New York, North Dakota) over a two-month period (October and November 2003). In the first set of experiments, sample members were asked to complete the BRFSS questionnaire via the Internet. In the second, the questionnaire was mailed to sample members. In both experiments, nonrespondents to the self-administered modes were followed up by telephone interviewers to complete the survey. The experiments were conducted in parallel with the regular monthly BRFSS data collection in each of the states, providing a baseline for comparison.

Because a mailing was required for both experiments, only address-matched sample was used. Separate state-specific samples were drawn for the two experiments following previously approved and monitored CDC BRFSS RDD sampling protocols. Each sample was address-matched by cross-referencing the telephone number with a database of known addresses. Cases without a valid mailing address were removed from the sample.

Cases for which a self-administered questionnaire was not completed by the 10th day of the month were loaded into CATI for telephone follow-up. In the mail experiment, all nonrespondents were followed up by telephone. In the Web experiment, a subsampling strategy was used for selecting cases for CATI follow-up to minimize costs. The initial Web experiment sample (N=9,629) was much larger than that of the mail experiment (N=2,406) to make up for the fact that not all households have access to the Internet and in anticipation of a lower Web response rate. Weighting adjustments were used in the analysis, when appropriate, to account for the subsampling.

To increase comparability across modes, only the core BRFSS component (the set of questions asked identically in all states) was used with the experiments. In developing the Web and mail versions of the BRFSS, we followed as close as possible the "unimode construction" approach outlined by Dillman—that is, "the writing and presenting of questions to respondents in a way that assures receipt by respondents of a common mental stimulus, regardless of survey mode" (2000, p. 232). Although designed as a CATI survey, no wording changes were required for the text of the 84-item questionnaire for the Web and mail modes.

There were, however, some differences in how the response options were handled. For the self-administered modes, "don't know" was made an option for six items that typically receive 10% or more "don't know" responses during regular monthly CATI data collection. "Refused" was not offered in either self-administered mode. Instead, it was assumed that respondents to the mail survey would simply leave a question blank, while Web respondents were provided with a "continue" button on each screen, allowing them to move to the next screen without entering a substantive response.

Households in these experiments were initially notified about the study via mail, with letters printed on state department of health stationary and signed by a state public health official. For Web survey households, a unique username and password were provided along with the Internet address (or URL) for accessing the Web site. Households in the mail experiment received a packet containing a cover letter, a copy of the questionnaire booklet, and a postage-paid return envelope. Initial mailings were sent three days before the start of data collection, with a reminder letter sent to all respondents on the 5th of the month.

For all telephone interviewing, a respondent was selected randomly based on the total number of males and females age 18 or older in the household. For the mail and Web surveys, households were asked to select an adult to complete the questionnaire.

For telephone contacts, the standard BRFSS 15-call protocol was followed for the baseline data collection, while the number of calls was reduced to 10 for the nonresponse follow-up. With this exception, all other BRFSS data collection protocols were in place for the nonresponse telephone follow-ups.

FINDINGS

Response Rates

One of the central questions addressed by this inquiry is if the use of alternative modes (Web and mail in particular) in conjunction with telephone follow-up of nonrespondents produces higher response rates than are obtained using telephone interviewing only. Because the experimental modes were limited to use of address-matched sample, we use only address-matched cases from the baseline data collection in our initial comparisons. Response rates were calculated following conventions established by the American Association for Public Opinion Research in response rate option #4 (AAPOR, 2000).

Table 1. Response Rates Among Address-Matched Sample Only, by State and Mode

State	BASELINE	WEB SURVEY EXPERIMENT			MAIL SURVEY EXPERIMENT		
	CATI only	Web only	CATI follow-up	Overall Web + CATI follow-up	Mail only	CATI follow-up	Overall mail + CATI follow-up
Mean	**40.1%[2,3]**	**15.4%**	**32.5%**	**47.9%[3]**	**43.6%**	**16.4%**	**60.0%**
	(1,378)	**(1,905)**	**(1,905)**	**(1,905)**	**(501)**	**(501)**	**(501)**
Arkansas	41.4%[2,3]	13.7%	34.0%	47.7%[3]	37.8%	21.8%	59.6%
	(1,314)	(2,139)	(2,139)	(2,139)	(473)	(473)	(473)
Indiana	38.3%[2,3]	15.7%	32.3%	48.0%[3]	43.3%	15.7%	59.0%
	(1,518)	(1,661)	(1,661)	(1,661)	(528)	(528)	(528)
New York	31.3%[2,3]	13.1%	24.7%	37.8%[3]	39.7%	12.6%	52.3%
	(1,833)	(2,102)	(2,102)	(2,102)	(539)	(539)	(539)
North Dakota	49.5%[2,3]	19.2%	38.8%	58.0%[3]	53.6%	15.3%	68.9%
	(846)	(1,720)	(1,720)	(1,720)	(464)	(464)	(464)

Note: Significance based on Chi-square test. Superscripts indicate significance ($p < .05$) with that cell and (2) the overall Web survey experiment results or (3) the overall mail survey experiment results. Estimated eligible households shown in parentheses. Data for the Web experiment CATI follow-up are weighted to adjust for subsampling of Web nonrespondents.

Table 2. Overall Response Rate Estimates, by State and Mode

State	BASELINE (CATI ONLY)		WEB SURVEY EXPERIMENT		MAIL SURVEY EXPERIMENT	
	Est. Eligible Sample (*N*)	Response Rate (%)	Est. Eligible Sample (*N*)	Response Rate (%)	Est. Eligible Sample (*N*)	Response Rate (%)
Mean	**1,673**	**48.8[2,3]**	**1,686**	**53.9[3]**	**1,734**	**61.9**
Arkansas	1,493	50.4[2,3]	1,507	53.9[3]	1,592	62.7
Indiana	1,883	47.1[2,3]	1,873	53.6[3]	1,940	60.4
New York	2,367	39.2[2,3]	2,446	43.4[3]	2,635	52.4
North Dakota	950	58.6[2,3]	920	64.5[3]	967	72.1

Note: Significance based on Chi-square test. Superscripts indicate significance ($p < .05$) with that cell and (2) the overall Web survey experiment results or (3) the overall mail survey experiment results. Data for the Web experiment CATI follow-up are weighted to adjust for subsampling of Web nonrespondents.

As Table 1 shows, among the address-matched sample, the response rates for the baseline telephone-only data collection ranged from 31.3% (New York) to 49.5% (North Dakota), averaging 40.1% across the four states. In comparison, both the Web survey and mail survey with CATI follow-up posted significantly higher response rates. The former averaged nearly an eight percentage point higher rate across the four states, while the latter showed nearly a 20% increase.

Looking closer at the Web experiment, on average 15.4% of respondents chose to complete the survey on the Web across the four states. Use of the Web survey was highest in North Dakota (19.2%) and lowest in New York (13.1%).

Response to the mail survey experiment was interesting in that the percentage of completed interviews from the mail survey itself was greater in three of the four states (Arkansas being the exception) than those obtained in the baseline phone interview. On average, the response rate for the mail survey was 43.6%, compared to 40.1% obtained in the CATI baseline. Response to the mail survey was highest in North Dakota (53.6%) and lowest in Arkansas (37.8%).

While the Web and mail experiments used address-matched sample only, we can simulate the overall effects on response rates by including nonaddress-matched cases based on the results obtained during the baseline data collection. If we assume that the final disposition of the nonaddress-matched cases in the baseline would not have been significantly different if these cases had been called as part of the Web or mail experiments, then we can

use these data to help simulate the expected overall response rate outcomes for the Web and mail experiments.

As shown in Table 2, the inclusion of nonaddress-matched households reduces the difference in response rates across the three groups in each state. A weighting adjustment was used to ensure that the address-matched cases in the Web and mail experiments contributed proportionately the same to the final response rate as the address-matched cases in the baseline data.

In each case, the Web and mail mixed-mode approaches produced higher response rates, but the impact of the alternative approaches was tempered. On average, the Web with CATI follow-up produced a five percentage point increase in response rates compared to the baseline figures, while the mail with CATI follow-up produced a thirteen percentage point increase.

Respondent Demographics

To determine if the increased response rates for the mixed-mode approaches led to a different mix of respondents, we compared selected demographic characteristics of address-matched respondents across the baseline and two experimental groups. We also compared the demographics of these groups to population estimates obtained by the U.S. Census Bureau in the 2002 American Community Survey (ACS).

Baseline respondents differed significantly from those in the Web and mail experiments in terms of sex, race, and age. A higher percentage of females were interviewed in the Web experiment (61.9%) and mail experiment (64.3%) than in the CATI baseline (58.1%). In terms of race, the proportion of respondents that were non-Hispanic White was higher for the Web (90.6%) and mail (90.1%) experiments than was the case with the CATI-only approach (86.4%). Additionally, significant age differences were noted across the three groups. Baseline respondents tended to be somewhat younger than those in the Web and mail experiments, with a higher percentage of 18–34-year-olds and a lower percentage of those age 65 or older.

In comparison to the ACS population characteristics, both the Web and mail with CATI follow-up approaches significantly overrepresented women, White non-Hispanics, and older individuals. These groups are traditionally overrepresented in RDD surveys as well. The net effect, thus, is that rather than help close the gap with some of these underrepresented groups (e.g., males, non-Whites, younger individuals), these particular mixed-mode designs actually exacerbated the problem of overrepresentation of these subgroups.

ANALYSIS OF SURVEY ESTIMATES

Perhaps the most important question to be addressed by these experiments is if moving from a CATI-only approach to a multimode approach involving a combination of self-administered and interviewer-administered modes influences the estimates obtained.

Bivariate comparisons of responses to fourteen key BRFSS health and behavior items showed significant differences between the baseline phone survey and the Web survey with CATI follow-up for nine of the items. The percentage point difference for these items ranged from 2.0% (“health care coverage”) to 5.7% (“trying to lose weight”). The Web multimode approach resulted in a higher percentage of “yes” responses for seven of the nine significant items and a significantly lower percentage of “yes” responses when asked the “HIV testing” and “condom use” questions.

Comparing the mail with CATI follow-up respondents to those interviewed by telephone only, we find a very similar pattern. There were significant differences in estimates for eight of the items. The range in differences between the mail experiment and baseline were somewhat higher than those seen in the comparisons between baseline and Web, ranging from 3.0% (“health care coverage”) to 7.9% (“joint pain”). Again, the direction of these differences was identical to the Web-baseline differences, as the mail with CATI follow-up produced higher estimates for six of the eight items and lower estimates for the “HIV testing” and “condom use” questions.

Interestingly, there were no significant differences in the estimates obtained for these items when we compare the overall responses to the Web and mail experiment. Despite the fact that the mail with CATI follow-up produced significantly higher response rates in all four states than did the Web with CATI follow-up and that significant differences were noted in terms of sex, age, and income, these differences did not appear to result in significantly different estimates for the questions examined.

DISCUSSION & CONCLUSION

These mixed-mode experiments show that Web surveys and mail surveys with telephone follow-up of nonrespondents are both possible alternatives to the current CATI-only approach for significantly increasing BRFSS response rates. Moreover, the mail with CATI follow-up approach produced rates significantly higher than either of the other two. In both sets of experiments, response rates were increased significantly in each of the four states examined.

However, the 5% increase in the expected overall response rate for the Web mixed-mode approach needs to be viewed with caution. Recent experiments examining the use of advance letters with the BRFSS population showed that prenotification alone can increase response rates by approximately six percentage points (Link, Mokdad, Town, Weiner, & Roe, 2003). Thus, it may be that the increase in response rates obtained in the Web experiments was due more to letters being sent to address-matched cases than to giving sample members an alternative mode. Also, the contribution of Web-completed questionnaires to the final Web experiment response rate was smaller than the contribution of returned mail questionnaires to the final mail experiment response rate. Web survey respondents made up just one-third of the Web experiment completions (15.4% of the 47.9% total), compared to mail survey respondents, who accounted for nearly three-quarters of the completed interviews in that experiment (43.6% of the 59.0% total). While both mixed-mode approaches increased response rates, the increases were not even across demographic groups. Respondents to the Web and the mail surveys differed significantly from those interviewed via telephone only on a number of important demographics, with sex, race, and age being perhaps the most significant. Telephone surveys typically overrepresent women in terms of completes, given that men are in general more difficult to contact at home and tend to be more reluctant to take part in surveys. One thought going into these experiments was that given the option to participate in the survey in a self-administered manner at the time of their choosing might encourage greater participation by men in the survey, but this was not the case. Future experiments might test alternatives for random selection of a respondent within the household for the self-administered component (rather than the self-selection approach used here) to attempt to correct for this male/female imbalance.

Both self-administered modes also were completed by a higher percentage of non-Hispanic Whites than was the case with telephone only. RDD surveys tend to overrepresent non-Hispanic Whites and underrepresent other races. Unfortunately, both the Web and mail surveys only increased this discrepancy.

The age of self-administered respondents differed significantly from those completing telephone only, in that they tended to be older. There were, however, clearly different groups responding to the Web and mail surveys. Web respondents were more likely to be middle-aged (35–54) than were those in either the mail or phone-only surveys. In contrast, a higher percentage of mail respondents were 65 or older than was the case with the other modes. This could be an encouraging finding, given the aging of the U.S. population. With the percentage of elderly increasing as the baby boom generation ages, mail surveys used in conjunction with telephone follow-up could increasingly become a viable means of collecting reliable and valid data from the general population.

Given that participation rates increased among some demographic groups more than others, it is not surprising that the unweighted

estimates obtained by these two approaches varied significantly from those obtained in the baseline. Interestingly, there were no significant differences in the overall estimates obtained by the two mixed-mode approaches when compared with each other, although there were differences in the demographic characteristics of these two groups. The question of interest is if the consistency of the results obtained using the multimode approach are reflective of more valid and reliable estimates or simply happenstance. On one hand, the level of nonresponse was reduced with the mixed-mode approaches, which should increase our confidence in the estimates. On the other, both approaches increased the differences (in terms of some basic demographic characteristics) between respondents and nonrespondents, thereby increasing the likelihood of bias in these estimates. Additional data analyses are planned to examine these aspects.

In conclusion, a great deal of work is required before sound recommendations can be made for moving an ongoing telephone-based surveillance like the BRFSS to a multiple mode approach. First, the findings highlight the need for a more complete test of alternative modes. Such a test should involve a sample more comparable to a typical BRFSS sample in that both address-matched and nonaddress-matched cases are included. Use of multiple modes also might be enhanced if used in conjunction with reliable, alternative sampling frames, thereby augmenting or even moving away from reliance on the current RDD sampling frame. Next, the Web and mail survey instruments should undergo more formal usability testing to optimize their formatting, layout, and navigation. There also is a need for a more detailed analysis of the data obtained through these various modes to ensure that differences introduced through the use of alternative (particularly self-administered) modes are the result of improvements in the validity and reliability of the data and not due to increased survey bias.

Perhaps most important is the need to recognize that issues involving the use of new methodologies and technologies (e.g., the Internet) need to be studied on an ongoing basis. Technology and communications are changing so rapidly that the research findings of today may not be relevant for long. Health survey methodologies are constantly evolving, and it is incumbent on researchers and methodologists to stay abreast of these changes, embracing through a rigorous testing and validation process the new technologies and approaches that improve the quality of the data and estimates they produce.

REFERENCES

American Association for Public Opinion Research. (2000). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. Ann Arbor, MI: Author.

Babbie, E. (1990). *Survey research methods*. Belmont, CA: Wadsworth.

de Leeuw, E. D. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT Publications.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. Little (Eds.), *Survey nonresponse.* New York: John Wiley & Sons.

Dillman, D. (2000). *Mail and Internet surveys: The tailored design method*. New York: John Wiley & Sons.

Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. (2002). Survey nonresponse in design, data collection, and analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. Little (Eds.), *Survey nonresponse.* New York: John Wiley & Sons.

Dillman, D. A., Sangster, R. L., Tanari, J., & Rockwood, T. (1996). Understanding differences in people's answers to telephone and mail surveys. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research: New directions for evaluation, No. 70* (pp. 45–62). San Francisco: Jossey-Bass.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews*. New York: Academic.

Link, M. W., Mokdad, A., Town, M., Weiner, J., & Roe, D. (2003). *Improving response rates for the BRFSS: Use of lead letters and answering machine messages*. Paper presented at the annual meeting of the American Association for Public Opinion Research, Nashville.

FEATURE PAPER: Don't Forget About Personal Interviewing

Patricia M. Gallagher and Floyd J. Fowler, Jr.
University of Massachusetts Boston

At the first Conference on Health Survey Research Methods, held at Airlie House in 1975, one session was devoted to the topic of whether statistical data could be collected by telephone. At that time, virtually all general population surveys conducted by the federal government or by academic organizations relied on area probability sampling and interviews conducted in person in people's households.

Although many caveats were mentioned, the participants certainly felt the telephone could be used for collecting data from people sampled from lists and for re-interviews, and the notion that general population data might be collected that would be credible using random-digit dialing was a possibility that was explored by some participants. Subsequent papers by Waksberg (1978), which demonstrated how to improve the efficiency of random-digit dialing, and by Klecka and Tuchfarber (1978), which demonstrated the similarity of data collected by random-digit dialing and in-person interviews, were critical to moving us to our current point where random-digit dialing has been a default option for most general population surveys. Meanwhile, in-person surveys of households declined markedly. There are several very large and extremely important surveys that are carried out using in-person interviews, such as the Current Population Survey (CPS), the National Crime Survey (NCS), the National Health Interview Survey (NHIS), the Medicare Current Beneficiary Survey (MCBS), and the National Survey of Drug Use and Health (NSDUH). The number of people interviewed in person is actually very large, but the number of different surveys using this methodology is comparatively small.

As we consider various options for how to routinely do general population surveys, telephone, mail, and the Internet are obvious candidates to play some role. However, as we are considering options, we thought it was important to also call attention to the potential of the old standard, the in-person household interview, as a potential player.

To do this, we thought one approach would be to remind participants of a design that was not uncommon in the 1970s and 1980s, when there was concern about the adequacy of sample frames for telephone surveys. The design we have in mind is based on an area probability sample. At the time of listing, interviewers collected names of those in multi-unit structures whenever they could. Reverse directories were used to identify telephone numbers associated with sampled housing units to the extent that it was possible. As many interviews as possible were conducted by telephone in order to realize the cost-saving advantages of telephone surveys. Then, for the remainder of housing units selected, in-person interviewers were used. A variation was that a mail survey could be used as a final step or for telephone nonrespondents prior to sending in-person interviewers to a household.

This is the methodology that was used by Hochstim (1967) in his oft-referenced comparison of modes of data collection. Thornberry (1976) used a parallel design in a health survey in Rhode Island, and Fowler and Mangione (1982) used such a design to study crime and fear in Hartford. Mangione, Hingson, and Barret (1982) used a similar design to study drinking behavior in a sample of greater Boston residents. In the later two studies, there was no mail component that followed the telephone phase; the designs both went directly from telephone to in-person interviews.

Table 1 summarizes the response rate experience for these various studies. The percentage of housing units that in fact could be interviewed by telephone varied markedly from study to study. Hochstim completed interviews by telephone with almost 80% of those who eventually responded. In contrast,

in the two studies that did not use mail as a second mode of data collection and that were done with central-city samples, only about 60% of the final respondents were interviewed by the initial telephone mode. It is worth noting, however, that the final response rates obtained in all these surveys ranged from quite acceptable (by 2004 standards) to terrific. Also, particularly for the two surveys that involved designated respondents in households and the typically more difficult population of central-city residents, the use of the in-person data collection made a large difference in moving the response rate from unacceptable to definitely acceptable.

We have a more recent example with another difficult-to-survey population: those on Medicaid. In a recent survey of Medicaid beneficiaries in Massachusetts about their experiences in getting health care, the CAHPS® protocol is to use mail surveys first, with telephone follow-up of mail nonrespondents. The sample frame is a list of those enrolled in Medicaid.

The experience throughout the country with CAHPS surveys of those on Medicaid has been that the response rates after the mail and phone phases have averaged in the 40% to 50% range (2003 range: 18% to 61%). In 2003, for more than half of all entities (states and individual health plans) reporting to the National CAHPS Benchmarking Database, the highest response rate for any of the participating health plans was under 40%. In a methodological experiment, we sent personal interviewers out to the addresses of those who had not responded by mail or by phone to try to complete an interview. The result was that we obtained responses from another 25% of the population, moving the overall response rate up to 72% (Table 2). Moreover, 12% of the remaining 28% essentially could not be found. Thus, we actually got returns from about 82% of those who could be found. In addition, administrative records for the total sample were compared with both the cumulative results for respondents from each phase of data collection and with those who never responded. Raising response rates is desirable in itself because it increases the credibility of data. However, the most important result is that improving response rates also reduced nonresponse bias. In each of the areas that we could assess using administrative data, we can demonstrate that the final group of respondents looked more like the total sample than the group of respondents that would have resulted based on either mail or mail-plus-telephone-returns alone. In most cases, the final group of respondents was virtually identical to the total sample in the ways that we could assess.

Table 1. Four Examples of Multimode Surveys

	Hochstim (1967)	Thornberry (1976)	Fowler & Mangione (1982)	Mangione et al. (1982)
Location	Alameda County	Rhode Island	Hartford	Greater Boston
Sample frame	Area probability	Area probability	Area probability	Area probability
Respondent	All household adults	Any responsible adult	Selected adult	Selected adult
Topic	Health	Health	Crime	Drinking
% sample responding by				
Phone	72%	62%	45%	55%
Mail	14%	2%	n/a	n/a
In-person interviews	5%	29%	30%	21%
FINAL RESPONSE RATE	91%	93%	75%	76%

DISCUSSION

Two of the major concerns about our current approaches to general population surveys are the eroding sample frames and the declining response rates. Area probability samples are a well-established approach to sampling general populations. Issues such as who lacks a household phone number or who has Internet service are not issues that affect area probability samples, nor will they be affected by whatever the latest technology is ten years from now. Another advantage is that such samples can be sent advance letters and other materials that may help with response rates. In addition, resources are growing for identifying who can be associated with particular addresses or housing units. Hence, more than in the 1970s and 1980s and for the studies described above, current researchers have several ways that they might be able to contact people in sampled housing units to collect information from them. For those people for whom the contact information is not adequate, given the data collection strategies being used, in-person interviewers offer an alternative way to approach those who have not been successfully contacted by other modes.

Another concern is nonresponse. People approached by telephone seem more and more to decline the opportunity to be interviewed. Mail surveys and Internet efforts to reach people and get them to respond likewise are limited in their effectiveness. Some people will play, while many others will not. An in-person interviewer may be the most effective way to enlist cooperation. Certainly, there are people who will respond to an in-person interviewer who do not respond to some of these other approaches to asking for help with a survey. Thus, adding in-person interviewers to the list of approaches that are used to enlist cooperation and collect data may be a very important strategy for bringing response rates up to a level at which they inspire confidence.

A final potential strength of in-person interviewing to note is that self-administered data collection can easily be included in the interview, either on paper or via computer-assisted self-interview (CASI). If the other modes used in a survey employ self-administration, that can help keep the measurement process consistent across all respondents and possibly reduce measurement error related to modes.

Table 2. Multimode Data Collection with Medical Sample

	Gallagher et al. (1999)
Location	Massachusetts
Sample frame	Medicaid member list
Respondent	Specific adult
Topic	Medical care experience
% sample responding by	
Phone	13%
Mail	34%
In-person interview	25%
FINAL RESPONSE RATE	72%

Obviously, we are aware that the reason alternatives to in-person interviews are so popular is that the unit costs for collecting survey responses tend to be lower when using mail, the telephone, or the Internet than they are for using personal interviewers. However, if adding in-person interviewers gets more valid estimates and more credible data, it may in fact make a lot of sense. Hence, as we move forward to think about how to collect data from general populations in the 21st century, we think it is very important that the old approaches, the area probability sample and the in-person interviewer, get serious consideration from researchers when they want to produce credible and valid statistics.

REFERENCES

Fowler, F. J., & Mangione, T. W. (1982). *Neighborhood crime, fear and social comfort.* Washington, DC: National Institute of Justice.

Gallagher, P. M., Fowler, F. J., & Stringfellow, V. L. (1999). *The nature of nonresponse in a medical population*. Paper presented at the International Conference on Survey Nonresponse, Portland, OR.

Hochstim, J. R. (1967). A critical comparison of three strategies for collecting data from

respondents. *Journal of the American Statistical Association, 62,* 976–989.
Klecka, W. R., & Tuchfarber, A. J. (1978). Random-digit dialing: A comparison of personal interviews. *Public Opinion Quarterly, 42,* 105–114.
Mangione, T. W., Hingson, R., & Barret, J. (1982). Collecting sensitive data: A comparison of three strategies. *Sociological Materials and Research, 10,* 337–346.
Thornberry, O. T. (1976). *An evaluation of three strategies for the collection of data from households.* Unpublished thesis, Brown University.
Waksberg, J. (1978). Sampling methods for random-digit dialing. *Journal of the American Statistical Association, 73,* 40–66.

SESSION 4 DISCUSSION PAPER: Of Frames and Nonresponse: Issues Related to Nonobservation

Mick P. Couper, University of Michigan

Collectively, the papers in this session present a remarkably pessimistic view of the future for health and other surveys in the next century. While they all address important research questions relating to data collection technology, the results are not particularly encouraging. The challenges the survey profession faces are daunting. Furthermore, none of the papers suggests a newly emerging technology that offers a panacea for our current woes. But I am more optimistic than this may suggest. While not downplaying the challenges, I believe the change that we necessarily face may lead us to approach these problems with greater creativity, which could lead to more insights into the extent of the problems and suggest possible solutions.

First, I believe that many of the problems we are facing in terms of errors of nonobservation (sampling, coverage, and nonresponse) may be offset at least in part by advances in terms of measurement. The quality of data we collect from respondents can be—and already is being—enhanced by the development and application of new technologies and methods. These include Internet data collection, audio- and video-CASI, and interactive voice response (IVR), to name but a few (see Couper, 2002, for a review). The Internet especially is allowing us to measure things in ways that were not possible—or, if so, were extremely difficult—in paper-based modes. So, while the golden age of telephone surveys may be over (as Frankel notes), I believe we may be entering a golden age of survey measurement. But this is Dillman's primary focus at this conference, so I will not say much more about this issue, except to note that if only we can solve the problems of how to identify, sample, make contact with, and obtain cooperation from members of the target population, we have an array of new and exciting measurement possibilities at our disposal.

Second, I believe the survey future appears daunting in part because while we think we see the possible demise—or at least degradation—of one of the most important data collection methods (telephone surveys), we cannot yet clearly see what is likely to replace it. While the future indeed appears uncertain, now is the time to focus our research energy on exploring alternative methods and evaluating alternative strategies to overcome the problems of sample representation that seem at present so intractable. The survey method has proven remarkably resilient over almost a century of use, and I believe the profession will rise to the challenge and find ways to overcome both current and future threats to the method.

Notwithstanding Gallagher and Fowler's important contribution on the role of face-to-face surveys, this session is primarily about two methods of survey data collection—telephone and Internet—and it is on these methods that I will focus my remaining remarks. Errors of nonobservation are typically grouped into sampling errors, (non)coverage errors, and nonresponse errors (e.g., Groves, 1989), and I will offer a few observations on each of these in turn.

SAMPLING ERRORS

Random-digit-dial (RDD) sampling for telephone surveys was a wonderful invention. It worked because of the structure of the U.S. telephone system and was so successful because of the high penetration of landline telephones in households. It offered great savings in cost over the development of area probability samples. As the number of nonworking nonhousehold telephone numbers has increased, the method has been adapted to optimize efficiency. However, RDD is a particularly American phenomenon. It is not widespread in other parts of the world. For example, there have only recently

been efforts to introduce RDD sampling in Britain (Nicolaas & Lynn, 2002). In addition, it is becoming increasingly difficult to implement and is a costly and inefficient approach to carry out (i.e., the yield of working household numbers is declining).

By far, the biggest threat to sampling for telephone surveys is the increasing penetration of cellular phones, particularly of cell-phone-only persons or households, as Blumberg, Luke, and Cynamon aptly note. While current RDD samples ignore cell phones, we increasingly will have to include them in our frames. From a sampling perspective, this will likely involve a separate stratum and the development of efforts to adjust for duplication of persons in both frames (landline/household and cell/person). I return to this issue under coverage below.

While RDD sampling revolutionized the survey industry, the same is not likely to happen for the Internet. I believe that no RDD-like equivalent is likely to emerge for the Internet in the foreseeable future, and anti-spam norms and legislation are likely to preclude any such efforts. Internet surveys thus will need to rely on other methods for sample selection (see Couper, 2000, for a review of approaches). Basically these can be grouped into the following strategies: (1) list-based samples, (2) transaction-based approaches, (3) recruitment through other means (e.g., face-to-face or telephone), or (4) Web as one alternative in a mixed-mode design (the approach used by Link and Mokdad is one example).

Another option is already widely used in the market research area—using nonprobability designs, such as Internet panels of self-selected volunteers. While it may be heresy to suggest considering such an approach for government and academic research, I believe it is time to focus research efforts explicitly on the trade-offs between such approaches and, say, a probability-based design using a much more expensive recruiting method (e.g., telephone) but achieving a very low response rate. Both raise concerns about representation. I am by no means advocating the wholesale abandonment of probability-based sampling methods. However, as with the current debate over response rates and nonresponse error (see below), we really need to evaluate the relative efficacy and quality of alternative approaches for different research topics. For example, is an RDD survey that achieves a response rates in the teens *necessarily* better than a study based on a sample of volunteers? What if the nonprobability design was one-third the cost of the probability-based design? But what if the goal were to make projections to a broader population with a high level of precision? I believe there are likely to be some health survey applications for which nonprobability designs may yield data of sufficient quality at lower cost than traditional methods. The Baker, Zahs, and Popa paper points to some of the dangers of this approach but also demonstrates the kind of research that we should be doing to understand when, how, and why we are likely to get different estimates.

COVERAGE

The proponents of Web surveys have long argued that even if Web access is not yet universal, it will approach such levels in the near future. I'm not so optimistic. Using data from Baker, Zahs, and Popa's presentation, the rate of growth of Internet penetration appears to have slowed in recent years. Whether this is just a dip in a continuing upward trend or the innovation adoption "S" curve has started to flatten is not clear. Prognostication is a risky business, but I believe that the Web—in its present form—will not reach the levels of penetration achieved by the telephone for many decades to come. The telephone is primarily a *communication* device, while the Web is still predominantly an *information* device, requiring literacy and relatively expensive equipment to fully exploit.

This simply means that high-quality probability-based Internet-only surveys of the general population are not likely soon. This will not stop people from trying—witness the contrasting efforts of Harris Interactive and Knowledge Networks. Nor will it stop others

from making claims of representativeness based on demographic matching or other strategies. However, because coverage of the general population is far from universal doesn't imply that we should abandon the method—there are an increasing number of populations of interest to researchers that have high rates of coverage and sometimes also a list of e-mail address for sampling. These include college students, health care providers, members of professional associations, users of Internet health care services (e.g., WebMD), and so on. I also believe Web surveys increasingly will be used as an alternative response mode for mail surveys, for follow-up studies of panel members with Internet access, or in other types of mixed-mode approaches. The same may be true of telephone surveys—the threats to RDD surveys do not apply to the same extent to mixed-mode or panel designs.

NONRESPONSE

While it may be hard to demonstrate empirically, I am convinced that telemarketing (and direct marketing in general) has much to do with declining response rates. I share Frankel's optimism that the Do Not Call (DNC) list actually may help surveys by dramatically reducing the number of unwanted calls and increasing the likelihood that household members we contact listen to our requests. However, I worry that the DNC list may have come too late. People already have changed their telephone answering behavior in response to the deluge of telemarketing calls, and even the introduction of the list may not get them to change back to the old ways again. This is an important area for research but unfortunately requires access to the DNC list. For example, what about those who have not (yet) signed up on the list? Is this because they are happy to receive telemarketing calls, are not aware of the list or how to register, or have found alternative (and possibly more effective) methods of screening unwanted calls? Understanding how telephone-answering behavior has changed and how those on the DNC list are different from those not on the list is an important step in understanding and adapting to the presence of the list.

Given that the list is closely held by the FCC, this may require some approach by a federal statistical agency or a coalition of the survey industry. Regardless, understanding what the DNC list is doing to telephone surveys should be a research priority.

We also need new strategies to tackle the problem of nonresponse in general. For too long our efforts relating to nonresponse have focused on **rates**. We expend enormous amounts of effort and money to move a response rate above a critical, usually predetermined, plateau. We need to turn both research and practice to a focus on **error**. Our strategies and designs must move from a rate-focused to an error-focused perspective.

I'm increasingly coming to believe—along with many other researchers in this field—that nonresponse error is not a certainty with low response rates. While high response rates certainly reduce the **risk** of high nonresponse error, a low response rate does not necessarily mean high nonresponse error. Recent papers by Keeter et al. (2000) and Curtin, Presser, and Singer (2000), among others, have demonstrated this. However, this also does not mean that we can conclude that nonresponse is ignorable, that there is no (or little) nonresponse error. It simply means that the rate is a poor proxy for error. Furthermore, there is no reason to expect a linear relationship between response rate and nonresponse error. This is an area ripe for further research.

This error-focused perspective is what we (Groves & Couper, 1998) referred to as "designing for nonresponse," which is now being referred to as responsive design. The approach has several ingredients, which include the following:

(1) Explicitly acknowledging the inevitability of nonresponse at the design stage.
(2) Setting aside resources for nonresponse-related strategies.
(3) Focusing those strategies not simply on increasing response rates but rather on minimizing differences between

respondents and nonrespondents or understanding how nonrespondents may differ from respondents.

The responsive design approach is greatly facilitated by the use of survey technology for survey data collection. For example, tracking both effort and costs on an ongoing basis is made easy with the use of computerized call records and cost reporting data. The information available to the designer is much richer than in the past and is available in time to make decisions in the middle of data collection activities. Similarly, the distributions on key variables can now be tracked on a continuous basis during data collection, allowing us both to see how estimates stabilize or change over time and to permit modeling of alternative end-game scenarios or strategies for later sample replicates based on the performance of replicates released earlier. To put it simply, working harder will no longer lead to the production or quality goals we desire—we must work smarter.

AN AGENDA FOR THE NEAR FUTURE

With the above observations in mind, let me offer a few thoughts as to what to do next. I believe it is vitally important to do the research now rather than waiting until the new technologies have emerged fully or events such as the DNC list have played themselves out. We need to anticipate the future trends and to prepare for changes that may be necessary in the design of many of our large and important survey data collection efforts. Several of the papers in this session demonstrate the kind of research that needs to be done. For example, Blumberg, Luke, and Cynamon reported on a set of questions relating to cell phone usage added to the NHIS to explore the correlates and potential consequences of increasing cell phone penetration on key health estimates. This ought to be expanded to include Internet access and use. For example, researchers at the University of Michigan and RAND (with funding from NIA) are engaged in research on the possible role of the Web in the Health and Retirement Survey, a large panel survey of persons age 50 and older. While this is not the first population that comes to mind when thinking of the Web, this research is not being conducted with the goal of replacing the interviewer-administered survey with a Web version in the near future but at understanding how the Internet may be used to supplement the ongoing data collection. More of this kind of research ought to be done. Similarly, Link and Mokdad give an example of the kind of work that can be done to explore mixed-mode approaches for surveys such as the BRFSS. Baker, Zahs, and Popa's paper is in the same spirit.

Now is the time to understand how telephone and/or Internet access and use is related to the key health outcomes we are interested in measuring. This cannot be done by closing ranks around the important data collection efforts we currently conduct, as if they were under siege. On the contrary, these national surveys—precisely because of their high quality—must serve as the platforms for exploring and understanding the alternative methods.

Despite the heady optimism of many in the market research industry, Web surveys are unlikely to replace high-quality federally funded or conducted surveys anytime soon. The Baker, Zahs, and Popa paper points to some of the reasons for this, as does the Link and Mokdad paper, bolstering the argument of Gallagher and Fowler for not abandoning traditional data collection methods. However, I believe that the Web has an important emerging role to play as a supplement to the existing methods of data collection. It remains to determine how and when to best make use of the strengths that the Internet offers in terms of survey measurement and how to avoid the potential pitfalls. The more we undertake studies like those described in this session, the better equipped we will be to face the challenges, with respect to both the Internet and telephone surveys.

We often think of technology as limiting, especially when we consider threats to coverage, sampling, and nonresponse. However, we also should think of technology

as enabling—e.g., providing technology to people in order to facilitate response (and accurate measurement). In this perspective, cell phones, PDAs, wireless Internet devices, etc., can be seen as ways to broaden the strategies available for keeping contact with and obtaining information from those we have contacted and enlisted in our survey efforts.

REFERENCES

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Couper, M. P. (2002, August). *New technologies and survey data collection: Challenges and opportunities.* Invited plenary presentation at the International Conference on Improving Surveys, Copenhagen (SMP Working Paper No. 115).

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, *64*, 413–428.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, *64*, 125–148.

Nicolaas, G., & Lynn, P. (2002). Random digit dialling in the UK: Viability revisited. *Journal of the Royal Statistical Society: Series A, 165,* 297–316.

SESSION 4 DISCUSSION PAPER: The Conundrum of Mixed-Mode Surveys in the 21st Century

Don A. Dillman
Washington State University

The four papers presented in this session are a stunning reminder to me of a discussion session at the first Health Survey Research Methods Conference held at Airlie House, Virginia, in 1975. I was invited to attend that conference because of concern about the increasing costs associated with face-to-face interviews and my development of a survey data collection center—the Washington State University Public Opinion Laboratory established in 1970 for conducting sample surveys by telephone (Dillman, 1978). Prior to the Airlie House Conference, I had presented results from those surveys at professional meetings, where the results often were questioned. Reviewers seemed to think that most people would not answer questions over the telephone, or if they did, the answers would not be valid.

At Airlie House, a round-robin exchange of information occurred in which several of us shared our telephone survey experiences. My main memory of that session was that one of the participants (I believe it was Seymour Sudman from the University of Illinois) summarized the discussion by noting that several participants had used the telephone to collect survey information, interviews of 10–20 minutes had been conducted without producing high refusal rates or even mid-interview cut-offs, and that one participant had even conducted interviews 45 minutes in length. He concluded that we appeared to have a new survey mode for health research that showed great promise for future use as an alternative to face-to-face interviews. I left that conference no longer feeling alone in my attempts to use telephone interviews for survey data collection.

Within a decade of the Airlie House Conference, the telephone interview would become recognized as an acceptable replacement for many face-to-face interview studies. An international conference devoted solely to the conduct of telephone surveys was held in 1987 (Groves et al., 1988), which attracted over 400 participants. In addition, a methodology journal devoted an entire issue to this new method of doing surveys (*Journal of Official Statistics*, Vol. 4, No. 4, 1989), and interest in telephone data collection came to dominate survey methodology in the early 1990s.

Yet, here we are, almost 30 years after the Airlie House Conference, questioning whether random sample telephone surveys are going to survive. We have listened to four papers, each of which gives a decidedly different image of the telephone interview than that which persisted only a few years ago. Frankel has reviewed the rise of random-digit dialing and the procedures that made random sample general public surveys by telephone possible. He predicts that the number of telephone interviews will decrease, and the conditions of interviewing (e.g. providing payments to respondents) also will change. However, he also notes that as of yet we have no alternative and that the use of RDD surveys will survive.

Baker and his coauthors have discussed the potential of Internet surveys for obtaining health information from respondents but conclude that the Web remains a niche methodology. Web users are different from nonusers, and although more aggressive weighting may help, it is not sufficient for allowing the Web to replace RDD surveys.

Blumberg, Luke, and Cynamon have described another threat to RDD surveying—the substitution of wireless phones for landlines. Although their investigation found that the exclusion of adults who could not be reached because of having "cut the cord" did not have a demonstrable effect on estimates of health and health care service use, it is a change that must be closely monitored. There seems to be little doubt that the percent of wireless-only customers will continue to rise,

although more in some categories of respondents than others.

Link and Mokdad noted a decline in RDD response rates and suggest that the use of Web surveys and mail surveys in conjunction with CATI may be viable alternatives for increasing participation in the Behavioral Risk Factor Surveillance System Surveys. However, they also note the existence of many barriers to conducting such mixed-mode surveys. Gallagher and Fowler also suggest that mixed-mode surveys offer promise for resolving the problem of RDD telephone surveys, noting that personal interviewing may be part of that mix.

This conference serves as a reminder, should one be needed, that particular modes of collecting survey data, whether face-to-face, telephone, mail, or even the new ones—the Internet and Interactive Voice Response (or touchtone data entry in an earlier form)—exist in a cultural and technological context. The threat to telephone surveying exists not because survey questions can't be asked and answered over the telephone but because society has new communication modes and ways of making them available to people. The sample frame convenience of random-digit dialing is under pressure because the telephone is becoming a personal device rather than household device, with different coverage qualities. The potential for talking with people on the telephone also is decreasing because society is changing with regard to how the telephone gets used, with greater emphasis on leaving messages and less on interactive conversations. The need to conduct sample surveys to discover unknown prevalence of health characteristics and their distribution patterns is not in decline, but our methods for obtaining these data seem destined to change. If there is any comfort in this state of affairs, I find it in the fact that once again survey methodologists are assembled from around the world sharing ideas about possible solutions to today's concerns.

One part of the solution for some surveys may be the use of mixed-mode surveys, a possibility suggested in the papers presented in this session. There are several reasons that much interest is being generated in conducting surveys in which data is collected from some respondents by a different mode than that used to collect data from other respondents. One reason is to reduce costs by getting some people to respond by less expensive modes so that use of the most expensive ones can be reserved for a smaller number of respondents. In addition, the introduction of a second or even a third mode has been proposed as a means of improving response rates and for obtaining data from different kinds of respondents than those who will respond to other mode(s). Thus, a major motivation for designing mixed-mode surveys is to reduce nonresponse error. Interest also exists in using second modes to contact people inaccessible by other modes.

Interest in mixed-mode surveys as a means of reducing nonresponse error introduces the possibility that another kind of error may negate improvements in the former (Dillman & Christian, in press). The use of second and third modes for collecting data raises the prospect of greater measurement error.

There is considerable evidence in the literature that different survey modes encourage respondents to answer questions in different ways. For example, beginning with the Hochstim (1967) study mentioned by Gallagher and Fowler, questions about respondent health consistently have produced less positive answers in self-administered surveys than when asked in interviews. The reason is social desirability. People's answers to such questions are culturally based, and answers are given to certain questions that tend to meet other's expectations. I was not surprised that this familiar pattern of more positive answers by telephone was reported for two different Web and phone survey comparisons in the paper by Baker et al.

Telephone surveys also are more likely to encourage "agreement" (acquiescence) to questions (e.g., Schuman & Presser 1981). In addition, evidence exists (albeit mixed) that respondents to self-administered questions are more likely to select from early-offered

answers to questions, while interview respondents are more likely to choose from among the last-offered choices (Dillman, Sangster, Tarnai, & Rockwood, 1996). I am not optimistic that such differences can be eliminated between interview and self-administered surveys, although data adjustment or indexing for certain frequently asked questions (for which considerable experimental comparison data exist) may be possible. So long as mixed-mode referred only to the mixing of telephone and face-to-face modes, measurement difference was not usually a serious issue. Most studies have not shown large differences between these two aural modes (see de Leeuw, 1992; Groves & Kahn, 1979). Although frequent differences between mail and telephone interviews have been clearly documented, such differences have not attracted a great amount of attention in health surveys because the major health surveys have depended much more on interview modes than on self-administered paper surveys. The introduction of the Internet, the use of which is likely to rise dramatically over the next decade, means this situation is about to change. Internet surveys depend upon visual communication and are self-administered, as are mail surveys; telephone interviews, on the other hand, depend only upon aural communication. It is becoming increasingly evident that these differences may be the basis for the occurrence of significant mode differences in the answers provided to survey questions.

Although the research done to date is limited, I suspect that just as social desirability, acquiescence, recency (tendency to select last rather than first categories), and other context effects (e.g., observance of a norm of even-handedness when answering related questions) are thought to occur with greater frequency in telephone interviews than in mail surveys, the same will happen for Web surveys. The theoretical rationales for their likely occurrence in mail surveys, based upon interviewer presence and control of the answering process, would seem to apply in similar ways to the processing of answers for self-administered Web surveys. Research on these issues should be a high priority for health researchers in the immediate future.

Another difficult problem that designers of mixed-mode surveys must contend with is that researchers tend to structure questions in different ways when asking questions by different modes. For example, designers of mail and Internet surveys often ask check-all questions, in which respondents are asked to read a list of items and mark all of the answers that apply to them. Such questions are invariably formatted when asked in telephone surveys, with the respondent being asked to indicate yes or no to each of the items. Evidence exists that yes/no queries are checked more often in self-administered questionnaires than when the same questions are asked in a check-all format (Dillman, Smyth, Christian, & Stern, 2003; Rasinski, Mingay, & Bradburn, 1994). Many other differences also exist in the way questions typically are asked in interview vs. self-administered surveys (see Dillman & Christian, in press, for additional examples).

Suggesting to survey designers that they ask questions in the same way across all modes often leads to resistance. For example, I once proposed to a national survey organization that they not use show cards for face-to-face interviews in order to get complete comparability with the telephone interviews. Their response was, "But this is the way we always do our face-to-face interviews." Similarly, proposing that "don't know" and "no opinion" answers be explicitly offered in telephone interviews, where they are available to the interviewer for use but not mentioned, and mail surveys, where they are most often omitted altogether, has led to resistance from designers for both survey modes. Unimode construction—i.e., the construction of questions for all modes in the same way (Dillman, 2002)—represents a significant challenge for those conducting mixed-mode surveys.

Another related source of potential mode differences, the effects of visual design and layout, has until recently received little attention. Evidence now exists that respondents to visually administered surveys

draw information not just from the wording of questions but also from symbols, numbers, and graphical layouts of questionnaire items (Redline & Dillman, 2002). Just as interviewer voice inflection, speed, and other voice characteristics act as an aural paralanguage and communicate information in addition to words, these additional sources of information act as a visual paralanguage to further define the meaning of questions in self-administered Web and paper surveys. Experimental evidence now exists that the modification of symbols and graphics on both mail and Web surveys may produce different answers to identically worded questions (Christian, 2003; Christian & Dillman, 2004; Dillman & Christian, in press).

An experimental comparison of responses to telephone, IVR, Web, and paper surveys showed that telephone and IVR respondents were much more likely to pick the most satisfied category on a five-point scale (in which 1 meant "completely satisfied" and 5 meant "not at all satisfied," and they also could choose any number in between). The percent choosing this category was telephone, 39%; IVR, 39%; mail, 21%; and Web, 26% (Dillman, Phelps, Tortora, Swift, Kohrell et al., 2001). These patterns of the aural modes producing similar results and the visual modes also producing similar results to each other existed for *the other opinion questions contained in that survey*. It was reasoned that differences in these modes of communication might account for the differences, with the intermediate boxes drawing more attention when appearing in the mail and Web formats. Thus, in a follow-up experiment, the visual format of a linear scale complete with numbers and check boxes for each of the five points was withheld in the visual form so that the stimulus delivered aurally would be the same as that delivered on paper. The respondents to the visual format were offered only a blank box in which the number of their scale value was to be recorded. The scale values obtained from the number box on both paper (Christian & Dillman, 2004) and the Web (Dillman & Christian, in press) were quite different from those obtained via a linearly displayed scale, each of which had its own check box.

More negative responses were consistently given on the number box format. One of the reasons for the observed differences appeared to be that respondents became confused on the direction of the scale when the visual support was replaced by only a number box, requiring the respondent to carry the meaning of numbers from the stem of the questionnaire to the blank where the number of their answer was to be recorded. The fact that the meaning of questions is communicated visually through symbols, numbers, and graphical features in addition to words raises the question of which visual formats translate best into equivalent aural formats for interviews and which do not.

Other experiments have shown that visually administered polar-point scales with words for only the extreme points produce different results than when each category receives a verbal label (Christian, 2003). Significant differences in category answers also occurred when the visual display was linear instead of double- or triple-banked, as is sometimes done for both paper and the Web. This research needs to be expanded in order to learn which visual displays translate best to aural survey modes so that equivalent answers may be obtained across all survey modes. The prevalence of opinion questions in health surveys suggests that a high priority needs to be assigned to such research.

The times in which we live and work as survey methodologists are much more complex than those that faced us at Airlie House in 1975. Many indications at this conference suggest we now face a decline in the capability of the telephone for the collection of health information and strike me as reasons for consternation. However, that concern is perhaps no greater than the distress felt by many who attended the Airlie House conference about the difficulties then facing face-to-face interviewing. It was important then, and it is important now, to respond by investing in the development of new survey mode capabilities. An increased emphasis on resolving problems associated with the use of

mixed-mode surveys seems an important part of that response.

REFERENCES

Christian, L. M. (2003). *The influence of visual layout on scalar questions in Web surveys.* Unpublished master's thesis. Pullman, WA: Washington State University.

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered surveys. *Public Opinion Quarterly, 68*, 58–81.

de Leeuw, E. (1992). *Data quality in mail, telephone, and face-to-face surveys*. Amsterdam: TT Publications.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: John Wiley.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method*. New York: John Wiley.

Dillman, D. A., & Christian, L. M. (in press). Survey mode as a source of instability in response across surveys. *Field Methods.*

Dillman, D. A., Sangster, R., Tarnai, J., & Rockwood, T. (1996). Understanding differences in people's answers to telephone and mail surveys. In M. T. Braverman & J. K. Slater (Eds.), *Current issues in survey research: New directions for program evaluation series* (pp. 45–62). San Francisco: Jossey-Bass.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., & Berck, J. (2001). *Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response and the Internet.* Paper presented at the annual meeting of American Association for Public Opinion Research, Montreal.

Dillman, D. A., Smyth, J., Christian, L. M., & Stern, M. (2003). *Multiple-answer questions in self-administered surveys: The use of check-all-that-apply and forced-choice question formats.* Paper presented at the annual meeting of the American Statistical Association, San Francisco.

Groves, R. M., & Kahn, R. L. (1989). *Surveys by telephone*. New York: Academic Press.

Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls II, W. L., & Waksberg, J. (1988). *Telephone survey methodology*. New York: Wiley-Interscience.

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

Rasinski, K. A., Mingay, D., & Bradburn, N. M. (1994). Do respondents really "mark all that apply" on self-administered questions? *Public Opinion Quarterly*, *58*, 400–408.

Redline, C. D., & Dillman, D. A. (2002). The influence of alternative visual designs on respondents' performance with branching instructions in self-administered questionnaires. In R. Groves, D. Dillman, J. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse*. New York: John Wiley.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys experiments on question form, wording, and context*. New York: Academic Press.

SESSION 4 SUMMARY

Richard Kulka, RTI International
Floyd J. Fowler, Jr., University of Massachusetts Boston

DISCUSSION

This session emerged from growing concerns about the potential viability of RDD survey methods for conducting high quality health surveys in future years. In the "keynote paper," Frankel traced the history of telephone surveys and the eventual emergence and dominance of RDD surveys as the primary mode for surveying the general population in national, state, and local surveys and the substantial impact such surveys have had on our society. He also described the emergence and growth since the 1990s of substantial impediments and threats to the continued health and effectiveness of RDD surveys, while also postulating that such surveys will remain viable due to the lack of any alternative low-cost method for conducting probability sample surveys. Blumberg, Luke, and Cynamon provided a detailed description and analysis of one of the key growing challenges to conducting RDD telephone health surveys in the years ahead—the increasing prevalence and impact of households abandoning residential landline telephones for wireless cell phones.

In the face of these and other threats, survey researchers are more aggressively exploring alternative modes to the telephone and RDD for conducting health and other surveys, most notably the Web. Baker and Zahs described several key known limitations of Web surveys and provided a detailed empirical comparison of a major telephone health survey conducted by RDD with comparable surveys administered online to sample members from two Web panels. The two other papers in this session explicitly addressed the growing need to consider mixed-mode survey methods, both as an alternative or complement to RDD surveys and as a valuable strategy in and of itself. To address the threat of declining response rates in a large ongoing state-based RDD telephone survey conducted by CDC, Link and Mokdad presented the results of experiments designed to determine whether conducting these surveys by either Web or mail with telephone follow-up of nonrespondents would constitute viable options to a single-mode RDD survey. Gallagher and Fowler presented both historical and recent data demonstrating the significant benefits on survey response (and other factors) of using area probability sample frames for multi-mode studies and including in-person interviews as a follow-up for those who cannot be successfully contacted or surveyed by other modes.

In combination, these brief presentations served to stimulate a very active and vigorous discussion on the viability of RDD telephone surveys, the potential for greater use of mixed-mode surveys, and the implications of these evolutionary changes for survey design and practice. This discussion was framed well by the two formal discussants. By previous agreement, Couper's discussion focused on issues of *nonobservation* (e.g., frame and nonresponse errors) associated with RDD and alternative modes and designs, while Dillman addressed issues related to *observation*, especially the potential consequences of mode changes or combinations of modes for the design of questions and questionnaires and related measurement issues.

The floor discussion was wide ranging but can mostly be captured by consideration of three general issues or categories:

(1) Increases in "frame erosion" and coverage problems with RDD sample surveys;
(2) Increases in nonresponse or declining response rates; and
(3) Increasing reliance on other modes, and especially multiple modes—combinations of survey modes used to address increasing challenges and threats to RDD surveys.

FRAME EROSION & COVERAGE PROBLEMS

Several participants spoke to the threats posed by the increasing penetration of cell phones, "Do Not Call" lists, and other significant threats to population coverage. That the challenges and threats to coverage to RDD surveys posed by Frankel and others were indeed real was generally conceded by most participants, and those who addressed the coverage issue generally spoke to the increasing decline in landline telephone use relative to wireless and wireless-only households and recent trends toward the merging of these telephone numbers (i.e., number portability). Blumberg noted that sampling and surveys of cell phone users can indeed currently be done legally (although autodialers cannot be used in such cases), but current practices in the U.S. to charge users for incoming calls (there is no way currently to get the caller to pay) have dampened their use in this regard. One participant observed that we must eventually use cell phones and noted that in Canada, there is no charge if the call is made by the government. Frankel re-emphasized that there are ways to sample cell phones, while also expressing optimism that the use of incentives to overcome call costs and the development of new statistical approaches to compensate for biases will be developed to address these threats.

Other participants emphasized a growing need to track and document changes in the population distributions of prevalence, use, and "rules" associated with these various forms of telephony as part of ongoing surveys conducted in person (e.g., the NHIS, the NSDUH) to understand the implications of these evolving changes for population coverage, potential biases, and weighting and other strategies for addressing these. One participant emphasized that, in essence, the continued use of some in-person interviewing is essential to the continued use of RDD to support weighting adjustments, and several others noted the continued use of in-person interviews as a baseline form of data collection for subsequent surveys or survey waves conducted by telephone.

The current state of our knowledge in this area encompasses the following:

(1) The sample frame for RDD surveys, based on household or "landline" phones, is likely to continue to erode. The time is not far off when RDD surveys will have to include individual cell phones in the sample frames.
(2) There is no obvious, readily available frame that would provide an alternative to RDD at this time for general population samples.
(3) Area probability frames are one appropriate basis on which to draw population samples. One of the important uses of in-person surveys based on area probability samples (such as the NHIS and CPS) is to provide sample frames for other surveys that may be conducted in modes other than in person. Address lists developed by the Post Office may provide an increasingly useful start for developing area probability samples (Iannacchione, Staab, & Redden, 2003).

The greatest need for research with respect to sample frames is to conduct studies of the comprehensiveness of alternative frames and to study the ways in which those excluded from various frames do and do not differ from the population as a whole. One of the important uses of in-person surveys based on area probability samples is to describe the people who would be included and excluded from less comprehensive frames, such as those who have household and/or wireless telephones or those who have or do not have Internet access.

NONRESPONSE

While threats to coverage were of significant concern, declining response rates and the potential for serious nonresponse bias in RDD surveys were by far the more significant concern as a source of nonobservation error, and these issues generated considerable discussion. In general, the discussion revolved around three basic topics:

(1) How to motivate people to respond in the face of these new challenges
(2) Does nonresponse really matter?
(3) The use of multiple survey modes to facilitate and motivate people to respond

In response to the question of whether "RDD as a single mode will be credible to collect population statistics five to ten years from now," it was emphasized that the answer depends on what and how much we are able to do (e.g., the use of Web sites and follow-up letters/strategies to legitimize surveys, even more extensive use of additional calls) and how much "heat" we are willing to take, including time and costs. For a number of reasons, however, including proscriptions by IRBs, as well as legal and technological barriers, the effectiveness of using such strategies is likely to diminish. Throughout the discussion, reference was repeatedly made to the likelihood of an increased need in the profession to use respondent incentives—i.e., to pay respondents for their participation. This issue was explicitly raised, with note being made that investigators and interviewers are being paid, substantial resources are being expended on technology and other survey features, and the use of respondent incentives can be very cost effective. Several participants noted the use and effectiveness of incentives in some key ongoing surveys, while others cautioned that this is not always the case and that some caution is necessary. In particular, they noted that money is only one reason that people participate in surveys. It is important to continue to do a good job of explaining to potential respondents how their goals and values will be enhanced by participating in surveys, as well as offering them monetary incentives sometimes.

Reflecting on the increasing use of Web surveys, there also was some discussion of the possible delivery and effectiveness of respondent incentives under that survey mode. One participant suggested the potential use of "electronic gift certificates," but current research evidence suggests that no electronic equivalent (e.g., PayPal) has the same effect as cash. However, an incentive offered in a survey of parents of children with ADD was especially effective in getting people to respond to a Web survey.

Inevitably, given this topic, reference was made to a number of recent papers on nonresponse in which surveys that varied widely in their response rates (e.g., 30% vs. 60%), showed few or no differences in their population estimates, thereby leading many observers to ask, "Do response rates really matter?" Couper emphasized that the relationship between nonresponse rate and nonresponse error is neither precise nor deterministic, so that the former should not be used as a proxy for the latter, and that the relationship is not necessarily (or probably) a linear one. He noted the need to coordinate and tailor efforts to increase response rates with careful monitoring of key survey estimates until they become stable, after which such efforts should be terminated. While lamenting recent tendencies to concentrate on some specific sources of error rather than on total survey error, a participant noted that although there are many measures for which substantial levels of nonresponse indeed make no difference, there also are notable situations in which even modest differences in response rate made a substantial and important difference (e.g., the National Comorbidity Survey). Gallagher emphasized that, in their use of multiple survey modes to increase response rates, each additional step brought their population estimates closer to the frame.

The state of our knowledge in this area includes the following points:

(1) Response rates for RDD-based surveys are certainly declining, and many surveys done by high quality organizations currently achieve response rates that a decade ago would have been considered unacceptable.
(2) Response rates and nonresponse error are not synonymous. The effect of response rates on survey estimates varies from

survey to survey and even from estimate to estimate within a survey.

(3) Prepaid monetary incentives raise response rates. Repeated contacts raise response rates. Effectively presenting the reasons for a survey probably improves response rates.

Research that is most needed includes:

(1) **Studies of the relationship between nonresponse and response error for all topics.** We need a much better understanding of when low response rates have important effects on estimates and when they do not.

(2) **Studies on alternatives to monetary incentives to motivate respondents to cooperate.** Some researchers have been experimenting with interviewer training as a way to improve response rates. There also have been studies of the value of various advance materials. However, the results to date have not been very definitive, and there is much more to be learned on these topics.

(3) **Studies of the effects on survey estimates of monetary incentives and other approaches to increasing respondent motivation to cooperate.** There are recurring questions about whether certain approaches to motivating respondents to participate improve or adversely affect survey estimates by increasing motivation or by inducing people to respond who really are not interested.

INCREASING RELIANCE ON MULTIPLE SURVEY MODES

A significant portion of the discussion focused on the increasing use of multimode surveys as an important tool in addressing the nonresponse issue. One key distinction was made between the mode of contact to elicit cooperation versus the particular survey mode used to participate in the survey—i.e., method of contact versus medium of measurement. Once sample members are successfully contacted and enlisted to participate, the potential use of a broad variety of data collection modes—mail, telephone, Web, PDAs, IVRs, etc.—is enabled. The significant success of the CPS as predominantly a telephone survey is facilitated by initial contact in the face-to-face mode.

More generally, several participants emphasized the increasing need, value, and trend toward using different (and sometimes) multiple survey modes—either sequentially or simultaneously—to approach different populations. Surveys largely have been built on the concept of one design for everyone, but there is increasing concern and evidence that one size does not fit all and that different modes and approaches may be needed to effectively survey different populations, subgroups, and individuals. The presumption is that certain modes and approaches fit certain people better, and potential respondents are more likely to participate if given a chance to respond in ways with which they are more comfortable. One participant described the analogy of changes in the banking industry, where they quickly discovered that they needed to keep "all channels open" (e.g., tellers, ATMs, and online banking). Another noted that modes need to reflect the ways in which people react to different approaches and technology, e.g., the ways in which cohorts comprehend information. It was postulated that an effective mixed-mode strategy might not follow a conventional "cheapest to most expensive" strategy but rather be tailored to each respondent in no particular order. Some suggested that face-to-face interviewing will probably always be a necessary mode in population surveys in order to reach less accessible members of the population, which for many purposes contain those most in need of measurement.

While changes in mode are likely to draw different people, such changes run the considerable risk of significant changes in the measurement process and measures, a theme explicated by Dillman in his formal discussion. Different modes offer significantly different stimuli, and a key challenge is to develop appropriate strategies to make them comparable or common in their measurement

properties and results, a process that can be described as making our surveys "mode-proof" or "unimode." However, data comparability across modes will not be achieved by "surface" comparability but rather will require different ways of presenting questions in both automated and nonautomated (and mediated and nonmediated) environments to achieve equivalence in the stimuli and their meaning. Dillman noted that we are in an age where we are rapidly moving to the use of all different modes simultaneously, using each as appropriate, thereby requiring the ability to make these come together seamlessly, all in the same survey organization. This requirement often runs directly counter to the organizational barriers and structures we have created to efficiently implement single mode surveys. Thus, there are both organizational or structural and methodological or design challenges to meeting the multimode and other demands of surveys in this new age and century.

The state of our knowledge on the significance of mode of data collection includes the following:

(1) Respondents differ in the mode of contact that is most likely to reach them and the mode of approach that is most likely to enlist their cooperation. The best mode (or combination of modes) will depend critically on the topic and the population. Using multiple modes of contact increases response rates.
(2) The mode in which people answer (ostensibly the same) questions can affect the answers. Many estimates have been shown to be very similar across modes. For example, Hochstim (1967) made over 1,000 comparisons of estimates obtained by phone, in-person interviewer, and self-administration and found only about 50 significant differences. On the other hand, we have good data showing that interviewers obtain more "socially desirable" answers than are provided via self administration, and there are a number of other differences that appear to be related to mode of administration as well (although sometimes the results are inconsistent).

Research that is most needed includes the following:

(1) **More and more comprehensive studies of how mode of contact and data collection affect who responds to surveys.** It is reasonable to think that nonrespondents to one mode may result in less nonresponse bias than those to another mode for a particular topic. The interaction between survey mode and nonresponse error is little understood but quite critical, as researchers explore alternative ways to collect data and reduce nonresponse.
(2) **How to collect comparable data across modes.** What are the principles of instrument design that will maximize comparability across modes? Are there topics (such as those that have a large component of social desirability) or approaches to designing questions that always will produce different results depending on mode?

REFERENCES

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

Iannacchione, V. G., Staab, J. M., & Redden, D. T. (2003). Evaluating the use of residential mailing lists in a metropolitan household survey. *Public Opinion Quarterly*, *67,* 202–210.

INTRODUCTION TO SESSION 5: Security and Privacy

Marcie L. Cynamon, National Center for Health Statistics

Issues surrounding privacy, confidentiality, and data security were a subject of discussion at the first conference. Almost thirty years later, these subjects continue to fuel challenges to survey research. Although researchers need the freedom and flexibility to conduct their business, the rights of study participants can never be overlooked or downplayed. Today, the basic concerns about the protection of the rights of human subjects are compounded by advances in legislation and technology that introduce new challenges to the complex contract between study participants and researchers. These developments have encouraged review boards to go beyond traditional safeguards and carefully consider the legal rights and potential social harms that may exist for survey participants. However, in doing so, review boards risk bureaucratic decision-making and heavy-handed interpretations of the Common Rule (45 CFR 46) and other relevant regulations. To combat those risks, researchers must provide review boards with sufficient background information and acceptable methodological alternatives so that these boards may reasonably interpret and apply the regulations.

The presenters in this session all share the common goal of conducting research while protecting study participants. Their presentations describe a wide range of recent experiences with the provision of information to study participants and review boards, and with identifying methodological alternatives to satisfy the concerns of review boards. Burt describes how procedures were adapted on a long-running family of surveys to address the new Standards for Privacy of Individually Identifiable Health Information (45 CFR Parts 160 and 164) that protect the sharing of medical record data, and the subsequent impact of these adapted procedures on response rates and data quality. Singer provides insight into the researcher-respondent interaction as introductory statements concerning the level of confidentiality change. Campbell describes limitations on creativity imposed on secondary analyses resulting from potentially overzealous protection of data access.[1] Dowd presents a successful, ongoing interaction with a review board regarding a highly sensitive longitudinal survey. The advice that each provides is for researchers to establish an interactive relationship with review boards so that procedures can be established in the best interest of the study participants and researchers.

[1] This paper was presented but not submitted for publication.

FEATURE PAPER: Incorporating HIPAA Privacy Rule into the National Health Care Survey

Catharine W. Burt, National Center for Health Statistics

The National Center for Health Statistics (NCHS) runs a family of health care provider surveys known as the National Health Care Survey (NHCS) to collect data from providers on patient encounters. The encounters sampled represent a broad range of service areas, from doctor visits and hospital discharges to nursing home stays. Sampled health care providers are asked to provide information for a sample of patient encounters to yield national estimates of utilization. While the surveys are authorized under the Public Health Service Act, which assures confidentiality in law, the newly effective Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) sets further standards for providers (i.e., covered entities) when disclosing protected health information for research or public health purposes.[1] The various surveys in the NHCS collect slightly different kinds of information and different patient identifiers.

This paper describes the ways survey procedures were modified to assist providers in participating in the surveys under the new regulations. The kinds of modifications varied across surveys but included obtaining or modifying Institutional Review Board (IRB) approval, creating data use agreements, completing accounting documents for disclosures made to assist the providers' record keeping requirements, creating new training materials for field staff, and new Web page materials for providers. The paper also discusses the process by which modifications were discussed and approved. The final sections address changes observed in response level and survey cost before and after the implementation date of April 14, 2003.

NATIONAL HEALTH CARE SURVEY & THE COLLECTION OF PROTECTED HEALTH INFORMATION

There are three major provider components to the NHCS: ambulatory care, hospital and surgical care, and long-term care. The various surveys within these components differ in regard to the type of protected health information (PHI) collected as specified by HIPAA's Privacy Rule (Table 1). Because the information collected differs, the level of review by IRBs differed prior to the Privacy Rule implementation. For example, the ambulatory care surveys rely solely on data already collected in medical records and no identifying patient information is collected, so they were exempt from IRB review for the protection of human subjects. However, because the long-term care surveys collect identifying data (e.g., Social Security number) to link to other databases, a full IRB review was required. The hospital and surgical care surveys collected a medical record number that could possibly, within the hospital, identify a patient, although outside of the hospital it was not an identifying piece of information. These surveys generally received an expedited IRB review. Because the Privacy Rule indicated that a full IRB review would be required if the surveys were to be used for research purposes, as we approached the Privacy Rule compliance date, we sought a full IRB review of all the provider-based surveys. Additionally, while NCHS requests the health care provider to complete or assist in record abstraction, in some cases the provider requests the data collection agent to abstract the data. When this occurs, the agent may see the patient's name and/or address, even though they are not collected. Under the Privacy Rule, this would be considered a disclosure of PHI. This situation might

[1] The Privacy Rule regulations are available at www.hhs.gov/ocr/hipaa/finalreg.html.

Table 1. Protected Health Information Collected or Planned for Collection in the National Health Care Survey Prior to the Privacy Rule

Data element	Ambulatory care	Hospital & surgical care	Long-term care
Birth date†	X	X	X
Encounter date†	X	X	X
ZIP code†	X	X	
Medical record number		X	
Social Security number‡			X
Medicare ID‡			X
Patient name‡			X

† Elements collected that are part of a limited data set.
‡ Elements planned for collection in the 2004 National Nursing Home Survey.

be interpreted as an "incidental" disclosure for which no accounting is necessary; however, the CDC's interpretation is that it is a disclosure for which the provider must account.

MODIFICATIONS MADE TO SURVEY PROCEDURES

With the publishing of the final rule in August 2002, NCHS staff evaluated what kinds of modifications would be needed to ensure that providers would continue to provide survey data and then developed and implemented such modifications. These included introductory letters, data use agreements, accounting documents, data modifications, and development of special training. In the case of the ambulatory surveys, survey protocols were developed to obtain a full IRB review with a waiver of patient authorization. The nature of changes to survey collection procedures follows.

Introductory Letter

The introductory letter was modified to include a paragraph indicating that there were several ways that the Privacy Rule allowed survey participation, including disclosure for public health purposes as well as for research, approval of the survey protocol by an IRB, and the collection of minimum necessary PHI to accomplish the survey objectives.

Question & Answer

Special Q & A's were developed to help address any concerns sampled providers might have about survey participation and their compliance with the Privacy Rule.

Accounting Documentation

Each survey created a one-page document that could be used by providers to assist in their accounting for disclosure requirements, whether the document was for each sampled encounter or patient or a summary accounting when more than 50 cases were abstracted.

Data Use Agreement

Because some surveys (e.g., the ambulatory care surveys) collected PHI designated by the Privacy Rule as part of limited data set (e.g., birth date, visit date, and residential ZIP code), survey-specific data use agreements were developed (where applicable) providing the necessary assurances of confidentiality.

Training

Special training modules were developed for field representatives to explain to the providers how they are able to participate and still be compliant with the Privacy Rule. In the case of the ambulatory care surveys, the training included a 30-minute PowerPoint slide show with audio that was distributed on a CD/ROM for field staff to view, which led them through the new procedures and survey materials, including a new chapter in the field manual.

Provider Materials

Special respondent Web sites were developed to display information about the surveys, including materials pertinent to the Privacy Rule (such as the IRB approval letters, Q & As, and data use agreements). In the case of the ambulatory care respondent sites, a seven-minute Flash presentation was developed to explain the survey and what the

provider must do to comply with the Privacy Rule.

Modifications to PHI Collected

There were two instances in which modifications were made related to the information collected: the hospital and surgical care surveys deleted the collection of medical record number so that the information collected met the definition of a limited data set, and the ambulatory care surveys allowed the respondents to enter patient age or month and year of birth rather than full birth date. The forms were not modified, but the field staff showed providers where to record age on the form if they objected to providing the full birth date. The other surveys already allowed the collection of patient age in place of birth date.

Other Modifications

There were several other modifications made to planned data collection activities to accommodate the implementation of the Privacy Rule. The National Hospital Discharge Survey (NHDS), which normally collected data from the previous year through April, changed the 2002 panel deadline to April 11, 2003, so that the providers and field staff would not have to worry about collecting data past April 14. NCHS also delayed a study to collect medication data in the NHDS from the spring until the fall of 2003 to permit hospitals time to become accustomed to reporting under the new HIPAA requirements. Finally, several of the surveys created toll-free telephone numbers for respondents to call in the event that they have any concerns about how survey participation is affected by the Privacy Rule.

REVIEW & APPROVAL PROCESS

After these modifications were developed, they were reviewed by Counsel at CDC in Atlanta, who ensured that the materials accurately reflect the regulation. There are still areas of concern among government agencies regarding the interpretation of some of the requirements, including the following: requiring individual accounting documentation when multiple records are disclosed for the same survey, and interpreting abstraction by our data collection agents as a disclosure rather than an "incidental" disclosure. Disclosures incidental to a permitted disclosure do not require documentation. For some of our surveys, this is the only disclosure that occurs; otherwise, the limited data set rule would apply.

EFFECTS ON SURVEY PARTICIPATION

While it is still too early to provide definite statements regarding the HIPAA Privacy Rule's effect on survey participation, we can state that no major participation problems were identified between April and December 2003. NCHS has received a handful of calls from doctors and hospitals expressing concern about the Privacy Rule and survey participation, but no effect has been noticed on general participation. The response rates for the ambulatory care surveys were essentially the same between January–March and April–July (NAMCS: 73% vs. 71% and NHAMCS: 95% vs. 91%). The field staff indicated that none of the hospitals' refusals in the second quarter were HIPAA-related. A National Nursing Home Survey pilot test in the summer of 2003 did not reveal any problems with participation, and the current research study on collecting medication data in the NHDS has not shown any compromise on hospital response.

Preliminary data collected from the 2003 NAMCS indicate that since the Privacy Rule implementation date, 82% of physicians accepted the data use agreement, and only 8% asked to see the IRB approval letter. Accounting documents were not left in a few cases in which our data collection agent abstracted the medical record data because the physician used electronic medical records or preferred office staff to place the document in the medical record. Anecdotal information also suggests that birth date and ZIP code may be missing more frequently from the abstracted data than it was in the past. Further analysis of the 2003 data will be required before definitive conclusions about the effects on participation can be made.

EFFECTS ON SURVEY COSTS

The creation of the new materials utilized about six full-time employees during December 2002–March 2003. Additionally, the field costs for the ambulatory care surveys ran higher than expected for the field staff due to extra time spent in training and the extra time spent in data collection necessitated by the explanation of the new Privacy Rule-related information to providers. The increase for the NAMCS and NHAMCS accounts for about 2% of the survey field costs. Budgeted costs for FY2004 also show an increase in the cost of data collection due to HIPAA.

SUMMARY

We believe that the preparation steps we took to clearly explain the interface between the Privacy Rule and participation in the NHCS, together with the new materials, helped to make a smooth transition from pre- to post-implementation.[2] Implementation of HIPAA's Privacy Rule has led to increased survey costs but has had less of an effect on survey response than originally conjectured. In some respects, the additional assurances of confidentiality made some providers more comfortable with providing patient data. However, the full impact on response rates will need to be measured over time as providers and survey organizations become more confident about the provisions in the Privacy Rule allowing continued survey participation.

[2] Helpful information about the Privacy Rule for health survey researchers planning record-based studies is available at http://privacyruleandresearch.nih.gov/research_repositories.asp and http://www.nahdo.org/memberaccess/webcall.htm. Privacy Rule materials used in the NAMCS and NHAMCS (e.g., data use agreements, IRB approval letters, questions and answers) can be found on our participant pages at www.cdc.gov/NAMCS and www.cdc.gov/NHAMCS.

FEATURE PAPER: Confidentiality Assurances and Survey Participation: Are Some Requests for Information Perceived as More Harmful than Others?

Eleanor Singer, University of Michigan

Breaches of confidentiality and their possible consequences pose perhaps the major risk of harm to social science research participants. To what extent are participants aware of this fact, and how does it affect their willingness to participate? A number of studies have documented the fact that confidentiality concerns affect participation in research (National Research Council, 1979; Singer, Mathiowetz, & Couper, 1993; Singer, Van Hoewyk, & Neugebauer, 2003). Further, if the topic of the research is sensitive, stronger assurances of confidentiality lead to more participation or better answers (Singer, Von Thurn, & Miller, 1995), whereas if the subject is not sensitive, such assurances reduce willingness to participate and/or increase perceptions of risk and harm (Berman, McCombs, & Boruch, 1977; Singer, Hippler, & Schwarz, 1992; see also Frey, 1986; Reamer, 1979).

Recent research indicates that the public overestimates the risk of confidentiality breaches associated with a given survey, but, given their perceptions of the risks, harms, and benefits involved, decisions about whether to participate are rational (Singer, 2003). Thus, the present experiment was designed to investigate whether attempts to strengthen "conventional" confidentiality assurances by adding a reference to a Certificate of Confidentiality would reduce public perceptions of risk and increase people's willingness to participate in the research and conversely, whether qualifying the conventional confidentiality assurance by adding a caveat—"to the fullest extent possible under state and federal law"—would increase perceptions of risk and reduce willingness to participate.

This research was made possible by funding by the National Center for Health Statistics to the Center for Excellence in Health Statistics at the Institute for Social Research, University of Michigan. I would like to acknowledge the assistance of Amy Corning in coding the open-ended responses and John Van Hoewyk in the analysis and to thank Mick Couper and Norbert Schwarz for their helpful comments on an earlier draft of this manuscript.

METHODS

Sample & Response Rate

The experiment was carried out by adding questions to two months (January and April 2003) of the Surveys of Consumers (SCA), a random-digit-dialed telephone survey of the national adult population fielded at the University of Michigan every month, primarily to measure consumer confidence. The monthly sample consists of interviews with 300 newly selected respondents, plus 200 re-interviewed after an interval of six months. The response rate to the survey was 58% in January 2003 and 60% in April 2003 for newly selected respondents and higher for re-interviews. The current study was carried out with first-time respondents; the Ns on which the analyses are based are 287 and 293, respectively.

Questionnaire & Variables

The questions for this study appeared at the end of the regular SCA interview, just before the demographic questions, and were introduced by interviewers as follows:

> *Now for something a little different. We are trying to learn how to better describe surveys to respondents . . . Imagine that the interviewer is talking with the respondent in person, in the respondent's home, and describes the first study as follows. . .*

Respondents were then presented with hypothetical introductions to one of two ongoing studies at the University of Michigan—the National Survey of Family Growth (NSFG) and the Health and

Retirement Study (HRS). Note that the data come from *respondents* to the SCA. It can't be assumed that they would generalize to nonrespondents as well; in fact, it is likely that willingness to participate would be lower, and perceptions of risk higher, among nonrespondents.

The introductions were very similar to the actual descriptions received by participants in these studies, but they tried to make the statements about risks and benefits as similar as possible. All respondents answered questions about both studies. A random half of the respondents were asked first about the HRS and then about the NSFG, and the other half were asked about them in the reverse order. Because of this, all analyses have been adjusted for the clustering of responses using the Jackknife regression procedure in *IVE*Ware (www.isr.umich.edu/src/smp/ive/).

Independent variables

The experimentally manipulated independent variable was the assurance of confidentiality given to respondents. In January, respondents received one of two confidentiality assurances, which differed only in whether they contained mention of a Confidentiality Certificate. Mention of such a certificate, which is obtained from the Department of Health and Human Services to provide additional protection against compelled disclosure of identified information, was intended to provide additional assurance of confidentiality to respondents, and we hypothesized that it would decrease their perception of risk and increase their willingness to participate in the study.

The relevant part of the introduction in the Control condition read as follows:

> *Your answers to our questions are used for research purposes only. Any information you give us will be kept confidential. The researchers who use our data see only statistics. We never give out names and addresses to anyone. You will not be individually identifiable in any reports.*

The introduction in the Certificate condition was identical, except that it added the following two sentences at the end:

> *To provide additional security, we have obtained a Certificate of Confidentiality from the Secretary for Health and Human Services. This certificate protects us from having to disclose the survey answers, together with respondents' names and addresses, to any person or organization, including the government.*

Because of increasing awareness of both legal and statistical threats to data confidentiality, there is a growing belief among researchers that absolute confidentiality should seldom, if ever, be promised to respondents. As yet, however, little is known about respondent reactions to such qualified assurances of confidentiality. Accordingly, in April, we repeated the January experiment, retaining the Control condition but substituting the following "Qualified" assurance of confidentiality, which is in fact being used or recommended by some survey organizations, for the Certificate condition:

> *Your answers to our questions are used for research purposes only. Any information you give us will be kept confidential to the fullest extent possible under state and federal law.*

Dependent variable

The dependent variable was willingness to participate, measured by a single question asked immediately after the description of the study had been read:

> *Please tell me how likely it is that you would take part in the survey I just described to you. Use a scale from zero to ten, where zero means you would definitely not take part and ten means you would definitely take part.*

This question was followed by a standardized probe, "Please tell me the reasons why you would (not) take part?"

Respondents who gave answers from 0 through 4 were asked why they would not take part; those who gave answers from 6 through 10 were asked why they would take part; and those who answered 5 were not probed.[1]

Intervening variables

Perception of risks. Four items assessed the perception of risks:

> *How likely do you think it is that each of the following people or groups would find out your answers to the survey questions, together with your name and address? Please answer using a scale from zero to ten, where zero means they will never be able to find out your answers, and ten means they are certain to find out your answers.*

Parallel items asked about four different groups: (1) family members; (2) business firms that might want to sell something; (3) employers; and (4) law enforcement agencies such as the IRS, the Welfare Department, or the police. Answers to the four questions were summed and averaged for a general measure of perceived risk. Note that there is no mention of the word "risk" in these statements.

Perception of harm. By "perception of harm," we meant how much people would mind if any of the four groups mentioned above gained access to their survey responses. The question was as follows:

> *Now I'd like to know how much you would mind if each of the groups I've just mentioned found out your answers to the survey, along with your name and address. Please use the same scale from zero to ten, where zero means you would not mind at all, and ten means you would mind a great deal.*

Answers again were summed over the four groups and averaged. In some cases, as noted below, we looked separately at sensitivity to individual groups. Of course, alternative definitions of "harm" are possible.

Perception of benefits to self and society. Benefits to society were measured by the following question:

> *Think again about the survey I just described. On a scale from zero to ten, where zero means not at all useful and ten means very useful, please tell me how useful each of the following groups would find the information from the survey.*

The question was asked about four different groups: the government agency sponsoring the survey, businesses planning new products, other researchers, and law enforcement agencies. Answers to the four questions were summed and averaged.

Benefits to self were assessed by the following question, asked immediately after questions about societal benefits:

> *Would you, yourself, get anything good out of the survey?* (Yes, No)

Perception of risks vs. benefits. The risk-benefit ratio was measured by a question that asked:

> *Taking it all together, do you think the risks of this research outweigh the benefits, or do you think the benefits outweigh the risks?*

We hypothesized that the Qualified introduction would increase perceptions of risk relative to the Control and the Certificate conditions and therefore reduce willingness to participate. On the basis of earlier research (e.g., Singer et al., 1992), however, we

[1] Responses to all open-ended questions were recorded verbatim and then coded. Reliability of the coding was assessed by having two people independently code a random sample of responses to each question. The percentage agreement on the detailed codes ranged from 77% for responses to "What would you, personally, get out of the study?" to 89% for the topic of the study described in the vignette, with most questions eliciting agreement of 85% or higher. Agreement on broader categories (e.g., on privacy as a whole vs. privacy of medical information, privacy of financial information, etc.) ranged from 89% to 94%.

expected that mention of a Confidentiality Certificate might increase respondents' perceptions of harm; if this occurred, mention of a Certificate might reduce willingness to participate. The precise ordering of the three conditions on willingness to participate would depend on whether respondents combined risk and harm additively or multiplicatively and whether they weighted them equally or weighed one more heavily than the other. We did not expect either assurance to affect perceptions of benefits.

Table 1. Predictors of Perceived Risk of Disclosure†

Variable	Parameter Estimate	Std. Error
Qualified	0.704	0.276**
Certificate	-0.365	0.276
Study (HRS)	0.062	0.101
Order	-0.072	0.101
Gender (Female)	-0.006	0.228
Education	-0.091	0.051*
Age (years)	-0.001	0.007
Intercept	5.189	0.867***
Effective *N*: 519		

* $p<.10$; ** $p<.05$; *** $p<.001$.
†OLS regression. We initially estimated all two-way interactions between assurance and the other independent variables then re-estimated after deleting nonsignificant interactions.

Table 2. Predictors of Perceived Risk of Disclosure†

Variable	Parameter Estimate	Std. Error
Qualified	-0.593	2.126
Certificate	4.102	2.043**
Study (HRS)	-0.533	0.121***
Order	0.339	0.121***
Gender (Female)	-0.100	0.237
Education	0.127	0.074*
Age (years)	0.020	0.009**
Qualified x Education	0.018	0.125
Certificate x Education	-0.195	0.130
Qualified x Age	-0.002	0.017
Certificate x Age	-0.034	0.018*
Intercept	3.554	1.276***
Effective *N*: 519		

* $p<.10$; ** $p<.05$; *** $p<.001$.
†OLS regression.

RESULTS

The results are quickly summarized. The Qualified assurance had the hypothesized effect on perceptions of risk: respondents in the Qualified condition perceived significantly more risk of disclosure than those in either the Control or the Certificate conditions, which did not differ from each other (Table 1). Mention of the Certificate of Confidentiality also had the hypothesized effect, in that it increased perceptions of harm relative to the Control and Qualified conditions, but only among younger respondents (Table 2). Perceptions of personal benefits did not differ by confidentiality assurance, but unexpectedly, perceptions of societal benefits did: respondents in the Qualified condition perceived significantly more societal benefits than those in the other two conditions, especially at low educational levels. Finally, in the Qualified and Certificate conditions, the ratio of risk to benefit was perceived as significantly lower among older than among younger respondents, whereas the reverse was true among respondents in the Control condition (Table 3).

Table 4, which replicates the results in Singer (2003), shows the effect of the intervening variables on willingness to participate. Risk, harm, and the risk-benefit ratio all have significant and strong negative effects on willingness to participate; perceived benefits for self and society have significant and strong positive effects.

The only other significant predictor of willingness is the study: respondents are significantly more willing to participate in the HRS, which is perceived as significantly less harmful than the NSFG and to have a lower ratio of risk to benefit. This result, which reverses the findings in Singer (2003), is attributable to the addition of linkage with medical records to the NSFG introduction. Apparently, disclosure of medical records is viewed as potentially more harmful than disclosure of Social Security records of

earnings, mentioned in the introduction to the HRS.

Beyond its effect on the intervening variables, however, the assurance of confidentiality adds nothing to the prediction of willingness to participate. Mean willingness scores are 4.11 in the Qualified condition, 4.34 in the Control condition, and 4.05 in the Certificate condition; these scores do not differ significantly from one another, even with the addition of controls for demographic characteristics.

Nevertheless, the reasons given by respondents for not being willing to participate are instructive. Whereas reasons pertaining to privacy and confidentiality are mentioned by 64.1% of those in the Control condition and 63.7% of those in the Qualified condition, they are mentioned by only 43.6% of those in the Certificate condition. Less than 2% of the sample mentioned protection of confidentiality as a reason *for* participation. Thus, it would appear that concerns about privacy and confidentiality can act as deterrents, but assurances of confidentiality cannot motivate participation.

Table 3. Predictors of Perceived Risk-Benefit Ratio†

Variable	Parameter Estimate	Std. Error
Qualified	1.007	0.586*
Certificate	1.798	0.708***
Study (HRS)	-0.203	0.114*
Order	0.058	0.114
Gender (female)	0.109	0.235
Education	0.113	0.033***
Age (years)	0.010	0.007
Qualified x Female	0.016	0.396
Certificate x Female	-0.721	0.458
Qualified x Age	-0.019	0.011*
Certificate x Age	-0.051	0.013***
Intercept	-3.103	0.604***
Effective *N*: 566		

* $p<.10$; ** $p<.05$; *** $p<.001$.
†Logistic regression.

Table 4. Perceived Risks, Harms, and Benefits as Predictors of Expressed Willingness to Participate†

Variable	Parameter Estimate	Std. Error
Study (HRS)	1.003	0.17***
Order	-0.089	0.169
Gender (female)	-0.320	0.251
Education	-0.053	0.061
Age (years)	-0.008	0.008
Risk	-0.102	0.046**
Harm	-0.161	0.044***
Societal benefits	0.161	0.054***
Personal benefits	1.445	0.260***
Risk-benefit ratio	-1.815	0.257***
Intercept	5.345	1.027***
Effective *N*: 548		

** $p<.05$; *** $p<.001$.
†OLS regression.

CONCLUSIONS

We conclude that even relatively subtle variations in assurances of confidentiality directly affect perceptions of the risk and harm of disclosure. A Qualified assurance increases the perception of risk, whereas mention of a Certificate of Confidentiality increases the perception of harm, especially among younger respondents. Perceptions of risk, harm, benefits, and the risk-benefit ratio all directly affect willingness to participate in a survey. But assurances of confidentiality, as operationalized here, have no direct effects on willingness, presumably because they have conflicting effects on perceived risk, harm, and benefits.

REFERENCES

Berman, J., McCombs, H., & Boruch, R. F. (1977). Notes on the contamination method: Two small experiments in assuring confidentiality of response. *Sociological Methods and Research, 6,* 45–63.

Frey, J. H. (1986). An experiment with a confidentiality reminder in a telephone survey. *Public Opinion Quarterly, 50,* 276–279.

National Research Council. (1979). *Privacy and confidentiality as factors in survey response.* Washington, DC: Office of Publications, National Academy of Sciences.

Reamer, F. G. (1979). Protecting research subjects and unintended consequences: The effect of guarantees of confidentiality. *Public Opinion Quarterly, 43,* 144–162.

Singer, E. (2003). Exploring the meaning of consent: Participation in research and beliefs about risks and benefits. *Journal of Official Statistics, 19,* 333–342.

Singer, E., Hippler, H.-J., & Schwarz, N. (1992). Confidentiality assurances in surveys: Reassurance or threat? *International Journal of Public Opinion Research, 4*, 256–268.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The role of privacy and confidentiality as factors in response to the 1990 Census. *Public Opinion Quarterly, 57,* 465–482.

Singer, E., Van Hoewyk, J., & Neugebauer, R. J. (2003). The impact of privacy and confidentiality concerns on participation in the 2000 Census. *Public Opinion Quarterly, 67,* 368–384.

Singer, E., Von Thurn, D. R., & Miller, E. R. (1995). Confidentiality assurances and survey response: A review of the experimental literature. *Public Opinion Quarterly, 59,* 66–77.

FEATURE PAPER: Human Subjects Issues in the National Survey of Child and Adolescent Well-Being

Kathryn Dowd, RTI International

This paper describes the human subject protection and research ethics experiences encountered in conducting the National Survey of Child and Adolescent Well-Being (NSCAW) from initiation through the 36-month follow-up. The research raised challenging issues for the IRB committees that required review, and the study team devoted considerable resources to addressing their concerns. The actual data collection effort has been monitored at an unprecedented level, with NSCAW interviews generating 215 incident reports. The lessons learned may benefit other projects, as IRB scrutiny of social behavioral research increases across the country.

PROJECT OVERVIEW

NSCAW is a national probability study of children in families investigated for child abuse and neglect. The NSCAW cohort includes children who had contact with the child welfare system selected from two groups: 5,501 from those entering the system through investigation for child abuse or neglect and 727 from among children who had been in out-of-home placement for about twelve months at sampling. These children were selected from 92 Primary Sampling Units (PSUs) sampled proportionate to size in 97 counties nationwide. The sample of investigated cases includes both cases that were receiving services and cases not receiving services, either because they were not substantiated or because it was determined that services were not required. This sample design required oversampling of infants, sexual abuse cases, and cases receiving ongoing services after investigation. The age of children was capped at 14 to increase the likelihood that youth could be located, a task made much harder when youth emancipate.

The study design includes three rounds of face-to-face interviews or assessments for children, their adult caregivers (e.g., foster and kin caregivers), teachers (for school age children not home schooled), and child welfare workers. As indicated in Table 1, there are four possible respondents for each case. Both those children who continue to receive child welfare services and those who leave the system are followed for the full study period. Nearly 68,000 interviews were completed through the fourth wave of data collection.

THE PROCESS OF OBTAINING IRB APPROVAL

From the outset, the issues presented by the NSCAW were anticipated to be quite challenging. To address these issues thoroughly, the project convened a group to recommend approaches to these challenges. The work group was aided by input from members of the NSCAW Technical Work Group and discussions with RTI's IRB Committee.

The study team began meeting with the RTI IRB Committee five months before the pilot study began. The committee identified a subset of its members to work closely with the study team to resolve issues of primary concern (described below). This subcommittee was engaged in the research design process to an extraordinary degree and met with the project director biweekly during the 15-month period prior to main study field training. In addition, the committee chairperson accompanied project staff to observe pilot study consent and data collection procedures in three households and monitored one call to an agency to report suspected ongoing child abuse. (A total of three reports were made during the pilot study.) The chairperson also provided support in obtaining approval from

Further information about the study design, sample design, questionnaire domains and content, and data collection procedures can be found at www.ndacan.cornell.edu/NDACAN/Datasets/Abstracts.

Table 1. Longitudinal Study Design: Sources and Modes of Data Collection

	Months After Close of Investigation			
Respondent	**~4 months**	**12 months**	**18 months**	**36 months**
Child	In-Person		In-Person	In-Person
Current Caregiver	In-Person	Phone/In-Person	In-Person	In-Person
Caseworker	In-Person	Phone/In-Person	In-Person	In-Person
Teacher	Mail		Mail	Mail

three site IRB committees and participated in conference calls with staff from the Office of Management and Budget's Office of the Special Counsel for Privacy regarding consent and assent form language.

Committees in three states (of the 36 states in which the study operated) required that their own IRBs review the study protocol. These committees varied in their procedures, foci, and concerns and required different levels of effort and negotiation. One state required submission of all adverse event reports indicating whether or not the event occurred in the state. This committee initially argued against the reporting of any suspected ongoing abuse or neglect, based on the concern that lay interviewers and survey data could not identify these situations accurately. Another state required reporting only if the event occurred in that state. One committee operates according to state (not federal) law, has very different requirements, and demanded the most effort and resources from study staff. For example, research involving human subjects is reviewed only once prior to initiation; no annual continuing reviews are conducted. This committee required customization of the consent and assent forms.

PRIMARY ISSUES

Agency Data for Sampling

The RTI IRB committee community member, who also served on the subcommittee, was very concerned about agencies providing data about completed investigations; these files formed the sample frame from which children were selected for participation in NSCAW. The families had not been informed of nor provided permission for release of sampling information to the study team. In a few states and counties, obtaining sampling frame files required considerable negotiation during the agency recruitment phase, although mutually satisfactory accommodations were reached in every site. The most serious adjustments in study procedures involved obtaining sampling data on only substantiated cases in two large states. (Weights were adjusted to statistically control for exclusion of unsubstantiated cases in these instances.) In the end, the committee accepted the decisions of the agencies regarding provision of frame file data.

Timing of Data Collection from Caseworkers

A second major issue discussed at length was the decision to attempt data collection from the investigative caseworker before contacting the family. Two factors contributed to this design decision. First, turnover among caseworker staff is extraordinarily high, and their caseloads are large. The study team judged that this timing would maximize the chances of locating the caseworker who conducted the investigation and of the caseworker remembering key aspects of the investigation. Second, the study team decided that, in order to maximize the likelihood of participation, families needed to be given approximately 60 days to process the investigation experience. We recognized that the experience varied tremendously across families depending on their particular circumstances. The 60-day time period was chosen as a trade-off between a need for "cooling off" for some families, the time required for foster parents to become acquainted with a child, and the improved chances of locating the family with contacting information from the agency.

However, these operational considerations and the planned timing of data collection from

the various informants caused concern among IRB committee members. While some of the data sought in the baseline caseworker interview centered on the substantiation decision process, the committee considered the data about the family's situation and the alleged abuse or neglect at the heart of the investigation to put the child and family at some risk, without the benefit of informed consent. The final judgment of the committee was to require the study team to retain and use only that caseworker data associated with families who participated in the NSCAW baseline, thereby indicating their consent. Note that the study team obtained permission from the IRB committee to use caseworker data about nonparticipating families for the purposes of adjustment of statistical weights and baseline nonresponse bias analysis, before the data were destroyed.

Risk/Benefit Ratio & Procedures for Gaining Cooperation

The calculation of the risk/benefit ratio for participants is at the heart of an IRB committee's assessment of the acceptability of a research project. NSCAW presented unusual challenges because of the variation of situations among children and families. With special protections of children in research, the assignment of risk had considerable consequences for gaining cooperation and data collection procedures and for the project's ability to obtain acceptable response rates.

From the committee's perspective, children already removed to foster homes were at reduced risk of further trauma than were children still in the custody of their "parents" (biological parents, grandparents, or other relatives) who could trigger another investigation and potential removal to foster care through certain responses to questionnaire items. Similarly, foster caregivers (who were not asked about child maltreatment, substance abuse, or domestic violence) were at reduced risk compared to custodial caregivers, who risked removal of one or more children and criminal prosecution through their family's participation.

While IRB committees generally assign risk to study participants at a global level, the committee reviewing NSCAW designated risk by respondent type specific to their circumstances. Children in foster care were determined to be at minimal risk with no special requirements beyond permission from their legal guardian (usually agency staff, the foster parent, or a family court judge, but sometimes the "parent" from whom they were taken). Children still in custody were judged at greater than minimal but less than substantial risk. The IRB committee required that both caregivers (if more than one resided in the household) be present for the explanation of the study prior to the informed consent process. An exception was allowed if the field interviewer detected one caregiver's concerns about the other's reaction to the study or concerns about intimate partner violence. (We recognized that contact with the family about the study could potentially trigger a violent reaction from an abusive partner, and the exception was granted to minimize that possibility.) While the study team was hesitant to burden lay field staff with these types of judgments, we designed, in consultation with the IRB committee, interviewer training modules to address these study procedures. Additionally, all consent and assent forms spoke to the limitations of promises of confidentiality in the plainest language possible.

Consent/Assent Forms

IRB committees devote much of their time to the detailed assessment of consent and assent form language. The purpose of this scrutiny is to assure that study respondents are adequately informed about the study and its purpose, burden, risks, and benefits. Both the specific content of the NSCAW forms and the language used to communicate the necessary elements were scrutinized in great detail. The forms underwent numerous reviews and countless iterations of revision, including changes necessitated by evolving human subject protection standards through the 53 months of data collection, changes in the composition of reviewing IRB committees, and the natural

progression of children and families through time.

One particular challenge faced was the need to standardize forms across sites to the greatest extent possible to maximize the comparability of results and produce truly national estimates. Achieving this objective required a great deal of negotiation between the various IRB committees reviewing the study protocol, necessitating a lengthy preparation period prior to implementation of the main study data collection. Customization of consent and assent forms and other study materials was necessary in only one state.

Embedded Reminders of Risk

As noted, the IRB committee determined that custodial caregiver respondents were at higher levels of risk for further actions by child protective services, and for negative social impacts if data confidentiality were breached and allegations of abuse or neglect became known in the community. The potential risks included another investigation, criminal prosecution, loss of employment, social stigmatization, and removal of one or more children from the home. To mitigate these risks, the committee required that reminders of the limitations on data confidentiality be inserted throughout the interview, especially immediately prior to the sections administered via ACASI (Audio Computer-Assisted Self Interview, where the computer "reads" the question and the respondent enters the response directly into the computer). The ACASI modules included those on substance abuse, involvement with the law, intimate partner violence, child injury, family tactics for resolving conflict, the child's experience of sexual abuse, and discipline and child maltreatment. The study team carefully negotiated the exact wording of these reminders to maximize the likelihood of accurate reporting of socially undesirable behaviors while still providing respondent protections.

Mandatory Reporting

One of the issues that required considerable thought and discussion was the obligation to breach respondent confidentiality to report apparent ongoing serious abuse or neglect to the appropriate authorities and to report to caregivers indicated suicidal intentions expressed by child respondents. Using an approach developed for the Longitudinal Study of Child Abuse and Neglect (LongSCAN), we narrowly defined "serious ongoing abuse" for the NSCAW baseline interviews, given the short time since the family had had an investigation conducted by a professional social worker.

By using a narrow definition, we were able to alert authorities to situations of a serious nature while not intruding on the process started by the child welfare investigation finished only weeks before the child and adult caregiver interviews. To define these threats more broadly at baseline would have put participating families at greater risk of losing custody of their children than nonparticipation, would have second-guessed the child welfare investigative process recently completed, and might have introduced a confounding intervention in a study that seeks to evaluate the very processes established to protect children. However, it was recognized that procedures in post-baseline interviews would need to be less narrow in definition because of the lag in time from the family's interactions with the child welfare system (assuming no further reports).

Using the definitions for serious ongoing abuse, the NSCAW study team identified questionnaire items that could elicit information requiring mandatory reports, developed scripted probes to help clarify the situation, and discussed ways field representatives were to interact with both respondents and local child welfare agency staff in mandatory reporting and other distressing situations. Items most likely to elicit reportable responses, as well as any scripted follow-up questions that were administered as a result of positive responses to potential report-triggering items, were generally placed in the ACASI modules of the instruments, which provides a more private setting to respond to the questions and minimizes field representatives' involvement in any resulting mandatory reports. Additionally, field

interviewers may report based on their direct observations while in the household.

The ACASI probes were designed to collect additional information on the frequency and recency of a positive report of maltreatment and whether the alleged abuse involved an adult caregiver living in the household. Field representatives were not notified about positive responses to mandatory report items triggered; instead, the data were transmitted to RTI daily, reviewed by members of the project team, and decisions made about the necessity of a report based on the responses to the interview questions, report probes, and guidelines established with the affected site. Mandatory reports were filed by project team members, in accordance with the procedures established with the individual sites. Copies of all reports were provided to the RTI IRB committee and, as required, to IRBs in two participating states.

An ACASI adaptation was developed for the administration of a cartoon-based measure of exposure to violence, as these questions are asked of children age 5–11 who do not get the more extensive ACASI module. The adapted method involves children wearing headphones and listening to the questions read by the computer and indicating their responses to the interviewer on a card so that she may enter them into the computer. The modified procedure was developed after staff administering pretest interviews noted that it is difficult to maintain a private interview setting when interviewing children this young.

Calls to authorities to report suspected abuse were made by study staff in the central office so that the appropriate context (e.g., responses to survey questions versus directly observed behavior or directly communicated information) could be conveyed in a standardized manner. Further, we minimized interviewer involvement in the process in order to both reduce the stress and potential legal entanglements (e.g., need to testify in court) and to avoid unconscious differential behavior possible if interviewers became aware of ongoing problems in the household. This arrangement is not wholly satisfactory to the agencies, which want as much specific information from direct observers as possible. However, all agencies initially accepted the study protocol as developed.

In the third year of data collection, a change in reporting procedures was implemented as a result of increasing centralization of reporting in participating agencies. Moves to state hotlines for reports greatly complicated the process, as we could no longer deal with one or two agency staff regarding mandatory reports nor rely on their understanding of the study protocol. In reporting to centralized hotlines and staff unfamiliar with the study, we now provide much more information about the study, as well as the specific questions that triggered the report and caveats about the nature of survey data (e.g., measurement error, respondent error). This new process was worked out in detail with the IRB committees.

Suicidal Intent

In addition to mandatory abuse and neglect report situations, the NSCAW interview for children as young as seven includes several questions about suicide ideation and intent, including probes about suicidal thoughts in the past two weeks and presence of a plan to commit suicide. In cases where the answer to both items is affirmative, the field representative is alerted at the end of the interview by the CAPI program and required to take steps in response to the situation. These include telling the child that his/her parent or caregiver would be told about the situation and reporting the situation to the parent or caregiver. Children were given the option of being present for the discussion with the parent/caregiver. The field representative was not allowed to leave the household until the report was made to the parent/caregiver, who was encouraged to alert appropriate mental health professionals or another service provider and was given a resource list. Field representatives were trained to handle spontaneous reports of suicidal attempts or threats that were not in response to interview questions, including those from adult respondents, in a similar manner.

Data Confidentiality & Release

One of the primary objectives of the Administration on Children and Families is to

make NSCAW data available to the research community for secondary analysis. The dataset is so rich that project resources could never adequately cover all the possible uses of the data across various fields of research. However, fulfilling this objective has been challenging, given the detailed and longitudinal nature of the data. The IRB committees and OMB required extensive assurances that data release procedures would not pose undue risk of re-identification to study participants. The study team consulted extensively with statisticians regarding statistical disclosure analysis techniques and with staff at the National Center for Education Statistics and the Committee on Data Access and Confidentiality (formerly the Interagency Committee on Data Access and Confidentiality Group) on data release strategies.

The process eventually agreed on for use with NSCAW data involves a tiered approach and licensing agreements through the National Data Archive on Child Abuse and Neglect (NDACAN) at Cornell University. The data are available at three levels of specificity: (1) a general release that has been reviewed carefully for disclosure risks and some variables deleted or recoded; (2) a restricted release containing all variables except identifying data such, as names, addresses, and agency names; and (3) an RTI restricted release that allows researchers to merge data from other sources to the NSCAW data and obtain a merged dataset of a subset of NSCAW variables plus their data. The levels of release have different requirements for obtaining a user's license based on the risk of re-identification. The general release requires only a completed application, IRB approval from the researcher's institution, and payment of a small administrative fee. The restricted release demands a completed application, IRB approval from the researcher's institution, approval of a data security plan, and payment of a $2,500 fee that covers one site visit to monitor compliance with the data security plan. The RTI restricted release mirrors the restricted release, with the exception that approval by the RTI IRB committee also is required.

CONCLUSIONS

The study has indeed succeeded past all the obstacles encountered to date, through patience, careful negotiation, and persistence. Sampling from all agencies was completed as designed. Acceptable response rates were achieved in the baseline, and response rates in the mid to high 80s have been achieved in post-baseline waves. The 36-month follow-up data collection was completed at the end of February 2004, with the highest response rates among the post-baseline waves. All agencies recruited have continued participation through the more than four-year duration of the study. A total of 215 incident reports have been submitted to the IRB—a total of 149 for suspected ongoing serious abuse based on questionnaire responses (16 from caregivers and 133 from children), 62 for child suicidal intent, three based on field interviewer observation, and one because of extreme distress experienced by both the child and caregiver respondents. No breaches of confidentiality have occurred. Data have been released to the research community, and 25 licensing agreements are in place.

The study team and the RTI IRB committee developed a close working relationship, and the committee has served as a resource of enormous benefit to the project. The NSCAW team has been contacted by other researchers across the country for information and assistance. Project consent and assent forms have been shared widely with others.

It is unclear whether NSCAW could successfully be initiated in the current environment, with the additional constraints imposed by implementation of HIPAA and other privacy protections such as state-level data security policies. NSCAW was fortunate in the timing of study initiation and in the level of dedication and commitment to the study objectives in participating organizations. The reviewing IRB committees generally have been very helpful and have demonstrated their interest in facilitating solid research while simultaneously ensuring that we protect the rights and safety of research participants.

SESSION 5 DISCUSSION PAPER: Security and Privacy: What Are We Doing Wrong?

Brad Edwards, Westat

This is the smallest of the five sessions at the conference, with four papers instead of five, but it hits some very important issues. I enjoyed all the papers. I will make some general observations, comment on each paper, and conclude with some speculative remarks.

As survey methodologists, we often chafe against privacy regulations imposed from outside the field. After all, there has been a strong tradition of confidentiality from the earliest days of survey research. The field is built on the assumption that the identity of individual respondents will not be revealed. When I began conducting surveys in the 1970s, I saw it as work that was somehow "purer" than most, untainted by special interests, with lofty goals of scientific advances and social good. Looking back on my youth, this seems naïve today, but I think in some way it is still a vision we share: discovering and reporting the truth, by keeping identity secret and sacred.

That is an important part of what I learned when survey research was a craft. It has become increasingly professionalized, with schools that grant degrees in the field. Many processes have become automated and transformed in the information revolution. If you consider surveys in the broadest sense, there are many more now than ever before, and not just expensive government-sponsored health surveys, but polls, assessments, and market research. The public is overwhelmed (and unimpressed, to judge by declines in response rates in the past decade, especially in phone surveys). Nonetheless, there has been no monumental breach of security or privacy in survey research. What are we doing wrong?

The concept of privacy in the American experience is closely related to the concept of property. An important part of privacy is control over your own back yard or "your space." During the past 50 years, with the explosive growth of information in general, personal information increasingly has come to be viewed as property. More and more people are unwilling to give away information about themselves freely, if at all. In an age when communication speed is accelerating, along with the volume of sensory input and data, we are continuously bombarded with "clutter" and strive to protect the most important information—our identities.

Despite the proliferation of surveys today, the field is still a little fish in a big pond. Our work exists in the much broader contexts of law, of medical research, and in the culture as a whole. And in those contexts, there have been sweeping changes in the past 30 years, following the end of the Tuskegee Syphilis Study. Exhibit A is a timeline that starts with that study, the best known example in the U.S. of research run amok, ruining lives without checks or accountability. Essentially a medical experiment, it still casts a long cold shadow over all research with human subjects in this country.

The Privacy Act of 1974 provided a broad legislative umbrella of protection for participants in federally-sponsored research. Five years later, the Belmont Report and the regulations that eventually became known as the Common Rule established IRBs, providing more specific protections for vulnerable populations and bringing health studies (including surveys) into an oversight structure of boards with broad discretion to implement general guidelines.

Just in the past three years, there have been a spate of major developments that affect the privacy of survey responses. For example, updates and expansions in the Common Rule; the passage of the Patriot Act in fall 2001 (which, among other things, made Department of Education surveys subject to search by the Attorney General); HIPAA, imposing privacy regulations on health care

Exhibit A. Law, Medicine, People

1930	1940	1950	1960	1970	1980	1990	2000
1932		*Tuskegee Syphilis Study*		*1972*			
				1974 Privacy Act			
					1979 Belmont Report		
						1991 Common Rule/IRBs	
							2001 Patriot Act
							2003 HIPAA
							"Do Not Call" List

providers and other holders of patient information; and the lightning-like passage of the Do Not Call list in Congress. The culture is increasingly alarmed about identity theft, and people seem to want to be left alone.

My own interest in privacy issues was heightened by an experience that began about five years ago with an unusual study that involved sampling newborns from birth records. Working with the National Center for Health Statistics, we developed a plan to negotiate with about 50 birth registration areas (mostly states, but also some cities and Indian reservations) to access the records.

Looking back on the experience, I think of this now as a large case study of what can happen when you go to fifty different entities with the same protocol, proposing to collect data from a vulnerable population, and ask for approval. NCHS was involved in a similar effort in the early 1990s and encountered some difficulty, but nothing insurmountable. We expected similar results a decade later, but these expectations were not quite met.

We found no restrictions in about 80% of the areas. But seven areas had restrictions imposed by law, and in three areas, IRBs imposed restrictions. We tried to negotiate solutions in all areas but were unable to reach a satisfactory solution in six of them. In what was perhaps the worst case, a dysfunctional IRB required four submissions and two personal appearances but never issued any written questions or findings. We withdrew the request when the children would have aged out of eligibility for the study, but the IRB contacted us two years later to ask if we would like to continue our request. We politely declined.

Five areas that did participate required changes in the protocol of one kind or another. This ranged from unique advance letters and consent forms, through prior passive consent (i.e., a postcard was sent to sampled families giving them the option to decline participation), up to (in one area) a severe prohibition on any refusal conversion activity.

This entire experience was not all bad. It resulted in some improvements in the protocol, and although we feared a negative impact on response rates, we were not able to detect any effect for many of the changes. The process was costly and took a lot of time. The whole experience calls into question the viability of national "followback" studies from vital records, given the patchwork of community standards that currently exists.

I've also been fortunate to participate in a small part of the effort that Burt describes so well in her paper on the National Health Care Survey (NHCS) program's response to the recent HIPAA legislation. As Westat's project director for the National Nursing Home Survey (NNHS), which is a component of the NHCS, I was able to observe the process unfold at NCHS last year.

NCHS/CDC determined that, for their agency, disclosures are allowed for public health purposes. Not every survey can fit under that rubric, and it only works if the provider's information to the patient about disclosure cites this condition, but it seemed to work on the NNHS field test. As Burt reported, we didn't see any major

participation problems when it was conducted last June and July. Studies that were in the field at the time HIPAA took effect in mid-April monitored response closely. On the MEPS Medical Provider Component's (MPC's) work with hospitals and on the Medicare Current Beneficiary Survey's long-term care facility component, we saw little or no impact on participation at Westat. The MEPS MPC's physicians were more problematic, but there was a sense last year that (as Burt notes) some providers may just have needed some time to become accustomed to the reporting requirements.

The NCHS approach (adapting materials for HIPAA, developing additional communication and training materials, and seeking HIPAA-specific IRB approval) seems quite sound. I found myself wondering if the IRB letter of approval worked like the Certificate of Confidentiality in Singer's experiment.

Singer's paper builds on a considerable body of work on informed consent in surveys. I think it is great to see this empirical data on the effects of introductions with different confidentiality and consent content. Given the legal tradition of the Common Rule and IRBs, there has been very little cognitive work on informed consent procedures. I often have been struck by how little many IRBs know about the survey participation process—in the case of face-to-face interviews, usually you are invited inside in the first minute or you are never invited in. It is only after you get inside and sit down that you have the opportunity to present the formal informed consent procedures. By then, most respondents have already agreed to participate. It is hard for most respondents to change their minds at that point without seeming inconsistent. Of course, the advance letter may incorporate all of the informed consent elements, but how often are the letters opened, read, and remembered?

Variation in IRB attention to informed consent also comes with survey mode. IRBs often spend a lot of time on the issue for face-to-face surveys but may ignore the issue for telephone surveys, or worse, impose a page of text to be read to the respondent before any interaction takes place between the interviewer and respondent. Since the decision to participate in a phone interview is usually made in the first five seconds, there is little opportunity to address informed consent issues in that mode before consenting, especially on random-digit-dial surveys. Mail also presents some interesting issues, since the respondent has the opportunity to view the entire data collection request in all its detail before deciding to respond, and by responding, documents his or her consent.

I would like to see a future experiment on the relationship between the certificate of confidentiality and willingness to participate, as Singer suggests.

Campbell[1] describes another important aspect of privacy and security for health surveys—confidentiality and public data files. He points to a power vacuum created by the wide variation in IRBs and the lack of input from the research community. This points to a structural difficulty with the U.S. system: the lack of a central statistical agency that could tackle this issue with a strong voice. NCES has provided some leadership in this area, developing a system of confidentializing public files and providing restricted access to a more complete version of the data set (e.g., with greater geographic detail or household composition data).

Campbell talks about variability among providers in approaches to public data files, but it seems to me the variation could be explained by study differences rather than house differences. For example, a major focus of the GSS is opinion data, which have no independent source that might be crossed with the survey data to identify respondents. On the other hand, a health survey that captures data on conditions and includes a fine-grained level of detail on geography risks disclosing the identity of people with relatively rare illnesses (like HIV) in counties with small populations.

There is a trade-off between access and risk of disclosure, as Campbell so rightly

[1] As noted in the Session 5 introduction, this paper was presented but not submitted for publication.

points out. Risks must be assessed in practical terms. Has the identity of any health survey respondent ever been discerned in a public use data set in the U.S.? Could the public data file be matched to another public file that would allow identification of a subject? Could a respondent or family member identify the respondent's own data in a survey? What is the risk of such an occurrence?

Do we need to worry about the possibility of additional data sources becoming available in the future that might be combined with the study's public use file to identify subjects? This is what Campbell calls re-identification. It is not far-fetched. I've encountered a couple of examples in the past decade where this was a risk. In one case, it caused a last minute change in dissemination procedures, with much more restricted access to a public data file.

Dowd's experience is closest to my own, in terms of the effort that often is required to clear complex sensitive surveys through IRBs and the kinds of problems encountered. She was dealing with a rare case—a survey that truly presents higher than minimal risk. The IRB process can be costly in such cases, but this is an example where it improved the survey protocol.

I also have had special difficulty with IRBs that aren't operating under federal law—in my experience, the process was much more political. It can present a considerable challenge to the study design when the human subject protection standards are in flux. IRBs have the right to revisit issues they may have decided years earlier on a continuing study. Different members may come onto the board and choose to review something the previous members approved.

One notion I've heard more than once from IRBs is that the survey interview somehow might upset a respondent so much that it would cause lasting psychological damage. If only most of our interviews were that meaningful and/or powerful! But in Dowd's case, the IRB had some reason to be concerned about the potential impact of the interview. Dowd also talks about a tension in reporting incidents to authorities, between the authority's desire to hear from the most direct source (the interviewer) and the study's need to provide the story in a standardized way. The most telling comment in her paper: it is unclear the National Survey of Child and Adolescent Well-Being would be successful in today's more difficult human subjects environment.

WHAT DOES THE FUTURE HOLD?

There is a trend toward health surveys incorporating other data sources (DNA, medical records, etc.). This trend moves us deeper into the stream of bioethics concerns. We are there, whether we like it or not.

Statements of best practices and dissemination of information about practical survey responses to human subject protection can only help. But we are in a world where highly sensitive surveys are in jeopardy. The majority may make it through the laws and regulations, but I think we will see increasingly tighter restrictions, and some studies will not make it off the ground. This session should serve as a call for much more empirical work on the effects of informed consent procedures, and the results need to be disseminated to the IRB community. The overarching goal of the health survey researcher—to design and implement sample surveys that can represent a population—are increasingly at odds with the trend toward greater privacy protection.

Just in the past five years, the Internet has transformed the way we access information and the amount of information available. Data sets that once were available only through hardcopy files in public agencies are increasingly online and searchable, making gigabytes of personal data part of the public domain and immediately available to anyone. For example, I understand one can access detailed marital history files for the State of Kentucky, so if marital history is part of the survey data collected, just knowing a Kentucky resident was part of the survey sample and knowing the individual's marital history (something any friendly neighbor might know) would allow a confidentiality breach. The Internet puts our personal

information out there and subjects us to increasing risk.

We have witnessed a powerful political phenomenon recently with the passage of the legislation authorizing the Do-Not-Call List. Americans do not like marketing calls intruding on home space. One or two calls a year might not be a problem for most people, but several calls a day are far too many, and 57 million Americans registered almost immediately for the list. The aggressiveness of the market sector in using the telephone system to sell products and services has finally been checked. But the phenomenon reflects how far we have come in the commercialization of private space.

It has become a routine aspect of American life to give up "private" information in exchange for some private benefit. For example, virtually every product for retail sale carries a bar code, and there is a scanner at every checkout counter. My supermarket chain offers discounts on many products for scanning my own barcode, which allows the grocery chain to match my buying behavior with my personal characteristics. I like the savings, and the supermarket likes the information I give in exchange. It is similar to the "welcome basket" most of us received when we checked into the conference hotel. In return for registering our preferences and a little personal information, the hotel chain designates us frequent visitors and gives us a personal gift.

I think the American public is signaling a need to renegotiate the survey contract. Declining response rates and increasing privacy concerns suggest that the traditional survey appeals to civic duty and the general public good do not work, and "token" incentives are less effective than they used to be. We need to find ways to individualize the exchange, and we need to offer more sophisticated benefits that are immediately perceived as worth something to the respondent. Survey methodologists rarely focus on these tasks, but I think we need to take a cue from marketers and make our surveys less obtrusive and more fun. There is a lot to be said for creating "interest-getting" questions and motivation-building designs. Most surveys may not pose more than minimal risk to respondents, in the sense that they are no more likely to harm respondents than is an everyday conversation. However, they are far different from everyday conversations, in the sense that they can be perceived as a waste of respondents' time—in other words, boring. We need to re-examine the risk/benefit ratio in this light.

SESSION 5 DISCUSSION PAPER: Back to the Drawing Board: Reactive Methodology

Joan E. Sieber, California State University, Hayward

Many cultural elements of the research environment create barriers to valid social and behavioral research. Not surprisingly, cutting-edge research often focuses on some current aspect of our culture and on development of methodology that responds to cultural impediments to valid research. I shall refer to such methodology as *reactive methodology*. "Reactive" refers to reactions to the research by respondents, their community, researchers themselves, and other environmental forces, such as the federal regulations of human research and Institutional Review Boards (IRBs). Reactive methodology seeks to resolve conflict between these players. Viewed in historical perspective, we see that reactive methodology has been increasing rapidly for the last 35 years and is unlikely to diminish.

In the 1950s and early '60s, social scientists sought to improve their status by emulating the hard sciences and behaving as though research participants were inert (nonreactive) substances. The implication was "grab the data and run." Social scientists assumed that they had no relationship with research populations. This approach began to diminish in the late '60s, when social scientists followed the funding for applied social research on such issues as early education, drug abuse, racial and inner-city concerns, and later the HIV epidemic. Upon venturing out into that world, social scientists discovered who they really were—investigators of dynamic human culture. They began learning to interact with context-specific, dynamic, reactive social phenomena, and recognizing that they bring their own baggage and are part of that dynamic. Since then, social scientists have been adjusting their methods to obtain valid data by creating relationships with subjects and their communities, including leaders, advocates, and experts (e.g., Melton, Levine, Koocher, Rosenthal, & Thompson, 1988). Later the need to learn to relate effectively to IRBs became apparent.

It turns out that social research is highly contextual. Much of the research in the social sciences is affected by scientific or ethical issues due to the following:

- The culture of researchers and participants and how their assumptions interact to affect the validity of the research and the applicability of the findings.
- The larger environment in which the research occurs. For example, "human research" to the layman, the legislator, and even the IRB often connotes biomedical research and the model of informed consent that implies. "Survey research" often connotes marketing research and telephone sales to the layman. These confusions produce misunderstandings and restrictions that hinder bona fide survey research.
- Conflicts between the goals of science (to gather valid data) and the values of society (e.g., to protect the privacy of respondents). As these conflicts are recognized by the media and the public, the government attempts to resolve them by imposing regulations, such as 45 CFR 46, HIPAA, and CIPSEA.
- The difficulty of studying people who have something to hide or who fear "the government" and research funded by government agencies.
- Research in cultures that the researcher does not fully understand.
- The legal and regulatory context of human research and how these are interpreted.

Ethics, broadly defined, is at the heart of reactive methodology. Ethics is about supporting important social values, such as respecting people and their communities and benefiting individuals and society. In the present context, this includes doing valid research, protecting and respecting research participants in all the ways that pertain to the

particular context, creating socially beneficial policies, and disseminating and applying findings effectively. There is no simple rule or formula that produces ethical social research. Each situation is unique. For example, when doing research on the health needs and health care of senior citizens who own homes or rent apartments in a senior community with assisted living accommodations, it makes sense for the researchers to present their project at the community center. They also might have it described (along with a photo of the researchers who will come to elderly people's doors to interview them) in the community's newsletter. Then, they might phone or write to make an appointment for a face-to-face interview. All of these steps will provide assurance for seniors who may fear the possibility of admitting to their homes someone who might harm them. These steps also will help clarify the situation for seniors who suffer mild dementia. The advantages of such a sensitive procedure are manifold: the researcher respects the needs, dignity, and vulnerabilities of elderly research participants. In response to the excellent communication comfort level that these procedures create, the researcher is able to develop a valid random sample of the community members and evokes candid and open responses from those interviewed. The results of the survey along with useful information might then be disseminated back to the members of this community—again via community meeting and community newsletter. These kinds of advance informing of prospective subjects would hardly be necessary or appropriate when surveying health needs of students living in a campus residence hall, although feedback via a residence hall meeting or newsletter probably would be highly appropriate. At the extreme other end of the spectrum, the freshman who uses his mother as a subject when conducting an N=1 interview surely need not compel her to sign a consent form before undertaking the interview.

Regulations are enacted in response to problems, with the goal of producing more ethical behavior. However, regulations are not to be confused with ethics! While ethical decision making should take into account the unique characteristics of the situation at hand, regulations are specific requirements—perhaps worded as though one size fits all. Most regulations, including 45CFR46, which governs human research, are written in such a way that they may be flexibly interpreted. However, that doesn't always happen! There are unintended consequences of regulations that create problems for social scientists. Regulations are enacted when social values are violated, the media become involved, and politicians regulate behavior to align it with social values. Regulations of human research enacted in response to breaches of public trust primarily in the biomedical sciences then may pose unintended and costly barriers to ethically conducted social research.

The studies of these barriers described in the papers in Section Five and throughout this conference are at the vanguard of social research methodology, for no amount of statistical or design elegance will enhance the power or validity of research when contextual forces are at work to deny the researcher valid data or any data at all. Every paper here has addressed issues of ethics, broadly defined, and all are contributing to a vibrant new reactive methodology. Yet, these papers are only scratching the surface of reactive methodology, and there currently exists no way of systematizing this methodology.

Needed is an organized body of knowledge that will enable researchers to predict, examine, and understand these reactive phenomena and to design more efficient, effective ways to gather valid data. Fortunately, the theories and methods of the social sciences are ideally suited to understanding how ethical issues arise and how methods may be created that will solve these problems.

How might those who do human research and develop reactive methodology develop a systematic, empirically based literature on ethical problem solving in human research? How might they arrange to share the useful tools they create? Those tools include such items as researcher training, approaches to

obtaining effective informed consent, approaches to increasing the benefit of research to subjects, communities and society, methods of ensuring that shared data cannot be re-identified to breach confidentiality, and so on. How might this empirical literature become the basis of better-informed IRB decisions, replacing the idiosyncratic and varied decisions that IRBs render based on anecdote and conjecture?

A PEER-REVIEWED ONLINE JOURNAL OF EMPIRICAL RESEARCH ON RESEARCH ETHICS

A quarterly international journal that will be concerned with all human research is on the drawing board. The first issue is planned for January 2005. It will have a prestigious international review board. Low-cost institutional subscriptions will be available to everyone on the institution's server; a sliding scale of subscription rates will accommodate poor institutions and poor countries. All issues will be archived and indexed so that researchers, students, IRBs, and others can readily access prior issues and locate articles of interest to them. (Contact the editor at jsieber@bay.csuhayward.edu for an editorial policy statement.)

The first quarterly issue of each year will be devoted to reviews of specific literatures. The other three issues will be devoted primarily to current empirical research on topics of research ethics. Table 1 presents the taxonomy of expected topics and indicates the conference papers that touch on each. Within these topics, literature reviews and empirical studies will reflect newly emerging research issues due to differences between cultures, topics, sensitivities, modes of research, technologies, legal and regulatory climate changes, and events within the global community.

TOPICS IN REACTIVE METHODOLOGY

The research reported at this conference indicates healthy growth of reactive methodology and promises more to come. Noteworthy, however, are the categories in Table 1 that were not addressed and that may emerge as important in the future of health survey research.

Deception includes topics that concern health survey researchers. Given the difficulty of fully informing respondents, omissions and self-deception need to be better understood. Debriefing rarely is exploited fully as an opportunity to learn what respondents think about the research experience. It is an opportunity to append cognitive interviewing onto the preceding research and to conduct other kinds of research into such issues as how respondents viewed the risks and benefits of the research participation.

Privacy – respondents' interest in controlling the access of others to themselves – is not understood adequately in relation to the survey researcher's desire to create a mutually satisfactory conversation. Privacy interests can be negotiated in relation to whether the researcher proffers benefits that cause respondents to want to grant access to themselves. What do researchers need to understand about a population to bring about respondents' desire to grant access to themselves? This question can be answered in a variety of venues, including community consultation and partnerships, focus groups within the research population, surrogate subjects, and debriefing interviews.

Issues of *taboo and controversial research* sometimes arise in the United States in relation to religious objection to federally funded research on topics concerning sexual behavior. In repressive regimes, there often is suppression of research that reflects negatively upon the government. In the global economy, where international trade and tourism bring the health problems of exotic cultures to Western society, suppression of information about health problems raging out of control ultimately results in costly quarantine of individuals and restriction of trade. How can the global community of health survey researchers foster openness, respond constructively to controversy, and emphasize to offending governments the ultimate costs of suppression of health research? Innovative efforts to promote greater openness to sensitive research and

Table 1. Taxonomy of Reactive Methodology and Papers that Address These Topics

Group	Topic	Papers
Communication with Subjects & Community	***Communication*** processes between researcher & participant, organization, community, or other entity. Informed consent, modes of communication, comprehension, trust, decision making, competence of subjects, context, etc.	2A, B,C,D,E; 3A,B; 5B
	Deception, intended, unintended, self-deception, perceived, concealment, mental reservations, omissions; evaluation of; desensitizing & debriefing; consent to be deceived, waive informing until later, efficacy of alternatives.	5B
	Cultural sensitivity, norms, language, meaning equivalency, cognitive interviewing, community consultation, partnership, advance planning, etc.	2A,B,C,D
	Relationships as a source of qualitative data, nature of relationship.	
Acquisition & Use of Data	***Privacy*** relative to stage of human development, learning, culture. How IRB & scientists' views on what is private may differ from that of subjects.	2E, 3B
	Confidentiality & relationship to what subjects will divulge, how researchers can keep confidentiality promises; methods of preventing breach.	2D; 5B,C
	Dissemination & data sharing, modes of dissemination; risks, responsible use, emphasis, omission, misinformation, data suppression or censorship, role of mass media; obligation to contribute to scientific literacy of society.	2A; 5A,C,D
Outside Influence on Research	***Government regulations***, interpretation, creative application, effects on research, how empirical evidence can influence the regulatory process.	5C
	IRBs, how they function & may be improved, their effects on research.	1B; 5A,C,D
	Taboo & controversial research that may cause harm or misleading results; whether research should be censored or results suppressed.	
	Scientific integrity, kinds of misconduct, causes, prevention; including faulty research design & statistics.	
	Ethics & politics, how political differences may underlie charges of scientific irresponsibility; controversial sponsorship & employment; classified research.	
Risk & Benefit	***Risk, wrong, & harm***, how evaluated, perceived by researchers, IRBs and subjects; how is safety judged; how procedures can be made safe	2E
	Benefit & promise of research, how they may arise, be estimated and maximized for subjects, communities, institutions, science, and society.	1C,D,E; 2A,B; 5B
	Risk/benefit, when knowledge may be gained at some risk; how justified?	
Theory, Method, & Design	***Epistemology*** – Is knowledge a form of intrinsic good or of power? Is it "out there" or a construction of the interaction between researcher & participant? What are the relative merits of various approaches to knowledge?	
	Validity, balancing rigorous design with satisfaction of other ethical concerns. Making valid designs more efficient to answer a wider variety of research questions; concepts of measurement & comparability over time.	1A,B,C,D,E
	Modeling. Theories/methods of modeling to improve accuracy & sophistication of prediction and empirical tests of complex ideas about health.	1C
	Equitable treatment of subjects, distributive & procedural justice in research planning, conduct, & application (e.g., use of placebos; withholding sponsorship information from subjects; selection of subjects); stratification, demographic challenges (e.g., mixed ethnic persons).	2C, 3A,D
	Technology, efficiency, & sampling, uses of new technologies to reach subjects, combinations of technologies, assumptions underlying such new designs.	3E, 4A,B,C,D,E
	Language & meaning. How can differences in language & meaning within & between cultures be bridged, yielding valid & comparable results?	3A,B,C,D,E

Key to Table 1: System for Identifying Papers
There were five sections, designated below as 1, 2, 3, 4, 5. Within each section, there were as many as 6 papers, designated by letter (e.g., A, B, C, D, E, F). Thus:
1A Wadsworth, 1B Correa, 1C Wolfson et al., 1D Ezzati-Rice, 1E Gfroerer et al.
2A Brown, 2B Cornelius, 2C Call et al., 2D Krauss et al., 2E Murphy et al.
3A Harkness, 3B Willis, 3C Morales, 3D Boyle et al., 3E Doyle et al., 3F Murray et al.
4A Frankel, 4B Blumberg et al., 4C Baker et al., 4D Link et al., 4E Gallagher et al.
5A Burt, 5B Singer, 5C Campbell, 5D Dowd.

evaluation of those efforts are urgently needed.

Scientific integrity is central to successful health survey research. Many people believe that researchers typically lie with statistics, have political agendas, and breach confidentiality. While largely untrue, this indicates the need to engage in wider public education about how health surveys are developed, administered, and used, and evaluation of the effectiveness of various approaches.

Ethics and politics are at the heart of controversies about Homeland Security, FOIA and the Shelby Amendment (e.g., see www.upenn.edu/almanac/v45/n26/shelby.html), and classified research (e.g., Moreno, 2002). Objective study and analysis of these controversies, rather than advocacy, are needed to keep such controversies from resulting in suppression of important health survey research.

Risk, wrong, and harm are largely in the eye of the beholder. The complaint that IRBs act on hunch and anecdote in assessing risk should be replaced with empirical research that can inform IRBs.

Risk/benefit assessment is fundamental to decisions by researchers, IRBs, and funders about whether and how to conduct sensitive and risky research. In critical matters of public health, risk/benefit assessment cannot be left to conjecture. A better understanding of how to decide when research is acceptably safe and acceptable (e.g., Lowrance, 1976) is vital to decisions about public health. Solutions will depend in part on survey research assessing and analyzing perceptions of risk and benefit.

Epistemology. What knowledge really is and how it should be used are at the heart of debates as diverse as those about feminist interview methodology, grounded theory, and Bayesian sampling strategies. Such debates cannot be settled unless researchers undertake empirical study of alternative methods to evaluate the validity of claims for and against each method.

Development or use of *modeling* techniques will enable health researchers to better use data to test theories about complex phenomena, such as the cost of health care and the course of epidemics.

Each of these kinds of situations will become better understood or resolved if there is a concerted focus on them via a well-developed literature on ethical problem solving in research.

WAYS TO DO EMPIRCAL RESEARCH ON ISSUES OF REACTIVE METHODOLOGY

Since empirical research on research ethics is not a field in which scientists typically work, how does such research get done? Several ways have become apparent, and more are likely to emerge.

- Ethical problems may need to be solved in order to ensure participant cooperation.
- New research problems sometimes pose a complex of methodological and ethical puzzles and produce whole new outlooks on research (e.g., AIDS research).
- One may work in a substantive area of social science that is relevant to research ethics, such as socialization of scientists, privacy as a behavioral phenomenon, or organizational behavior.
- Methodologists may create methodological solutions to ethical problems.
- One may find oneself embroiled in ethical problems (e.g., accused of insensitivity to or betrayal of a research population) and examine issues causing this turmoil.
- Researchers puzzled by certain ethical problems over the years may find that research and scholarship about those problems become the next logical career step.

Given such motives, how does one then begin the empirical research? There are several ways:

- Pilot studies may be built into a larger study to answer an ethical question.
- Post-study evaluation, perhaps as part of debriefing or through follow-up contact with participants, can be used to answer questions about participants' perceptions, concerns, and/or reactions.

- Surrogate subject studies may be conducted, asking people how they would feel about being in a study in which such-and-such would occur.
- Experiments within studies (e.g., a survey may randomly assign respondents to different approaches).
- Stand-alone experiments, such as that reported by Eleanor Singer.

Each of these approaches may focus on a specific procedure or population to discover the best way to conduct a specific kind of research. Alternatively, one may employ such approaches in pursuit of answers to more general ethical questions. Or, one may review studies of an ethical issue over diverse topics, procedures, or populations to seek general principles of ethical problem solving.

Not all research on research ethics is empirical in the strictest sense. *Methodological* research can provide a basis for solving ethical dilemmas. *Modeling* can direct more efficient empirical answers to questions. *Theoretical* research typically involves methodological or empirical study as well but is primarily focused on conceptualizing and synthesizing what is already known about a given problem.

As Norman Bradburn has metaphorically expressed it, ethical dilemmas in research are not the showy annuals that thrill us but the hardy perennials that keep showing up—like weeds. We would rather they were not there. Not surprisingly, we have been slow to acknowledge the range of ethical problems out there and to tackle them systematically. However, social scientists have the tools and theories that can be used to identify, understand, and solve these problems. Perhaps we are ready to regard empirical research on research ethics as a vital part of doing our work and creating better science.

REFERENCES

Lowrance, W. W. (1976). *Of acceptable risk: Science and the determination of safety*. Los Alto, CA: Kaufmann.

Melton, G., Levine, R. Koocher, G., Rosenthal, R. & Thompson, W. (1988). Community consultation in socially sensitive research: Lessons from clinical trials on treatments for AIDS. *American Psychologist, 43*, 573-581.

Moreno, J. (2001). *Undue risk: Secret state experiments on humans*. New York: W. H. Freeman.

SESSION 5 SUMMARY

Larry Osborn, Abt Associates, Inc.
Marcie L. Cynamon, National Center for Health Statistics

Session 5's presentations and discussion centered on three interrelated themes: (1) incorporation of increasing confidentiality and data security regulations into the research process while maintaining study integrity; (2) the need for empirical research regarding best practices for provision of information about study confidentiality, risks, and benefits to respondents; (3) and a re-emphasis on the importance of the respondent-researcher contract that is at the heart of survey research.

INCORPORATION OF INCREASING CONFIDENTIALITY & DATA SECURITY REGULATIONS

The incorporation of confidentiality regulations, in particular the recent Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) and strictures applied by Institutional Review Boards (IRBs), into research protocols has created challenges. Such incorporation often requires creative thinking, a willingness to experiment with procedures to produce a mutually satisfactory protocol, forging positive collegial working relationships with IRB and other regulatory board members whenever possible, allowance of time for negotiations in a project schedule, a knowledge of the regulations that apply to any given project, and the implementation of methods to clearly explain the meaning of the regulations to respondents. One discussant coined the phrase "reactive methodology" to describe this interactive process. Researchers can play an important role in shaping future policy regarding the use of data. However, current standards (or the lack thereof) for release of public-use data for secondary analysis and the barriers to development of a common set of regulations for data release have hampered important and timely research.

IRBs typically provide guidance (and strictures) on a case-by-case basis, rather than basing such guidance on informed decisions backed by empirical evidence. Therefore, it is incumbent upon researchers to work with IRBs and regulatory bodies in ways similar to those described by the presenters. The presentations and discussion suggest that it is, or has so far been, possible to successfully incorporate mandated confidentiality and data security regulations into research protocols.

During the ensuing discussion, it was noted that there were a number of audience members who were members of IRBs and other regulatory boards. The need for these boards to allow experimentation in order to produce a mutually satisfactory protocol, rather than just imposing constraints, was stressed, and it was suggested that those who are members might try to promote that orientation within their own boards.

NEED FOR EMPIRICAL RESEARCH REGARDING PROVISION OF REQUIRED INFORMATION

Constraints imposed as protections to confidentiality often are imposed without empirical evidence to support them, highlighting the need for research to guide protocol and questionnaire design and the subsequent review of study protocols by regulatory boards. One presentation described the results of an experiment investigating the impact of varying degrees of confidentiality assurance provided during a survey introduction on respondent perceptions of risk and subsequent likelihood of participation. This study provides an excellent example of the sort of empirical research that is currently lacking. The presentation suggested that respondents are able to rationally consider the risk/benefit ratio of a study based on information provided to them in the study's introductory scripts and presumably make decisions regarding participation on the basis of that consideration. Therefore, the question was

asked whether it was ethical to attempt refusal conversion for respondents who have passed the point in a questionnaire where risk/benefit information is provided unless we are able to offer some alteration in the risk/benefit ratio. It was agreed that respondents appear to be able to assess risks and benefits based on the information that is provided, but what information is provided and how it is provided must be considered. Information regarding risks and benefits must be provided in a way that is clear yet not burdensome. Interviewers could be prepared to provide supplementary information regarding confidentiality assurances, if asked. In any case, the ability of respondents to assess risks and benefits should not preclude attempts at refusal conversion.

The process of determining how to provide confidentiality and data security without risking study integrity should involve further experimentation, and an archive of empirical research regarding procedures, both successful and unsuccessful, should be amassed in order to guide future study design and IRB review of those designs. To supplement efforts to work with regulatory boards to develop workable research protocols, and to build an archive of empirical research dedicated to ethical issues related to data collection and distribution, an online journal regarding such issues will be launched in the near future.

RE-EMPHASIS ON THE RESPONDENT-RESEARCH CONTRACT

This session, along with the others during the conference, highlighted the importance of the respondent-researcher contract and the need for researchers to consider their responsibilities in terms of fair treatment of respondents. While there appear to have been no major publicized breaches of the survey contract (such as those seen in medical trials, for example) occasional problematic cases do come to light. An example of an ethical abuse of data confidentiality was offered in which data gathered by a research firm were shared with a retail firm. Public awareness of such situations is likely to affect respondent trust in other studies. Regardless, it is essential to treat respondents well, recognizing their contributions in terms of time and information. Over time, a greater focus has been placed on improving interaction with respondents at each phase of a project, including questionnaire development, interviewer training, and data collection itself. It is becoming increasingly important to thoroughly examine how we engage with respondents, in order to make their participation as beneficial and positive as possible. The ultimate goals of both researchers and regulators should be to ensure that respondents are provided with the accurate information needed to make reasoned decisions regarding participation in research, to maximize the benefit of participation in the process, and to follow through with the implicit (or explicit) promise to respondents that their data will be used beneficially.

PARTICIPANT LIST

Lu Ann Aday
The University of Texas School of Public Health
P.O. Box 20186
Houston, TX 77225
Phone: (713) 500-9177
E-mail: laday@sph.uth.tmc.edu

Reg Baker
Market Strategies
20255 Victor Parkway, Suite 400
Livonia, MI 48152
Phone: (734) 542-7600
E-mail: Reg_Baker@marketstrategies.com

Timothy Beebe
State Health Access Data Assistance Center
University of Minnesota
2221 University Ave. SE, Suite 345
Minneapolis, MN 55414
Phone: (612) 624-1406
E-mail: beebe002@umn.edu

Stephen Blumberg
National Center for Health Statistics
3311 Toledo Rd., Room 2112
Hyattsville, MD 20782
Phone: (301) 458-4107
E-mail: sblumberg@cdc.gov

John Boyle
Schulman, Ronca & Bucuvalas, Inc.
8403 Colesville Rd., Suite 820
Silver Spring, MD 20910
Phone: (301) 608-3883
E-mail: j.boyle@srbi.com

Norman Bradburn
National Science Foundation
4201 Wilson Blvd., Room 905
Arlington, VA 22230
Phone: (703) 292-8700
Fax: (703) 292-9083
E-mail: nbradbur@nsf.gov

E. Richard Brown
UCLA Center for Health Policy Research
10911 Weyburn Ave., Suite 300
Los Angeles, CA 90024
Phone: (310) 794-0812
E-mail: erbrown@ucla.edu

Catharine Burt
National Center for Health Statistics
3311 Toledo Rd., Room 3407
Hyattsville, MD 20782
Phone: (301) 458-4126
E-mail: cburt@cdc.gov

Vicki Burt
National Center for Health Statistics
3311 Toledo Rd., Room 4211
Hyattsville, MD 20782
Phone: (301) 458-4127
E-mail: vburt@cdc.gov

Christine Bush
Health Resources and Services Administration
3830 S. Lowell Blvd.
Denver, CO 80236
Phone: (303) 781-3837
E-mail: cbush@hrsa.gov

Kathleen Cagney
Department of Health Studies
The University of Chicago
5841 S. Maryland Ave., MC 2007
Chicago, IL 60637
Phone: (773) 834-3924
E-mail: kcagney@health.bsd.uchicago.edu

Richard Campbell
Health Research and Policy Centers
University of Illinois at Chicago
850 W. Jackson Blvd.
Chicago, IL 60607
Phone: (312) 850-2257
E-mail: dcamp@uic.edu

Anne Ciemnecki
Mathematica Policy Research
600 Alexander Park
Princeton, NJ 08540
Phone: (609) 275-2323
E-mail: aciemnecki@mathematica-mpr.com

Steve Cohen
Agency for Healthcare Research and Quality
Center for Financing, Access and Cost Trends
540 Gaither Rd.
Rockville, MD 20850
Phone: (301) 427-1466
E-mail: scohen@ahrq.gov

Fred Conrad
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106
Phone: (734) 936-1019
E-mail: fconrad@isr.umich.edu

Llewellyn J. Cornelius
School of Social Work
University of Maryland
525 W. Redwood St.
Baltimore, MD 21201
Phone: (410) 706-7610
E-mail: lcorneli@ssw.umaryland.edu

Adolfo Correa
Centers for Disease Control and Prevention
National Center on Birth Defects and
Developmental Disabilities
1600 Clifton Road, MS E-86
Atlanta, GA 30333
Phone: (404) 498-3811
E-mail: acorrea@cdc.gov

Mick Couper
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106
Phone: (734) 647-3577
E-mail: mcouper@umich.edu

Marcie Cynamon
National Center for Health Statistics
3311 Toledo Rd., Room 2113
Hyattsville, MD 20782
Phone: (301) 458-4174
E-mail: mcynamon@cdc.gov

Michael Davern
State Health Access Data Assistance Center
University of Minnesota
2221 University Ave. SE, Suite 345
Minneapolis, MN 55414
Phone: (612) 625-4835
E-mail: daver004@umn.edu

William Davis
National Cancer Institute
6116 Executive Blvd., Suite 504, MSC8317
Bethesda, MD 20892-8317
Phone: (301) 594-3582
E-mail: davisbi@mail.nih.gov

Don Dillman
Social and Economic Sciences Research Center
Washington State University
Pullman, WA 99164-4014
Phone: (509) 335-1511
E-mail: dillman@wsu.edu

Kathryn Dowd
RTI International
Survey Research Division
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: (919) 541-6262
E-mail: kld@rti.org

Melanie Doyle
National Centre for Social Research
35 Northampton Square
London EC1V 0AX, England
Phone: 44-020-7459-9501
E-mail: m.doyle@natcen.ac.uk

Brad Edwards
Westat
1650 Research Blvd.
Rockville, MD 20850
Phone: (301) 294-2021
E-mail: bradedwards@westat.com

W. Sherman Edwards
Westat
1650 Research Blvd.
Rockville, MD 20850
Phone: (301) 294-3993
E-mail: shermedwards@westat.com

Jack Elinson
Mailman School of Public Health
Columbia University
722 W. 168th St.
New York, NY 10032
Phone: (212) 305-4027
E-mail: je7@columbia.edu

Trena Ezzati-Rice
Agency for Healthcare Research and Quality
Center for Financing, Access, and Cost Trends
540 Gaither Rd.
Rockville, MD 20850
Phone: (301) 427-1478
E-mail: tezzatir@ahrq.gov

Jacob Feldman
NORC
7500 Old Georgetown Rd., Suite 620
Bethesda, MD 20814
Phone: (301) 951-5071
E-mail: feldman-jack@norc.net

Floyd J. Fowler, Jr.
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
Phone: (617) 367-2000
E-mail: fjfowler@fimdm.org

Martin Frankel
Baruch College and Abt Associates, Inc.
14 Patricia Lane
Cos Cob, CT 06807
Phone: (203) 253-2541
E-mail: martin_frankel@abtassoc.com

Patricia Gallagher
Center for Survey Research
University of Massachusetts Boston
100 Morrissey Blvd.
Boston, MA 02125
Phone: (617) 287-7200
E-mail: patricia.gallagher@umb.edu

Joe Gfroerer
Substance Abuse and Mental Health Services Administration
5600 Fishers Lane, Room 16-105
Rockville, MD 20857
Phone: (301) 443-7977
E-mail: jgfroere@samhsa.gov

Pamela Giambo
Abt Associates
1110 Vermont Ave., Suite 610
Washington, DC 20005
Phone: (202) 263-1826
E-mail: pamela_giambo@abtassoc.com

Janet Harkness
ZUMA
Post Office Box 122155
D68072
Mannheim, Germany
Phone: 49-621-1246-284
E-mail: harkness@zuma-mannheim.de

Bradford Hesse
National Cancer Institute
Executive Plaza North, Room 4068
6130 Executive Blvd., MSC 7365
Bethesda, MD 20892-7365
Phone: (301) 594-9904
E-mail: hesseb@mail.nih.gov

Elizabeth Jacobs
Collaborative Research Unit
Cook County Hospital
1900 W. Polk St., 16th Floor
Chicago, IL 60612
Phone: (312) 864-7311
E-mail: ejacobs@rush.edu

Timothy Johnson
Survey Research Laboratory
University of Illinois at Chicago
412 S. Peoria St., 6th Floor
Chicago, IL 60607
Phone: (312) 996-5310
E-mail: timj@srl.uic.edu

Graham Kalton
Westat
1650 Research Blvd.
Rockville, MD 20850
Phone: (301) 251-8253
E-mail: grahamkalton@westat.com

Judith Kasper
Health Policy and Management
Bloomberg School of Public Health
Johns Hopkins University
624 N. Broadway, Room 641
Baltimore, MD 21205
Phone: (410) 614-4016
E-mail: jkasper@jhsph.edu

Dan Kasprzyk
Mathematica Policy Research
600 Maryland Ave., SW, Suite 550
Washington, DC 20024
Phone: (202) 264-3482
E-mail: dkasprzyk@mathematica-mpr.com

Alice Kroliczak
HIV/AIDS Bureau
Health Resources and Services Administration
5600 Fishers Lane, Room 7-90
Rockville, MD 20857
Phone: (301) 443-3592
E-mail: akroliczak@hrsa.gov

Richard Kulka
RTI International
P.O. Box 12194
Research Triangle Park, NC 27709-2194
Phone: (919) 541-7008
E-mail: rak@rti.org

James Lepkowski
Institute for Social Research
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106
Phone: (734) 936-0021
E-mail: jimlep@umich.edu

Marsha Lillie-Blanton
The Henry J. Kaiser Family Foundation
1330 G St., NW
Washington, DC 20005
Phone: (202) 347-5270
E-mail: mlillie-blanton@kff.org

Michael Link
RTI International
3040 Cornwallis Rd.
Research Triangle Park, NC 27709-2194
Phone: (919) 990-8462
E-mail: link@rti.org

Donna McAlpine
Division of Health Services Research and Policy
School of Public Health
University of Minnesota
420 Delaware Ave. SE, MMC 729
Minneapolis, MN 55455
Phone: (612) 625-9919
E-mail: mcalp004@umn.edu

Colleen McHorney
Regenstrief Institute for Health Care
1050 Wishard Blvd.
RG, Sixth Floor
Indianapolis, IN 46202
Phone: (317) 630-7664
E-mail: cmchorney@regenstrief.org

Peter Ph. Mohler
ZUMA
Post Office Box 122155
D68072
Mannheim, Germany
Phone: 49-621-1246-172
E-mail: director@zuma-mannheim.de

Ali Mokdad
Behavioral Surveillance Branch
Centers for Disease Control and Prevention
3005 Chamblee Tucker Rd., MS-K66
Atlanta, GA 30341
Phone: (770) 488-2524
E-mail: amokdad@cdc.gov

Joseph Murphy
RTI International
203 N. Wabash, 19th Floor
Chicago, IL 60601
Phone: (312) 456-5261
E-mail: jmurphy@rti.org

Mary Cay Murray
Abt Associates, Inc.
640 N. LaSalle, Suite 400
Chicago, IL 60610
Phone: (312) 867-4049
E-mail: mary_cay_murray@abtassoc.com

Kathleen O'Connor
National Center for Health Statistics
3311 Toledo Rd., Room 2114
Hyattsville, MD 20782-2003
Phone: (301) 458-4181
E-mail: kdo7@cdc.gov

Colm O'Muircheartaigh
NORC
University of Chicago
1155 E. 60th St.
Chicago, IL 60637
Phone: (312) 759-4017
E-mail: colm@norc.uchicago.edu

Diane O'Rourke
Survey Research Laboratory
University of Illinois at Chicago
505 E. Green St., Suite 3
Champaign, IL 61820
Phone: (217) 333-7170
E-mail: dianeo@srl.uic.edu

Larry Osborn
Abt Associates, Inc.
640 N. LaSalle, Suite 400
Chicago, IL 60610
Phone: (312) 867-4071
E-mail: larry_osborn@abtassoc.com

Joanne Pascale
U.S. Census Bureau
Statistical Research Division
Room 3134-4
Washington, DC 20233
Phone: (301) 763-4920
E-mail: joanne.pascale@census.gov

Colleen Porter
Department of Health Services Administration
University of Florida
P.O. Box 100195
Gainesville, FL 32610-0195
Phone: (352) 273-6068
E-mail: cporter@phhp.ufl.edu

D. E. B. Potter
Division of Statistical Research and Methods
Center for Financing, Access and Cost Trends
Agency for Healthcare Research and Quality
540 Gaither Road, Suite 500
Rockville, MD 20850
Phone: (301) 427-1564
E-mail: dpotter@ahrq.gov

Marsha Reichman
National Cancer Institute
6116 Executive Blvd., Suite 504
Bethesda, MD 20892
Phone: (301) 594-7032
E-mail: reichmam@mail.nih.gov

Robert Santos
NuStats
3006 Bee Cave Rd., Suite A300
Austin, TX 78746
Phone: (512) 306-9065 x2235
E-mail: rsantos@nustats.com

Joan Sieber
California State University, Hayward
Home address: 2060 Quail Canyon Court
Hayward, CA 94542
Phone: (510) 538-5424
E-mail: jsieber@bay.csuhayward.edu

Eleanor Singer
Institute for Social Research
Survey Research Center
University of Michigan
Box 1248
Ann Arbor, MI 48106
Phone: (734) 647-4599
E-mail: esinger@isr.umich.edu

Tenbroeck Smith
American Cancer Society, BRC
1599 Clifton Rd., NE
Atlanta, GA 30329
Phone: (404) 327-6442
E-mail: tenbroeck.smith@cancer.org

Tom Smith
NORC
1155 E. 60th St.
Chicago, IL 60637
Phone: (773) 256-6288
E-mail: smitht@norc.uchicago.edu

Yonette Thomas
Division of Epdemiology, Prevention, and Services Research
National Institute on Drug Abuse
6001 Executive Blvd., Room 5166
Bethesda, MD 20892
Phone: (301) 402-1910
E-mail: yt38e@nih.gov

Michael Wadsworth
Medical Research Council
National Survey of Health and Development
Department of Epidemiology
University College London Medical School
1-19 Torrington Pl.
London WC1E 6BT, England
Phone: 44-0-20-7679-1734
E-mail: m.wadsworth@ucl.ac.uk

Joseph Waksberg
Westat
1650 Research Blvd.
Rockville, MD 20850
Phone: (301) 251-1500
E-mail: josephwaksberg@westat.com

Dan Waldo
Center for Medicare and Medicaid Services
U.S. Department of Health and Human Services
7500 Security Blvd., C3-16-27
Baltimore, MD 21244-1850
Phone: (410) 786-7932
E-mail: dwaldo@cms.hhs.gov

Elizabeth Ward
American Cancer Society
1599 Clifton Rd. NE
Atlanta, GA 30329
Phone: (404) 327-6552
E-mail: elizabeth.ward@cancer.org

Richard Warnecke
Health Research and Policy Centers
University of Illinois at Chicago
850 W. Jackson Blvd.
Chicago, IL 60607
Phone: (312) 355-1167
E-mail: warnecke@uic.edu

Gordon Willis
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
6130 Executive Blvd., MSC 7344, EPN 4005
Bethesda, MD 20892-7344
Phone: (301) 594-6652
E-mail: willisg@mail.nih.gov

Karen Wilson
523 Garnsey Rd.
Fairport, NY 14450
Phone: 585-223-8267
Fax: (585) 242-9733
E-mail: karen_wilson@urmc.rochester.edu

Michael Wolfson
Statistics Canada
RH Coats Bldg., 26K
Ottawa, Ontario, Canada K1A 0T6
Phone: (613) 951-8216
E-mail: wolfson@statcan.ca