

Survey Methodology Research FY 1999 - FY 2013 Awards List

Awards from this competition were jointly reviewed and supported by NSF's Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies represented by the Federal Committee on Statistical Methodology ([FCSM](#)). The following agencies provided direct financial support for these awards:

[National Agriculture Statistics Service](#), USDA

[Bureau of Justice Statistics](#), DOJ

[National Center for Education Statistics](#), DOE

[Bureau of Labor Statistics](#), DOL

[National Center for Health Statistics](#), DHHS

[Department of Transportation](#)

[National Center for Science and Engineering Statistics](#), NSF

[Economic Research Service](#), USDA

[Social Security Administration](#)

[Energy Information Administration](#), DOE

[U.S. Census Bureau](#), DOC

[Internal Revenue Service](#)

[U.S. Department of the Treasury](#)

FY2013

Fractional Imputation for Incomplete Data Analysis

1324922

Jae-Kwang Kim

Iowa State University

Amount: \$249,997

Total Award Duration: 36 Months

Incomplete data frequently are encountered in survey data due to nonresponse, inaccurate measurement, and two-phase sampling, among other things. Any form of incomplete data can damage the representativeness of a sample, and a naive analysis with incomplete data can lead to biased estimation. Imputation is a process of assigning values to the missing items with the objective of reducing bias and improving the efficiency of the resulting estimators. This project will develop fractional imputation methods as a tool for handling incomplete data for general-purpose estimation. These methods will serve as important building blocks for the establishment of a complete statistical package for analysis of incomplete data that ultimately can be applied to problems in a variety of disciplines, including the social, behavioral, and economic sciences. In particular, the project will develop fractional imputation to address several important problems with incomplete data, including (1) likelihood-based inference, (2) robust estimation using fractional hot deck imputation, (3) synthetic imputation for survey integration, and (4) statistical matching technique.

Because fractional imputation is a relatively new approach for handling incomplete data, there is a critical need for theoretical and methodological development. The advantages of the fractional imputation approach lie in its computational simplicity, wide applicability, and its statistical validity. By using fractional weights, fractional imputation avoids the burden of iterative computation, such as Markov Chain Monte Carlo, for the evaluation of conditional expectation associated with missing data. The proposed approach can be used to estimate parameters consistently and efficiently. The fractional imputation approach can be applied to nonstandard situations such as measurement error models, regression analysis combining two different surveys, and causal inference from observational studies. The impact of the proposed research is therefore substantial because the proposed approach can be used as a general methodology for incomplete data. Because of the computational simplicity and statistical validity of the fractional imputation approach, the results of the proposed research should have wide applicability. It also should have a major impact in providing complete data sets for analysis and new data products combining information from different surveys.

[Additional Information](#)

National Historical Geographic Information System

1324875

Steven Manson

University of Minnesota-Twin Cities

Amount: \$869,999

Total Award Duration: 60 Months

The National Historical Geographic Information System (NHGIS) is the nation's most comprehensive source for statistical data, geographic data, and metadata describing spatial characteristics of the American population from 1790 to the present. With 265 billion data points, NHGIS is the largest publicly accessible social science database in the world and is used by thousands of researchers. This project will implement four major improvements to the infrastructure of NHGIS. NHGIS will create data tables for the 1790-1940 decennial censuses from a massive new individual-level database. The project will develop new GIS data identifying locations of incorporated places and MCDs for the 1790-1970 censuses. Census unit boundaries frequently change from one census to the next, making it difficult to analyze population changes over time and space. NHGIS will interpolate census statistics from 1980, 1990, and 2000 and the American Community Survey to produce estimates for 2010 census areas at several geographic levels, including census tracts, places, county subdivisions, and metropolitan areas. Finally, the project will sustain and expand NHGIS education and outreach efforts, offering individual user support and in-person workshops, online tutorials, and web-based community tools.

NHGIS democratizes access to the census, the fundamental source of data about the U.S. population. The proposed improvements will be used for academic research and also for social science training, journalism, policy research at the state and local levels, and private sector research. This infrastructure will allow thousands of investigators in disciplines from economics and ecology to environmental and health policy to address changes across the broad sweep of time at fine levels of spatial organization. NHGIS provides a unique laboratory for the spatial analysis of economic and social processes and offers the empirical foundation needed for developing and testing models of society. Creating new spatiotemporal data for the period 1790 through 1940 will open new opportunities to investigate profound social and economic transformations of American society, including industrialization, immigration, westward migration, and urbanization. Small-area census data are the primary source for investigating such critical issues as suburbanization, the decline and rebirth of central cities, residential segregation, immigrant settlement patterns, rural depopulation, agricultural consolidation, and population shifts from the rust belt to the Sunbelt. The improvements to NHGIS will allow researchers to apply new and powerful approaches to familiar problems by broadening the scope of local and regional analyses to explore variations across time and space simultaneously.

[Additional Information](#)

Reduction of Survey Length through Split Questionnaire Design: Consequences for Nonresponse and Measurement Error

1259985

Andrey Peytchev

Research Triangle Institute

Amount: \$580,000

Total Award Duration: 36 Months

Much research in the social sciences and development of government policy relies on survey data, and the demand for survey data continues to grow. The need for more data has led to longer surveys, increasing the burden for survey respondents in terms of time and effort. Empirical evidence shows a positive correlation between survey length and survey nonresponse, which threatens the representativeness of the survey estimates. There also is evidence that measurement (reporting) error increases as respondents are asked to answer more questions in the survey. Collecting fewer variables may not satisfy a given study's objectives, however. This research project experimentally evaluates the ability to collect all desired data through a split questionnaire design in which respondents are asked only a subset of the questions. The project will use a multiple imputation method to complete the data in the sections that are not asked of particular respondents. The investigators' will extend current imputation methods to include semi-parametric and parametric models. The main hypothesis is that the split questionnaire design approach will yield estimates with less bias and even less total error compared to deploying the full questionnaire.

This project evaluates a method that essentially transfers part of the time and effort to complete the survey from the individual to the researcher. It also evaluates the ability to collect higher quality data as a result of this reduction in respondent burden. Finally, the study aims to extend the employed statistical methods to better preserve the properties of the data. The results will help to provide an alternative methodology for a wide array of surveys, improve split questionnaire design methodology itself, and provide information regarding the circumstances under which implementing such designs can be beneficial.

[Additional Information](#)

Mobile Devices for Survey Data Collection

1261340

Roger Tourangeau

Westat Inc

Amount: \$519,492

Total Award Duration: 36 Months

This project will examine mobile devices, specifically smartphones and tablet computers, as vehicles for survey data collection. The appeal of these devices for survey researchers is clear. Because they are lightweight and relatively inexpensive, they make it easier to collect data using such existing survey modes as computer-assisted personal interviewing. The research will examine three issues raised by use of such devices. First, the input methods that these devices permit (such as touchscreen interfaces) are relatively unfamiliar to many users and may create response problems. Although these interfaces are sometimes used on laptops, tablets and smartphones require them, making usability concerns more central. Second, the screens on tablets and smartphones are considerably smaller than those on laptop or desktop computers. Experiments on web surveys demonstrate the importance of "visual prominence." Any information that respondents need to use should be immediately visible to them without their having to perform any action (such as a mouse click) to make the information visible. Even the need for an eye movement may effectively render information invisible. Because of the small screens on mobile devices, it may be much harder to make all of the potentially useful information visible to respondents than it is with a laptop or desktop computer. The final issue is the perceived privacy of data collected on these devices. Respondents are willing to reveal sensitive information about themselves when a computer administers the questions, and web surveys seem to retain the advantages of earlier methods of computerized self-administration. But it is unclear whether respondents will display the same level of candor when the survey is administered over the Internet on a tablet computer or a smartphone. Two realistic field experiments and a usability study will examine these issues. Both experiments will be conducted in a single, face-to-face survey. The first experiment will compare laptop computers with tablets and smartphones and will examine the effects of both screen size and input method on breakoffs, missing data, completion times, and indicators of the quality of the responses. The second experiment will compare the same three data collection platforms as vehicles for collecting sensitive information. The experiment will ask respondents to assess the sensitivity of the questions, because item sensitivity may vary as a function of the device used to collect the data.

Surveys are a central tool for social scientists and policymakers in the United States, and survey research is a multi-billion dollar industry in the United States alone. Any set of technological advances, such as the widespread adoption of smartphones and tablet computers, is likely to have a major impact on how surveys are done. Although mobile devices will be widely used for surveys regardless of whether this research is done, the work will produce practical guidelines

for using such devices to collect survey data and will alert survey researchers to some of the potential pitfalls of these devices.

[Additional Information](#)

FY 2012

Interactional Influences on Survey Participation

1230069

Nora Schaeffer

University of Wisconsin-Madison

Amount: \$260,596

Total Award Duration: 24 Months

This research focuses on the critical challenge of recruiting participants for social research surveys, a problem highlighted by the recent National Research Council's Panel on the Future of Social Science Data Collection. The refusal component of nonresponse has grown steadily in recent years, threatening the ability of surveys to represent the populations from which the samples are drawn. The proposed research examines the moment-by-moment unfolding reciprocal influence between interviewers and sample members. For example, when an interviewer engages in persuasive actions, is it because she has somehow perceived cues that this sample member is persuadable? When the interviewer engages in actions that appear to be successful in persuading the sample member, is it simply that a persuadable sample member has provided her the opportunity to do so? The research uses new substantive and methodological approaches, drawing on theories of social exchange and reciprocity; conversation analysis and the interaction order; and content analysis. The empirical investigations exploit a unique collection of pairs of acceptances and declinations in a case-control design that uses observations from the Wisconsin Longitudinal Study that are matched on the sample members' propensity to participate. The matching of observations provides for comparison between calls that end in acceptances and declinations of the request for the interview. The project fills a gap in knowledge about whether sample members' voices provide interviewers with information that they can use to rapidly form accurate assessments of the likelihood that a call will lead to an interview. An acoustical analysis of the sample member's initial speech provides variables such as pitch that can be used with the case-control design to estimate whether these properties of the sample member's voice predict acceptance of the request for participation. By selectively and strategically adding to the existing data transcripts of subsequent contacts and attempts to obtain cooperation, the research also expands previous descriptions of ways interviewers tailor and maintain interaction to describe whether empathic responsiveness and interactional responsiveness (as specific forms of tailoring) make a difference for acceptances of the request to participate.

The intellectual merit and importance of the project are rooted in its contributions to improving the accuracy of social scientific research, which is key to the nation's data infrastructure. The quality of information that describes an entire population can be compromised if participation is restricted to highly selected segments within a sample. This research expands theories of survey participation and their application to the development of practical methods for increasing participation in survey interviews for social research. The broader impacts of the research include identifying specific techniques for improving recruitment to participation in social research, disseminating information about these techniques in refereed journals and presentations at conferences, and training of both graduate and undergraduate students as they participate in

the research effort. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Diagnosis and Reduction of Bias in Respondent-driven Sampling

1230081

Krista Gile

University of Massachusetts Amherst

Amount: \$199,938

Total Award Duration: 36 Months

Respondent-Driven Sampling (RDS) is a type of link-tracing network sampling used to study hard-to-reach human populations. Beginning with a convenience sample, respondents are given uniquely identified coupons to distribute to other population members, making them eligible for enrollment. This is effective at collecting large diverse samples from many hard-to-reach populations. Despite often highly-effective sampling, statistical inference from RDS data is under-developed. Estimates are based on strong assumptions allowing the data to be treated as a probability sample. This project develops methodology for RDS in three ways: (1) Identifying sources of bias and elevated variance in real populations; (2) Extending the network model-assisted inferential framework to address additional sources of bias; and (3) Expanding the network model-assisted framework to arbitrarily large populations. In (1), the research will use simulated RDS samples from fully-observed network data to both explore features of real-world network structures that induce bias or elevated variance in estimates, and also to calibrate diagnostics of such features based on sampled data. Parts (2) and (3) involve advances of the network model-assisted estimator for RDS data. This is the most flexible estimator currently available for RDS data, but is quite computationally burdensome. In (2), the project will extend the existing network model-assisted estimator to adjust for additional features of the network and sampling process. In (3), the project will develop a new variant of this estimator that is more computationally practical in large populations.

Hard-to-reach populations are often at the low end of disparities in wealth, opportunity, and acceptance and at higher risk for negative social and health outcomes. These populations couple reduced social and health opportunities with a difficulty in collecting traditional probability samples. Thus, RDS is widely used and is of special interest to social and behavioral scientists as well as public health officials. This project will allow practitioners to confidently use RDS data in a wider variety of contexts. Specifically, researchers will (a) better understand which network and sampling conditions might induce bias into their estimators and have better tools to find evidence of these features in their samples, (b) have more options for adjusting their estimators to account for non-ideal network and sampling conditions, and (c) be able to compute the resulting estimators with greater efficiency, making the methods practical even in cases of very large target populations. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

The Validity of Markov Latent Class Analysis for Evaluating Measurement Errors in Complex Panel Surveys

1229222

Paul Biemer

Marcus Berzofsky

Amount: \$175,000

Total Award Duration: 24 Months

Markov latent class analysis (MLCA) comprises a broad class of models and techniques for analyzing categorical longitudinal data subject to misclassification. An important application area is exploring data quality issues in panel surveys. Because MLCA does not rely on gold standard or replicate measurements, it can be applied to virtually any panel survey. For data quality evaluations, MLCA has been used to compare interview modes and alternative questionnaire designs, estimate measurement bias, investigate the causes of misclassification, and investigate many other measurement error issues. Despite its many potential applications in survey work, MLCA has not enjoyed widespread use among survey methodologists because practical guidance on fitting MLC models to complex survey data is lacking. This project will: (1) evaluate the magnitude of the model bias when one or more MLCA assumptions fail when analyzing complex survey data under a wide range of conditions; (2) identify and evaluate the current strategies for diagnosing and repairing MLC model failure and misspecification; (3) address the limitations of current methods by developing improved strategies for diagnosing and repairing MLC model failure and misspecification, particularly in applications to complex surveys; and (4) apply the most effective diagnostic and remedial approaches to real panel survey data to demonstrate the range of modeling issues that can arise in practical applications as well as how to deal with them effectively. As part of the application of these approaches, at least 10 years of data from several national panel surveys will be analyzed to identify temporal trends in measurement error for key national statistics.

This research has important implications for MLCA in all branches of science where classification error is an issue, including social science, epidemiology, clinical research, educational testing, and psychology. The project's impact will be felt in at least four ways. The research has particular relevance for complex survey applications because of the emphasis in this research on modeling cluster-correlated data selected with unequal probabilities and subject to nonresponse and measurement error. It also has important implications for disadvantaged and minority populations whose data may be differentially affected by measurement error. In addition, the evaluation of error trends will provide important information on current and historical levels of measurement error for three important federal statistical programs. Finally, theories regarding the relationship between measurement error and survey participation will be formulated and tested. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Doctoral Dissertation Research: Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys

1238612

Richard Valliant

University of Michigan Ann Arbor

Amount: \$15,153

Total Award Duration: 12 Months

Sequential mixed-mode surveys use a mix of modes or data collection methods such as mail, telephone, in-person, and web to increase the number of people who respond to a survey. In sequential designs, there is usually no control in assigning subgroups of respondents to modes. As a result, nonrandom assignment of modes is an inherent characteristic of sequential mixed-mode surveys. This design is important since there are usually limited funds to probe people to respond. While the goal of using mixed modes is clear, one compelling research question is how the nonrandom mix of mode impacts survey data and how these effects should be handled in estimating survey population characteristics such as mean income, and health insurance coverage. To date, since the nonrandom mix of modes poses a challenge in evaluating the mode effects, the existing inference methods assume that mode effects can be ignored in sequential mixed-mode surveys despite their unknown impact on the quality of the survey estimates. This research develops and evaluates the statistical inference methods accounting for nonrandom mode effects to test the comparability of the survey estimates from the different modes. In parallel, this project also develops statistical inference methods accounting for both nonresponse and nonrandom mode effects in the presence of nonignorable mode effects. The public-use Current Population Survey (CPS), 1973, and Social Security Records Exact Match, and the nonpublic-use American Community Survey (ACS) data will be used to conduct empirical and simulation evaluations.

This research provides federal agencies, survey organizations, research centers, and other data producers assessment and inferential methods that adjust for both nonresponse and nonrandom mode effects in the context of sequential mixed-mode surveys. Some large surveys have employed some variation of mixed-mode surveys in order to meet budget constraints. On the other hand, in the presence of nonignorable mode effects, the bias properties for the survey population characteristics are not known and the existing assessment and inferential methods do not control for the nonrandom mode effects. This research produces sequential mixed-mode assessment methods which will test the ignorability of the mode effects which can be a threat for the quality of survey data. In parallel, this research also produces methods of inference which will yield higher quality survey estimates in the presence of nonignorable mode effects.

[Additional Information](#)

FY 2011

Modernizing Occupational Coding In The General Social Survey

1123510

Michael Hout

Peter V Marsden

National Opinion Research Center

Amount: \$500,000

Total Award Duration: 24 Months

In the past thirty years, changes in technology, business, and government practice have substantially altered the American occupational structure. Our project provides a foundation to understand the consequences of new occupations on the current economy and contemporary society, and to preserve unique data key to documenting these fundamental historical changes. Specifically, this project modernizes the occupational and industry data in the General Social Survey (GSS) from the 1970s to the present time.

The project has several goals. They dovetail recent key NSF recommendations that encourage large infrastructure data sources such as the GSS to facilitate increased data access and dissemination. This can be done by presenting data and metadata according to a well-defined protocol, which will allow desirable modes of data access, search, downloads, and documentation. The project also meets the NSF challenge to retrofit historical or legacy data and metadata to become machine readable. This will possibly open up vast amount of data for dissemination and analysis once issues of confidentiality and disclosure are resolved.

To accomplish this goal, this project will (1) retrieve GSS respondents' detailed verbatim descriptions of their work activities, occupations, and industries from the physical questionnaire manuscripts from early GSS waves, (2) convert them into machine-readable form, (3) recode them to reflect 2010 occupation and 2007 industry categories developed by the U.S. Census, and (4) attach external data such as socioeconomic scores and prestige assessments to the recoded categories.

The intellectual merit of digitizing occupational information and recoding occupational and industry categories in the process is that it enables researchers to use the full potential of the occupation and industry information recorded in the GSS over time. Doing so will enhance the value of the GSS as a resource for comparative and contemporary research on social inequality, mobility, and other fields and preserve its growing value as a historical database describing trends in U.S. society over two generations. Ensuring the longevity of such legacy data by converting hand-written text into machine-readable text, the project also develops an archive of verbatim descriptions that will allow future researchers to code them using other standards, including U.S. Census standards that may become available in upcoming decades.

Broader Impacts

The GSS is a public resource as well as a scientific one. Public media, especially newspapers,

make extensive use of the GSS. By improving the quality of occupational and industry information in the GSS and ensuring that it is coded in a consistent way over time, this project will help journalists and citizens make sense of social trends and patterns. Also, high schools and colleges make extensive use of the GSS as a teaching tool. Teachers and students will get more out of these exercises from the new data products this project will produce when data reflect contemporary distinctions among occupations and industries as accurately and precisely as possible.

Collaborative Research: Best Predictive Small Area Estimation

1121794

Ahmad Hakim-Elahi

Amount: \$69,462

Total Award Duration: 36

1122399

Jonnagadda S. Rao

University of Miami School of Medicine

Amount: \$78,632

Total Award Duration: 36

1118469

Thuan Nguyen

Oregon Health and Science University

Amount: \$85,800

Total Award Duration: 36

Surveys usually are designed to produce reliable estimates of various characteristics of interest for large geographic areas or socio-economic domains. However, for effective planning of health, social, and other services and for apportioning government funds, there has been a growing demand to produce similar estimates for small geographic areas and subpopulations, commonly referred to as small areas. This research project aims at developing a new method of small area estimation that potentially will lead to a dramatic improvement in accuracy over the traditional methods in practical situations. Model-based small area estimation utilizes statistical models, such as mixed effects models, to "borrow strength." In particular, the empirical best linear unbiased prediction (EBLUP) is a well-known model-based method that has had dominant influence in small area estimation. From a practical point of view, however, any proposed model is subject to model misspecification. When the proposed statistical model is incorrect, EBLUP is no longer efficient or even effective. In such cases, a new method, known as observed best prediction (OBP), may be superior. This project involves several important research topics on OBP, including theoretical developments, assessment of uncertainties under weak model assumptions, and implementation of the OBP via user-friendly software. The research largely will expand the results of our earlier studies, and contribute to making the OBP method more effective, practical, and easy to use.

The research introduces a completely new idea and method to model-based statistical methods in survey sampling. It is expected to impact other scientific areas where statistical methods have been used for prediction problems. The project will develop and freely disseminate R code to implement the OBP method. The education component of the project will introduce the OBP method into courses at the investigators' institutes. These courses are expected to draw students and researchers from statistics, biostatistics, genetic epidemiology, animal and plant sciences, educational research, social sciences, and government agencies. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Adjusting for Unmeasured Confounding Due to Cluster with Complex Sampling Designs

1115618

Babette Brumback

Amount: \$ 310,000

Total Award Duration: 36

In social epidemiology, a geographic neighborhood or cluster is viewed as an important determinant of health behaviors, mediators, and outcomes. One may be interested in the effects of measured or unmeasured neighborhood characteristics or in turn on individual effects that have been disentangled from neighborhood effects. Analyses of nationally representative surveys, such as the National Health Interview Survey, provide a means of estimating these effects. This project will develop statistical methods that can account for the complex sampling design of such surveys at the same time as disentangling individual effects from neighborhood effects. These methods will be applied to analyze data from the National Health Interview Survey. Furthermore, in global health and several other fields, community randomized trials with complex sampling designs are used to estimate the effect of one or more interventions versus a standard or control condition. Scientific interest often focuses on the relative effects of community-level adherence to the intervention on individual-level outcomes. The project also will develop statistical methods that can account for the complex sampling design of such trials while simultaneously extracting the effect of intervention adherence, as opposed to that of intervention intention on individual-level outcomes. These methods will be applied to analyze data from randomized trials designed to study effects of school-level sanitation, water safety, and hygiene on individual education outcomes.

This project involves collaborative research across the fields of biostatistics, social epidemiology, and global health. The research will advance statistical methodology as well as improve the capability of researchers in social epidemiology, global health, and other fields to address important scientific questions. Disentangling individual-level effects from neighborhood-level effects will be useful in understanding the relative roles of the individual versus society and environment in health behaviors and outcomes, which will be useful in designing interventions. Going beyond simple comparisons of treated and untreated individuals in randomized clinical trials and estimating the effects of community-level intervention adherence on individual-level outcomes will further understanding of the effects of interventions in global health and other fields. Illustrating and communicating the new statistical methods for joint estimation of individual-level and neighborhood-level effects on outcomes using nationally representative surveys will provide the federal agencies that sponsor these surveys with enhanced options for creating public-use datasets to facilitate these analyses. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

Collaborative Research: Responding to Surveys on Mobile Multimodal Devices

1026225

Frederick Conrad

Total Award Duration: 36 months

Amount: \$705,410

1025645

Michael Schober

Amount: 36

Total Award Duration: \$258,335

Collecting survey data of national importance (for example, on employment, health, and public opinion trends) is becoming more difficult as communication technologies undergo rapid and radical change. Important basic questions about whether and how to adapt data collection methods urgently need to be addressed. This project investigates how survey participation, completion, data quality, and respondent satisfaction are affected when respondents answer survey questions via mobile phones with multimedia capabilities (e.g., iPhones and other "app phones"), which allow alternative modes for answering (voice, text) and can allow respondents to answer questions in a different mode than the one in which they were invited. Two experiments will compare participation, completion, data quality, and satisfaction when the interviewing agent is a live human or a computer and when the medium of communication is voice or text, resulting in four modes: human-voice interviews, human-text interviews, automated-voice interviews, and automated-text interviews. The first experiment randomly assigns respondents to one of these modes; the second experiment allows respondents to choose the mode in which they answer. Results will shed light on whether respondents using these devices agree to participate and answer differently to human and computer-based interviewing agents, and whether this differs for more and less sensitive questions. Results also will shed light on how the effort required to interact with a particular medium (e.g., more effort to enter text than to speak) affects respondents' behavior and experience, and whether the physical environment that respondents are in (a noisy environment, a non-private environment, a brightly lit environment with glare that makes reading a screen difficult) affects their mode choice and the quality of their data. Finally, the results will clarify how allowing respondents to choose their mode of response affects response rates and data quality.

These studies are designed to benefit researchers, survey respondents, and society more broadly. For researchers, the benefit is to allow them to adapt to the mobile revolution as they collect data that are essential for the functioning of modern societies, maintaining high levels of contact and participation while gathering reliable and useful data. For survey respondents, the potential benefit is the design of systems that make it more convenient and pleasant to respond and that enable them to choose ways of responding appropriate to their interactive style, the subject matter, and their physical environment. For society more broadly, it is essential that the survey enterprise is able to continue to gather crucial information that is reliable and does not place undue burden on citizens as their use of communication technology changes and as alternate sources of digital data about people proliferate. More fundamentally, the results will add to basic understanding of how human communication is evolving as people have expanded ability to communicate anytime, anywhere, and in a variety of ways. The project is supported by the

Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

FY 2010

Collaborative Research: Responding to Surveys on Mobile Multimodal Devices

1026225

Frederick Conrad

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$525,873

1025645

Michael Schober

New School University

Total Award Duration: 36 months

Amount: \$156,258

Collecting survey data of national importance (for example, on employment, health, and public opinion trends) is becoming more difficult as communication technologies undergo rapid and radical change. Important basic questions about whether and how to adapt data collection methods urgently need to be addressed. This project investigates how survey participation, completion, data quality, and respondent satisfaction are affected when respondents answer survey questions via mobile phones with multimedia capabilities (e.g., iPhones and other "app phones"), which allow alternative modes for answering (voice, text) and can allow respondents to answer questions in a different mode than the one in which they were invited. Two experiments will compare participation, completion, data quality, and satisfaction when the interviewing agent is a live human or a computer and when the medium of communication is voice or text, resulting in four modes: human-voice interviews, human-text interviews, automated-voice interviews, and automated-text interviews. The first experiment randomly assigns respondents to one of these modes; the second experiment allows respondents to choose the mode in which they answer. Results will shed light on whether respondents using these devices agree to participate and answer differently to human and computer-based interviewing agents, and whether this differs for more and less sensitive questions. Results also will shed light on how the effort required to interact with a particular medium (e.g., more effort to enter text than to speak) affects respondents' behavior and experience, and whether the physical environment that respondents are in (a noisy environment, a non-private environment, a brightly lit environment with glare that makes reading a screen difficult) affects their mode choice and the quality of their data. Finally, the results will clarify how allowing respondents to choose their mode of response affects response rates and data quality.

These studies are designed to benefit researchers, survey respondents, and society more broadly. For researchers, the benefit is to allow them to adapt to the mobile revolution as they collect data that are essential for the functioning of modern societies, maintaining high levels of contact and participation while gathering reliable and useful data. For survey respondents, the potential benefit is the design of systems that make it more convenient and pleasant to respond and that enable them to choose ways of responding appropriate to their interactive style, the subject matter, and their physical environment. For society more broadly, it is essential that the survey enterprise is able to continue to gather crucial information that is reliable and does not place

undue burden on citizens as their use of communication technology changes and as alternate sources of digital data about people proliferate. More fundamentally, the results will add to basic understanding of how human communication is evolving as people have expanded ability to communicate anytime, anywhere, and in a variety of ways. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)
[Additional Information](#)

Bayesian Empirical Likelihood and Penalized Splines for Small Area Estimation

1026165

Malay Ghosh

University of Florida

Total Award Duration: 36 months

Amount: \$169,991

This research will develop new semiparametric Bayesian methods for small area estimation based on empirical likelihood and penalized splines. The approach also can be adapted for certain random and fixed effects models. These methods will allow for greater flexibility in handling problems where assumptions of normality of the likelihood, the linearity, or both can be subject to question. Empirical likelihood dispenses with any parametric structure of the likelihood, while penalized splines can avoid the assumption of a specific functional relationship between the response and the covariates. The introduction of Dirichlet process mixture priors for random effects overcomes the unverifiable and sometimes questionable assumption of Gaussianity of random effects. Further, these priors are particularly helpful when one requires clustering of small areas for administrative purposes. The research will examine the robustness of the new methods through simulation studies. The integration of penalized splines with empirical likelihood also will be investigated.

Small area estimation has become vital for every Federal agency in the United States. Examples include the Small Area Income and Poverty Estimation (SAIPE) project of the United States Bureau of the Census, local area unemployment rates as needed by the Bureau of Labor Statistics, small area agricultural cash rent program of the United States Department of Agriculture, and the estimation of children under poverty in K-12 grades at the school district level, which are useful for the Department of Education and many other projects. Small area estimation also is important for the private sector; for example, in aiding the decision making of local businesses. The unifying theme in all these is that one needs reliable estimates at lower levels of geography, such as counties, subcounties, and census tracts. The varied nature of problems and the associated complexity in this regard demands a continuous enhancement of existing methods and the development of new techniques. The development of new small area estimation methodology therefore holds great promise for real life applications. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Doctoral Dissertation Research: Clarifying Survey Questions

1024244

Stanley Presser

Cleo Redline

University of Maryland College Park

Total Award Duration: 12 months

Amount: \$12,000

Although comprehension is well recognized as a critical component of the survey question response process, much about it remains unknown. Past research has shown that ambiguous or vague concepts can be clarified through the use of definitions, instructions, or examples, but respondents do not necessarily attend to these clarifications. The aim of this doctoral dissertation research project is to investigate where and how to present clarifying information so that respondents will recognize it as essential to their answering survey questions correctly. A key issue is whether sensory channels (aural versus visual) make different demands on comprehension. The answer to this question is largely unknown, because sensory channel is often confounded with the presence of an interviewer in so many of the relevant studies. This project will help to establish whether respondents anticipate the end of a question and are more likely to interrupt clarifying information that is placed after a question than before and whether this harms survey estimates. It will help to confirm whether incorporating the clarifications into the questions and asking a series of simpler questions, as suggested by many researchers, is even more effective. Finally, understanding will be gained regarding whether respondents are better at comprehending complex clarifications in the visual channel than the aural, and whether channel interacts with the method of clarification.

The goal of this project, to gain a better understanding of how to reduce ambiguity and vagueness in survey questions across survey modes, is especially relevant given the current debate over how to design questions for mixed-mode surveys. This project has the potential to lead to more accurate survey estimates, to better descriptions of the nation based on survey data, and to better decision-making policies. As a Doctoral Dissertation Research Improvement award, support is provided to enable a promising student to establish a strong independent research career. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Latent Space Models for Aggregated Relational Data in the Social Sciences

1023176

Tian Zheng

Andrew Gelman

Columbia University

Total Award Duration: 36 months

Amount: \$350,000

How does the understanding of social networks contribute to social science? In particular, (1) which network features or observable characteristics encode social structure; (2) how do these features contribute to the formation of connections or social ties; and (3) how does network structure impacts diffusion, specifically the spread of influences, opinions, and diseases? A key difficulty in studying these questions is that most contributions to current understanding in this area come from a small number of applications where full network data are readily available. Collecting complete network data is typically financially and practically impossible, while sampled network data are hard to collect and usually require special statistical modeling considerations. In this project, latent space approaches will be applied to aggregated relational data (questions of the form, "How many X?s do you know?") collected using surveys on non-network samples.

The project's methods development and data analysis should make information about more complicated network structure available to the multitude of researchers who cannot in practice collect data from complete networks. By analyzing existing data using this framework, the project will estimate aspects of social structure in personal acquaintances networks, variation of social structure across different sub-networks (family, friends, coworkers, etc), estimate homogeneity of groups, and estimate individual and population gregariousness, all of which could benefit social science researchers interested in understanding relationships between groups of individuals in the population. The methods also are potentially useful for research on groups that are difficult to reach with traditional surveys, such as those with HIV/AIDS, the homeless, or injection drug users. The project is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

FY 2009

Some Issues in Analytic Inference from Complex Surveys

0922142

F. Jay Breidt

Jean Opsomer

Colorado State University

Total Award Duration: 36 months

Amount: \$447,887

This award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).

Surveys represent a key source of data in a wide range of scientific disciplines, and are often an important input in public and corporate decision-making processes. While the primary use of survey data is to describe the population from which the sample was drawn, analysts are increasingly using those data to make inference about statistical models. This use of survey data in model estimation and analysis is referred to as analytic inference for survey data. Survey data producing organizations usually recommend that survey weights be used in all model fitting with survey data, in order to account for the sampling design and the post-sampling adjustments. This often significantly complicates the work of data analysts, in large part because the effect of survey weighting in analytic inference is still incompletely understood. This research will explore some long-standing open issues in analytic inference for surveys. First, within the survey-weighted mode of inference, the PIs will investigate improved variance estimation methods and distributional theory. Second, they will evaluate a number of alternatives to survey weighting in model fitting and inference. This research will advance the state of knowledge in how to properly conduct analytic inference for survey data by bringing together concepts, methods and approaches from survey statistics and other areas of statistics. Ultimately, it will enhance the ability of data analysts to conduct perform analytic inference for survey data. Because of the importance of survey data collection and analysis in so many disciplines, the results of the proposed research will be widely applicable.

[Additional Information](#)

Integration of Spatial and Social Network Analysis in Vaccine Trials

0924479

Michael E. Emch

University of North Carolina at Chapel Hill

Total Award Duration: 30 months

Amount: \$301,183

This award is funded under the American Recovery and Reinvestment Act of 2009 (Public Law 111-5).

In 1985, a community-based individually-randomized oral cholera vaccine trial was conducted in Matlab, Bangladesh. Existing methods for measuring the effectiveness of this vaccine trial to the transmission of the disease are not robust. New methods for spatial analysis and social network analysis are needed to better understand the impact of vaccine trials and disease transmission. This project builds on earlier work that developed theory and methods for incorporating neighborhood-level variables into "ecological vaccine trials".

Dr. Michael Emch at the University of North Carolina at Chapel Hill will reanalyze the cholera vaccine trial data from the previous study using new measures of social and environmental connectivity developed with social network analysis software, a geographic information system (GIS), and satellite remote sensing technologies. These new measures will help determine: (1) how placebo incidence and protective efficacy (PE) vary within social networks; (2) how placebo incidence and PE vary spatially by different environmental contexts; and (3) how spatial, environmental, and social network information can be used simultaneously to assess the effectiveness of vaccines and disease risk. Four data sets, already collected by the investigators, will be used for in this study. These datasets include: (1) a large cholera vaccine trial database, (2) a comprehensive longitudinal demographic database from which the vaccine trial participants were selected, (3) an accurate household-level GIS database, and (4) high resolution Quickbird satellite imagery. Kinship ties between all households will be measured using a complete network design as well as environmental connectivity using satellite imagery and GIS techniques. These GIS and satellite databases, in conjunction with the demographic, social network, and vaccine datasets, will facilitate adding integrated spatial, environmental, and social components to this vaccine evaluation study. Reanalyzing the Matlab oral cholera vaccine trial using a spatial analytical framework and social and environmental connectivity measures can give new insight into vaccine effectiveness evaluation and disease transmission modeling. As a result, the theoretical and methodological issues addressed in this project will have important implications for future vaccine trials.

This study will reanalyze an existing cholera vaccine trial database using new measures of social and environmental connectivity developed with social network analysis software, a geographic information system (GIS), and satellite remote sensing technologies. These methods will help determine: (1) how placebo incidence and efficacy vary within social networks; (2) how placebo incidence and efficacy vary spatially by different environmental contexts; and (3) how spatial, environmental, and social network information can be used simultaneously to assess the effectiveness of vaccines.

[Additional Information](#)

Motivated Underreporting

0850445

Roger Tourangeau

University of Michigan Ann Arbor

Total Award Duration: 24 months

Amount: \$565,102

0850999

Frauke Kreuter

University of Maryland College Park

Total Award Duration: 24 months

Amount: \$92,804

This research examines three forms of survey measurement error and investigates the relations among them. The first form of measurement error affects questions designed to identify members of the population eligible for a given survey (for example, persons over 65 years old). Several studies find that members of the eligible population are underreported in screening interviews. Although no survey perfectly covers its target population, surveys aimed at specific subpopulations seem especially prone to undercover that particular population. The second form of measurement error involves filter questions. These are questions that, depending on how they are answered, either lead to additional follow-up questions or to the respondent's skipping out of the follow-up items. Many survey researchers believe that respondents are likely to give false answers to the filter questions in order to avoid the follow-up questions. As a result, many surveys ask the filter questions at the beginning of the questionnaire and administer the follow-up questions later on rather than interleaving the filter and follow-up questions. The final form of measurement error involves conditioning, or time-in-sample, effects. Over the last forty years, many survey researchers have suggested that respondents in ongoing panel surveys report fewer relevant events across waves of the panel survey and across time periods in a diary survey.

What the three phenomena appear to have in common is underreporting motivated by the desire to reduce the effort needed to complete the questionnaire. But it is not clear whether these forms of error result from something the interviewers do, something the respondents do, or both. The proposed studies use both new experiments and analyze existing data to try to pinpoint the locus of these effects (interviewers versus respondents) and to explore the effectiveness of different methods for reducing these errors. The project will contribute to the improvement of various national statistics that are derived from survey items affected by these problems. The project also will further the training of graduate students and contribute to the professional training of survey researchers at both institutions. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

[Additional Information](#)

Calibration with Estimated Controls

0924250

Richard L. Valliant

University of Michigan Ann Arbor

Total Award Duration: 24 months

Amount: \$ 125,000

The results obtained from surveys are used to estimate characteristics of populations. Calibration estimators use auxiliary information to improve the efficiency of these survey estimates. Calibration is used in the U.S. in the Census Bureau's Current Population Survey (CPS), the Bureau of Labor Statistics' Consumer Expenditure Survey, and a variety of other Federal surveys as well as surveys administered by other nations. Some of the key uses of calibration are to (i) reduce variances, (ii) correct for nonresponse, and (iii) correct coverage deficiencies of a sample frame.

A primary assumption with calibration is that the sample estimates are calibrated to true population values known without error. Often, however, these population values are estimates obtained from other surveys which could possess sampling variance. These estimated controls add variability to survey estimates and may produce bias if the source of controls is inaccurate. The use of estimated controls in surveys is common but is virtually never accounted for in survey variance estimation. This research will investigate methods that will account for this extra imprecision in variance estimates and also study the effects of bias in the population values and in the analytic survey itself. The specific goals of this research will be to (1) analyze the effect of benchmarks and analytic survey estimates that are biased due to coverage errors with extensions done for general regression (GREG) estimators and other calibration estimators; (2) develop variance estimators for GREG and other calibration estimators that account for estimated benchmarks and undercoverage by sample frames; (3) develop some practical rules for determining whether estimated benchmarks are too variable to be worth using; and (4) illustrate theoretical findings via simulation studies based on real survey data set.

[Additional Information](#)

FY 2008

Estimating RDD Survey Bias Using ZIP Code Matching to Census Data

0818931

Paul P. Biemer

Andrey Peytchev

Research Triangle Institute

Total Award Duration: 30 months

Amount: \$143,835

Low and declining response rates in household surveys provide the potential for nonresponse bias, but assessing the nonresponse bias is a challenging problem when characteristics of nonrespondents are unknown. For example, for random-digit-dial (RDD) surveys, information about nonrespondents may be limited to just the telephone number. To address this issue, the census geocoding (CG) method has sometimes been used. This method appends aggregate census information for small geographic units (for example, census blocks, block groups, or tracts) to nonresponding cases located in those units. These census data can then be used to evaluate nonresponse bias or adjust for it. Little is known about the effectiveness of the CG approach, but it is often used in nonresponse evaluations, particularly for RDD surveys. While the CG method is an important tool for recovering information from nonrespondents, it relies on the degree to which its assumptions are met. This research project will investigate the efficacy of the CG method using RDD and face-to-face survey data and will develop guidelines for implementing it so as to maximize its effectiveness for use in household surveys. One of the goals of the proposed project is to investigate the minimum amount of information required to apply the CG approach effectively. As an example, if only a telephone number is known, census aggregate data closer to the telephone exchange-level must be used rather than at the block-level. Finding a method that can be used effectively to evaluate and correct for nonresponse bias in household surveys will be an important contribution to survey research.

Much of social science relies on probability-based surveys, which in turn rely on the ability to obtain survey measures from all sample members. Nonresponse threatens probability-based inference. This project will advance the knowledge about methods for evaluating and adjusting for nonresponse bias in surveys. The project will not only evaluate the CG method for use in telephone surveys but for virtually any survey where the CG method can be used. This includes any survey where information on nonrespondents is limited to an address, ZIP code, or telephone number and very little else. Many household surveys use address frames where all that is known for a frame unit is the household family name, the address and, in some cases, a telephone number. The CG approach can be applied in these situations to evaluate the bias due to nonresponse. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Acoustic Properties, Listener Perceptions, and Outcomes of Interactions between Survey Interviews and Sample Persons

0819734

Frederick G. Conrad

Robert Groves

University of Michigan Ann Arbor

Total Award Duration: 24 months

Amount: \$125,780

0819725

Jose R. Benki

Michigan State University

Total Award Duration: 24 months

Amount: \$108,801

While the sample survey is one of the cornerstones of social science research methods, the tool is experiencing large declines in participation of sampled persons. This project studies participation in cross-section Random-Digit Dialed (RDD) telephone surveys. The declines have led to such significant research cost increases that serious consideration of terminating basic surveys is occurring.

One large source of variation in RDD cooperation rates is the interviewer. In centralized telephone interviewing facilities interviewers are often assigned similar mixes of cases, yet obtain very different response rates. All the stimuli yielding a decision to participate in an RDD survey must be delivered through audio channel through words, pitch, inflection, and pacing of the interviewers' speech. These attributes are not the traditional objects of study of survey methodologists. Hence, the project combines insights from speech science, phonetics, psycholinguistics, and survey methodology. A central conceptual framework utilizes the companion notions of tailoring, convergence, and similarity in cooperative dyadic communication.

5,000 digital audio recordings of RDD telephone interviewer introductions (from two data collection organizations, 6 different surveys, and 165 different interviewers) are transformed into a quantitative data set suitable for dynamic and static statistical models of response propensities. Acoustic measures motivated by concepts from speech science are extracted using Praat acoustic software; raters make judgments of perceived attributes of speakers using both concepts central to leverage-salience theory and the social psychology; and word and disfluency rates are extracted from transcriptions of the audio files.

The resulting data set is a relational one with records at the levels of the interviewer, respondent (case), contact, and conversational turn. Hierarchical and survival models are built using the combined conceptual frameworks predicting the final participation decision of a sample case.

While the research emphasizes basic science, it has potential practical outcomes. Academic survey research is a crucial tool for a \$15 billion commercial research industry in the US; this project's success can contribute to the health of this commercial sector.

[Additional Information](#)
[Additional Information](#)

Accommodating Individual Differences in Software Designed for Location-Based Survey Tasks

0822002

Sarah M. Nusser

Leslie Miller

Elizabeth Murphy

Iowa State University

Total Award Duration: 36 months

Amount: \$230,189

The use of digital maps, imagery and global positioning system (GPS) receivers to support a range of field survey activities is perceived as an area that has potentially high payoff in terms of reduced costs and improved data quality for national surveys. At present, however, appropriate survey methods for effectively using these resources are lacking. The long-term goal of this research is to develop a theoretically grounded yet practical approach for designing computer-assisted survey instruments that rely on geospatial resources to support location-based field tasks. The project will contribute to this goal by developing naturalistic decision-making models for distinct location-based survey tasks that reflect the diversity of strategies that result from individual differences (e.g., in spatial ability), and by developing model software interfaces that accommodate variation in strategies for using computers and integrating environmental and digital geospatial information. The work will begin with building naturalistic decision-making models to describe a survey task referred to as address verification, which involves evaluating whether the location of a housing unit on the ground is accurately depicted on a map. The model will describe how task implementation varies in relation to individual abilities, such as mental rotation of spatial objects, perspective taking, and spatial memory. The approach involves outlining a high-level naturalistic decision-making model to describe cues (information) available to and selected by a field staff member, hypotheses (if any) that are formulated from this information, and the action selected that leads to completion of the task. Using this template, a series of more specific models will be created that describe how variation in spatial ability affects the decision-making process. Expert evaluation and field-based cognitive walk-through methods will be used to further build and refine models. With the understanding gained from these decision-modeling investigations, models of two user-interface views will be developed for the address verification task. The first user-interface view will be a basic interface that provides the required map functionality without taking any user differences into consideration. The second user-interface view will be designed to be an inclusive interface using what was learned in our modeling research. Information on variation in field behaviors (e.g., preferred cues, map use, task order) in relation to spatial ability will be used to create the set of screens for the inclusive user-interface views. Survey field experts and cognitive psychologists will evaluate these draft screens in relation to their knowledge and experience with field staff behaviors. This research is an important first step of an iterative process to develop interface design principles for location-based survey field tasks that rely on an inclusive software design approach to accommodate individual differences.

This research will make several substantive contributions to both basic and applied science. First, the project will begin developing a theoretical approach for effectively incorporating geospatial information resources in computer-assisted instruments for field surveys. Through this process,

naturalistic decision-making theory will be extended to include heterogeneity in individual abilities, specifically spatial ability. In addition, user interface design approaches will be broadened to handle the context of geospatial data for a task involving scientific protocols. The practical impact of the research will be felt in fielding surveys with improved field operations efficiency and enhanced quality of data and survey estimates for the national statistical infrastructure and other large-scale surveys. This project will provide students in human-computer interaction, computer science, and statistics with a substantive experience with interdisciplinary and inter-organizational research collaboration. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Differential Item Functioning (DIF) Testing with Estimation of the Latent Densities

0818722

Carol Woods

Washington University

Total Award Duration: 24 months

Amount: \$67,012

One important application of item response theory (IRT) is for identification of differential item functioning (DIF). DIF occurs when an item has different measurement properties for one group of people than another (e.g., men versus women), irrespective of group-mean differences on the latent variable. Typically, in IRT-based DIF testing, the latent variables are assumed to be normally distributed in the populations of people. However, this assumption may be untenable for some variables. The present research develops statistical methodology and software for IRT-based DIF testing in which the latent distribution for each group of people is estimated simultaneously with the item parameters.

This research fills a gap in extant methodology and is aimed at improving the match between data and the statistical models used to assess DIF. Questionnaires measuring latent variables which are potentially nonnormally distributed are used as part of making, for example, psychological and medical diagnoses and treatment plans, and hiring or admission decisions in businesses or universities. Assessment instruments constructed based on majority-group individuals may provide misleading results for underrepresented groups. The proposed methodology could be applied in these areas to promote fair, accurate, and inclusive measurement. The research is supported by the Methodology, Measurement, and Statistics Program, the REESE Program, and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Reverse Geocoding, Geographic Masking, and Geospatial Privacy

0814537

Paul A. Zandbergen

University of New Mexico

Total Award Duration: 30 months

Amount: \$121,980

The widespread availability of powerful geocoding tools in commercial Geographic Information Systems (GIS) software and the interest in spatial analysis at the individual level have made address geocoding a widely employed technique in many different fields. However, when locations of individuals and/or households are made public as published maps, the addresses associated with these locations can be determined using a technique known as "reverse geocoding." Therefore, individual-level data can often not be released due to geospatial privacy concerns and/or legal requirements. Techniques to preserve geospatial privacy exist and are collectively referred to as "geographic masking." The goal of this research is to develop robust methods for the protection of geospatial privacy of individual and/or household level data. The research will: 1) investigate the capabilities of reverse geocoding in identifying individuals and/or households based on point locations and associated attribute data; 2) determine which geographic masking techniques are most effective in meeting standards for geospatial privacy protection; and 3) determine the impact of geographic masking techniques on spatial analysis derived from point locations. The overall research design relies upon an empirical evaluation of reverse geocoding and geographic masking using a large sample of residential addresses in twelve different US Counties. Several types of geocoding will be employed, including address point and street geocoding using GIS software as well as commercial geocoding firms. Various existing and newly developed geographic masking techniques will be applied and evaluated for their robustness. This will provide an assessment of the type and magnitude of masking necessary to effectively preserve geospatial privacy. The effect of geographic masking will be determined using several large simulated datasets of typical individual-level data. This will provide insight into which geographic masking techniques provide sufficient levels of geospatial privacy protection, while at the same time maximizing the robustness of spatial-statistical analysis of the masked data.

The increased interest in individual level data in a number of fields and the widespread availability of high quality spatial data and software tools present a major challenge to the protection of geospatial privacy. Current practices in the use of geographic privacy protection techniques do not consider the abilities of software tools, and are not based on theoretical or empirical estimates of the likelihood that geospatial privacy is breached. This research will expand existing privacy concepts to geographic data in order to provide a consistent quantification of the degree to which geospatial privacy is protected. The research will provide specific guidelines for the use of specific geographic masking techniques in the light of the

current and future capabilities of reverse geocoding. This will provide much needed confidence in deciding whether and how to release individual-level locational data.

[Additional Information](#)

FY 2007

Doctoral Dissertation Research: Extending Respondent-Driven Sampling to Social Norm Research and Development of Web-Based RDS Software

0718377

Douglas Heckathorn

Cyprian Wejnert

Cornell University

Total Award Duration: 24 months

Amount: \$12,000

This project will make several contributions to Respondent-Driven Sampling methodology (RDS), a new statistical method now widely used in the study of hidden and hard to reach populations, through analysis of web-based RDS (WebRDS) samples of undergraduate students from two U.S. universities of varying selectivity. First, the project extends RDS to the study of social norms within the context of social network structure by mapping enforcement, compliance, and perception of knowledge-seeking norms onto the underlying social network structure of each population. RDS is especially well suited for such analysis because it provides a random sample of behaviorally defined ties from within a social network. The overall social network structure will be analyzed based on these ties and combined with survey questions regarding the social norms of respondents who make up those ties using the RDS homophily and affiliation indices. Consequently, the research will provide important insights into the interaction between social norms and social networks. Second, the project develops and tests WebRDS software, a fully automated online variant of RDS, for distribution to the general scientific community. Finally, while the RDS estimator has been shown to be asymptotically unbiased analytically and computationally, the hidden nature of most RDS study populations has prevented large scale, multi-site empirical validation. This project will provide such validation through comparison of RDS estimates with institutional data from two large undergraduate populations.

The larger research community will benefit from the methods developed in this project. First, the development of WebRDS software will provide for very fast and efficient sampling of electronically connected institutions. This will be especially useful for case control studies of epidemiological outbreaks where a reduction in time spent on data collection can prevent additional infections. The method of social norm research developed in this project can be applied to any outcome variable that is potentially spread through interpersonal contact. More immediately, the project will study knowledge-seeking norms in university students, a population recently identified by the U.S. Department of Education as declining in quality. As a Doctoral Dissertation Research Improvement award, this award also will provide support to enable a promising student to establish a strong independent research career. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Cultural Variability in Survey Question Processing and Response Behaviors

0648539

Allyson L Holbrook

Timothy Johnson

University of Illinois at Chicago

Total Award Duration: 24 months

Amount: \$549,036

Although considerable evidence suggests that cultural, ethnic, and racial background may have important effects on survey measurement error, the processes by which culture influences the survey respondents' behavior remain largely unexplored. To address this important problem, the study has four objectives. These include (1) investigating the effects of respondent race/ethnicity and culture on respondent verbal and nonverbal behaviors as indicators of response processing difficulties in survey interviews, (2) the effects of race/ethnicity on survey response styles, (3) the degree to which the effects of race/ethnicity on cognitive processing is moderated by question design features such as topic and format, and (4) the degree to which race/ethnicity effects can be accounted for by individual differences in cultural value orientations. Questions from major NSF-funded surveys, including the General Social Survey, the National Election Survey, and other major national social surveys, will be used to examine these objectives. Survey interviews will be conducted in the Chicago metropolitan area with a total sample of 800 adults from four distinct cultural groups: African Americans, Korean Americans, Mexican Americans and non-Hispanic Whites. Interviews will be conducted in English, Korean, and Spanish. Interviews will be audio and video recorded and subsequently behavior coded to identify verbal and nonverbal markers of cognitive processing difficulties, including problems with question comprehension, memory retrieval, response mapping, and socially desirable responding. The response latencies associated with each answer also will be electronically recorded. Hierarchical linear modeling will be used to analyze these data and address each study objective.

As the U.S. ethnic and racial structure continues to diversify, understanding how this demographic transition may complicate the interpretation of survey results will become an increasing priority. This research will contribute to the understanding of the mental process respondents engage in as they answer survey questions. The study also will provide crucial insight into racial/ethnic differences in these processes. It will offer insights about the nature of racial/ethnic variability in question comprehension across a variety of common types of survey questions, as well as racial/ethnic variability in response styles. It will address the role of race/ethnicity through the lens of cultural values, unpacking known racial/ethnic differences in extreme response style, acquiescence, socially desirable responding, and other such differences by linking them to cultural value orientations. The study also may lead to recommendations regarding best practices for the design of survey questions that minimize cultural differences in comprehension and response styles. The research is supported by the Methodology, Measurement, and Statistics Program, the Political Science Program, the Sociology Program, and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Can Internet Surveys Be Trusted?

0720101

Alan J. Krupnick

Resources For the Future Inc

Total Award Duration: 12 months

Amount: \$45,714

Probability-based internet survey panels potentially provide survey responses from a sample representative of the general population at low cost and with the ability to provide complex information treatments. Despite using random digit dialing to recruit the panelists, this technique raises concerns about the representativeness of survey responses given the multiple stages of panelist recruitment and retention. Relatively little also is known about how mode of administration, such as taking a survey on a computer, affects survey response. This research project will pursue answers to the following questions: (1) What nonresponse biases are introduced by mode-of-administration and sampling strategies? Specifically, do probability-based internet panels introduce more bias than other modes/sampling strategies? (2) After controlling for sample selection, do the mail and internet modes of administration affect the variable of interest and question response patterns? (3) Does the survey perform equally well for each mode/sampling strategy as gauged by standard tests of validity? To address these questions, four treatment approaches will be used to administer a survey on the willingness to pay for ecosystem improvements in the Southern Appalachian Mountains, crossing two sampling strategies (panel, random digit dialing) with two survey modes (paper, computer) and a fifth treatment in which respondents recruited through random digit dialing may take the survey on a computer (by internet) or by mail. If people who self-select into a mail (or internet) survey answer differently than people who are assigned to a mail (or internet) mode, this would imply that there are characteristics of people that can lead to biased willingness to pay depending on mode. State-of-the-art techniques will be used for evaluating non-ignorable nonresponse (or selection on unobservables) along with a rich variety of frame variables including Census tract characteristics. The project will provide greater understanding of the factors affecting response propensity and whether response propensity is correlated with the survey outcomes, focusing on the internet/panel combination.

The U.S. Office of Management and Budget has recently issued guidance that requires a non-response bias analysis for surveys expecting a response rate less than 80 percent. The guidance does not clarify how such analyses are to be conducted, however, and there is relatively little survey methodology research regarding how such analyses should be conducted and the factors that affect survey response. This project proposes and evaluates methods for testing the effects of nonresponse on sample representativeness. These issues also are of concern to the U.S. Environmental Protection Agency, which co-funded this project through a competitive grant from the National Center for Environmental Economics.

[Additional Information](#)

Responsive Design for Random Digit Dial Surveys Using Auxiliary Survey Process Data and Contextual Data

0719253

Sunghee Lee

Thomas Belin

University of California-Los Angeles

Total Award Duration: 12 months

Amount: \$90,000

This study develops a framework for a tailored random digit dial (RDD) telephone survey design that responds to findings from nonresponse bias studies. The lack of data for nonrespondents, a major challenge in RDD nonresponse studies, is overcome by using two types of data available regardless of response behavior - paradata and contextual data. Paradata are data created by the survey process itself and include the history of all calls made to each sampled number (e.g., number of calls placed, calling dates and times) and the survey design features used for each number (e.g., advance letter, monetary incentives, refusal conversion). Contextual data come from external sources (e.g., decennial census data) and are created by linking the geographic identifier (e.g., census tract, ZIP code) of all sampled telephone numbers to the external data at the corresponding geographic level. These data include various characteristics of the corresponding location, such as demographics and socio-economics, which are assumed to approximate the characteristics of individuals residing in that location. The study regards survey response behaviors as a stochastic process influenced simultaneously by the traits of the sample, the survey features, situational circumstances, and the perceived importance of these factors. Response behavior is modeled with variables in the paradata and contextual data and their interactions. The fitted model is used to predict how response behaviors change given hypothetical calling schedules and survey design features, allowing design tailoring for any subsequent surveys. An estimate of a bias indicator can be calculated for respondents and nonrespondents separately. The magnitude of nonresponse bias can be diagnosed by comparing these estimates.

Survey data are a vital source for quantifiable information about the population and are widely used by government agencies, policy makers, and social, political, and health science researchers. Advancing the current RDD survey practice is important given the popularity of these surveys in spite of ever-decreasing response rates. This study will provide a new design framework that takes a holistic approach to nonresponse and incorporates the understandings of nonresponse. The major element of the design tailoring is that it aims to increase response rates and decrease potential nonresponse bias. The results of this study will help organizations conducting RDD surveys improve their design process. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

A Probability-Sampling Framework for Modeling the Impact of Time-Varying Covariates on Event History Data

0720195

Stephen L Rathbun

University of Georgia Research Foundation Inc

Total Award Duration: 36 months

Amount: \$236,240

Event history data is comprised of the times of repeated events and can be found throughout the social and behavioral sciences, including investigations of substance abuse, child behavior, criminology, social unrest, and organizational ecology. Using methods collectively known as Ecological Momentary Assessment (EMA), recent advances in electronics have made it possible to collect richer kinds of event history data and at a much higher volume, improving the quality of data and offering challenges and opportunities for statistical research. In particular, methods are required for modeling the impact of time-varying covariates on the timing of repeated events, complex dependence relationships among repeated events, and variation among subjects in their responses to time-varying covariates. A probability-sampling framework will be developed for modeling the impact of time-varying covariates on event history data, supporting the construction of point process and survival models for data from an EMA of smoking behavior. The proposed framework may be applied whenever the estimating equations for a statistical model involve the integration of some function (e.g., hazard) of time-varying covariates over a sampling domain. The sampling domain is treated as a population of points, and the covariates are taken to be unknown but deterministic functions of time. A probability-based sampling design is used to sample the covariates from which a design unbiased estimator for the integrated function of the covariates may be obtained. Substituting this design-unbiased estimator into the estimating equations and solving for the parameter yields the proposed parameter estimator. This research aims to investigate the inferential properties of the estimator, to develop point process and survival models for event history data with random subject effects, and to investigate efficient probability-sampling designs for time-varying covariates.

New statistical methods will be developed for analyzing the impact of time-varying covariates such as psychological state and the subject's environment on the pattern of repeated behavioral events using observations collected by electronic devices. This research will support the development of models designed to improve our understanding of the mechanisms underlying addictive behaviors. Aside from the behavioral models considered here, the statistical framework may be applied more broadly in time series, geospatial statistics, and spatial epidemiology. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

FY 2006

Structural Misspecification in Latent Variable Models: Symptoms, Consequences, and Diagnostic Tests

0617276

Kenneth A. Bollen

University of North Carolina at Chapel Hill

Total Award Duration: 36 months

Amount: \$177,619

0617193

Stanislav Kolenikov

University of Missouri-Columbia

Total Award Duration: 36 months

Amount: \$122,146

Structural misspecifications refer to flaws in a statistical model such as omitted variables, having the erroneous number of latent variables to represent a concept, or formulating the incorrect set of relationships between variables. Structural misspecifications are understudied, particularly considering their frequency and their serious consequences for explaining, predicting, and understanding outcome variables. The project addresses four common structural misspecification problems that emerge in latent variable Structural Equation Models (SEMs) for which there have been no widely accepted solutions. These are: (1) negative sample estimates of variances, (2) sample correlation estimates with absolute values greater than or equal to one, (3) tests of dimensionality of latent variables, and (4) tests of the presence of latent variables such as random effects or method factors. The first two problems reflect either sampling fluctuations or structural misspecification. The last two problems are checks on the necessity for latent variables. Each of these problems present conditions under which the usual significance tests are not justified by classical maximum likelihood theory and significance tests. This research project examines the robustness of the usual classical significance tests for such problems and develops alternative significance tests that should be robust to these conditions in large samples. The project uses analytic results to justify the robust significance tests and employs empirical examples and Monte Carlo simulation techniques to examine the finite sample performance of the classical and the robust significance tests for a variety of correct and incorrect models. The project will lead to recommendations of the conditions under which researchers should employ classical and robust significance tests.

This project will provide researchers with diagnostic tools to assess the quality of their statistical models. For example, the project will provide the best way to test whether improper solutions such as negative error variance estimates or correlation estimates whose absolute value exceeds one are due to sample fluctuations or due to a more serious error in the model. It also will provide tests of whether two latent variables are really the same variable or whether some latent variables are really needed in a model. Due to wide proliferation of SEMs as an analytical tool in sociology, psychology, education, marketing, and other social and natural sciences, the refinements in the methodology of SEMs developed in this project will improve the quality of research and validity of the findings in those areas of science. This will lead to better

understanding of social and natural processes studied by means of structural equation models. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

[Additional Information](#)

Spatial and Small Area Estimation Problems with Application to Large-Scale Surveys

0604373

Sharon Lohr

Arizona State University

Total Award Duration: 36 months

Amount: \$210,612

The new American Community Survey (ACS), which contacts approximately 250,000 households across the United States each month, collects an unprecedented amount of information about a large number of households distributed spatially across the United States. Even with this large national sample size, however, the sample sizes in many geographic regions such as census tracts are too small for estimates to have acceptable variances. This project will result in new small area estimation methods that take advantage of the ACS's spatial structure and ongoing data collection to give more accurate estimates of characteristics such as poverty rate for geographic areas with insufficient ACS sample size. A new multivariate approach with continuous and binary variables will combine information from different regions, time periods, and data sources to yield more accurate small area estimates without additional data collection cost. New multivariate Bayesian spatial models, allowing nonstationarity in the estimation, will exploit the spatial information in the ACS to model spatial and temporal patterns in the data, improve small area estimates, and enable detection of changes over time. The investigators will study properties of computer-intensive methods for estimating mean squared errors of estimators, and develop numerically stable and computationally efficient methods for calculating mean squared errors.

Estimates from the ACS are used for income and poverty assessments, funding allocation, transportation planning, allocation of resources for the disabled, studying population patterns and migration, and many other purposes. The new statistical methods are expected to give more precise small area estimates from the ACS and other surveys that have spatial information, thereby improving the quality of the information available for making resource allocation decisions. The methods also will allow researchers in many subject areas to take advantage of the spatial and temporal aspects of survey data to study phenomena such as spatial distribution and local discontinuities in poverty, distributional changes following events such as hurricane Katrina, relationships between variables measured in other data sets (for example, environmental contaminants or criminal victimization) and data from the ACS, and changes in transportation patterns. The statistical methods may be used to model the spatial distribution of pollutants, detect environmental or introduced contaminants, and model effects of interventions in education, among many other applications. Large-scale surveys such as the ACS are expensive; the statistical methods developed in this project will help researchers extract more information from them without additional budgetary costs. The research is supported by the Methodology, Measurement, and Statistics Program, the Statistics and Probability Program, and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Dissertation Research: An Investigation of the Nexus of Survey Nonresponse and Measurement Error

0620228

Trivellore E. Raghunathan

University of Michigan Ann Arbor

Total Award Duration: 12 months

Amount: \$12,000

This project will examine the nexus between nonresponse and measurement errors in sample surveys. Recent research has not demonstrated a strong relationship between nonresponse rates and nonresponse bias. Nonetheless, best practices argue that researchers should attempt to maximize response rates. One voiced concern about practices involving nonresponse reduction is that reluctant people in the sample, successfully brought into the respondent data set through persuasive efforts, may provide data filled with measurement error. However, no study has looked at links among the propensity to be a respondent, true values on a question of interest, and the respondent's measurement error properties for that question. The research will fill this gap by addressing three questions. First, under what circumstances is nonresponse propensity related to the survey variables of interest? Is noncontact or refusal nonresponse more likely to induce nonresponse bias? Second, what is the relationship between nonresponse propensity, nonresponse bias, and measurement error? In particular, how do properties of questions and characteristics of respondents affect the nexus between nonresponse bias and measurement error? Is nonresponse propensity arising from noncontact or that from refusal more susceptible to a correlation with measurement-error bias? Third, can traditional statistical adjustments or analytic techniques remedy the problem? Data from two national and two regional surveys will be analyzed to answer these questions. The research will demonstrate when placing resources into nonresponse reduction is outweighed by increases in measurement error on key statistics. The research also will suggest changes to be made during field collection or in postsurvey adjustments that jointly account for two error sources: nonresponse and measurement error.

Sample surveys are the mechanism by which key indicators of the nation's well-being are created and through which mechanisms of personal and societal changes are tested. Understanding error structures of sample surveys - both in terms of when errors occur and the mechanism behind their occurrence - is critically important because conclusions made from surveys could be dramatically wrong. As nonresponse rates continue to rise, survey organizations are increasing their persuasive efforts in an attempt to maintain a representative sample on which inferences can be made. If these persuasive efforts actually lower the quality of data, then the additional funds for these efforts are misguided. The findings from this research will inform survey practitioners in how to conduct surveys and obtain better quality data and the scientific community that uses surveys to understand when survey-derived findings may be at risk. As a Doctoral Dissertation Research Improvement award, this award also will provide support to enable a promising student to establish a strong independent research career. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Latent Class Models of Measurement Error

0549916

Roger Tourangeau

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$241,897

0550002

Frauke Kreuter

University of Maryland College Park

Total Award Duration:

Amount: \$59,291

One of the most crucial activities in mounting a survey is the development and testing of the survey questions. Unfortunately, this process largely remains a qualitative endeavor, one that features reviews of the questions by experts, focus group discussions with a handful of participants, and small numbers of intensive "cognitive" interviews. Many researchers have questioned the effectiveness of these methods for identifying problem items. In addition, there is a disconnect between the qualitative data produced by these conventional questionnaire pretest techniques and the quantitative standards (such as reliability and validity) that the data are meant to address. This project will systematically assess the potential of a quantitative method -- latent class analysis (LCA) -- for use in developing and testing survey questions. The project seeks to answer several specific questions about the application of LCA models as a tool for evaluating survey questions by conducting a series of new experiments and analyses of existing data. The experimental studies will compare results from the LCA models against "gold standards," where true values for the variables being assessed are known. These studies will compare the conclusions from the LCA method against those from more conventional analyses. The analytic studies will apply LCA models to existing data sets and also use simulations to assess the robustness of the LCA method to violations of its underlying assumptions.

This project will advance basic knowledge about various strategies, including the use of latent class models, for questionnaire development. It will show whether these models can assess the measurement characteristics of survey items even in the absence of external validation data (such as administrative records). The project will compare the latent class models to conventional questionnaire development techniques and determine whether they can yield better questionnaires, reduced questionnaire development costs, or both compared to the traditional methods. The results of this research will be of value to the survey community, including the federal statistical agencies.

[Additional Information](#)

[Additional Information](#)

The Effects of Survey Presentation on Nonignorable Nonresponse and Measurement Error
0550385

Roger Tourangeau

Robert Groves

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$600,000

Every request to take part in a survey is framed in some way. This project consists of a set of experiments that investigate how the presentation of the survey request affects nonresponse and measurement error. The experiments are guided by a theory of survey participation (the salience-leverage theory) that claims that people decide whether to take part in a survey based on whatever aspects of the survey are made salient in the presentation of the survey request and on how they evaluate those features. Two initial experiments randomly vary the description of the topic and sponsor of the survey, with hypothesized effects both on nonresponse propensities and on reporting. In the third experiment, survey design features that can mediate or reduce the error-producing influences of the survey topic and sponsor will be examined. Thus, the project experimentally tests mechanisms producing nonresponse bias and measurement errors and, once these effects have been documented, provides guidance to the survey practitioner about how to reduce their impact.

While the research is theoretically motivated and features experimental control, there are important practical implications of the work for the federal statistical agencies and the larger survey community. Sometimes estimates of key social indicators (e.g., the prevalence of rape or the frequency of defensive use of handguns) vary widely across surveys. The effects explored in this project may help explain these discrepancies. In addition, this work will a) help agencies conducting surveys anticipate when different sponsors may obtain different results, b) provide evidence about potentially harmful effects on nonresponse error and measurement error of emphasizing a single purpose of a survey, and c) produce evidence regarding design features that can reduce the effects of the presentation of the survey on nonresponse and measurement error.

[Additional Information](#)

Regression Diagnostics in Survey Data

0617081

Richard L. Valliant

University of Michigan Ann Arbor

Total Award Duration: 24 months

Amount: \$150,000

Diagnostics for linear regression models are included as options in many statistical packages and now are readily available to analysts. However, these tools are generally aimed at ordinary or weighted least squares regression and do not account for stratification, clustering, and survey weights that are features of data sets collected in complex sample surveys. The ordinary least squares diagnostics can mislead users because the variances of model parameter estimates will usually be estimated incorrectly by the standard procedures. The variance or standard error estimates are an intimate part of many diagnostics. This research will adapt existing diagnostics for use with survey data, and, where necessary, develop new ones. This project also will study the properties of existing linear regression diagnostics when they are applied to complex survey data. Extensions are needed to cover both clustered and unclustered data. The particular techniques to be studied are: leverages for linear regression and their heuristic cutoffs for influence; distributions of leverages, including histograms and quantiles; modification of unit-deletion measures of influence on model parameter estimates and predicted values and the rules-of-thumb used to identify influential observations; change in standard error estimates due to deletion of an observation or groups of observations; and extension of collinearity diagnostics, including variance inflation factors and variance decompositions for parameter estimates.

The data collected in many surveys sponsored by U.S. government agencies and other domestic and international organizations are used to fit statistical models. These models are used to understand the correlates of disease, unemployment, education achievement levels, and other topics. The surveys are typically stratified, single or multistage surveys where units can have substantially different survey weights. Some examples of substantive areas are medical conditions, expenditures for medical care, the social welfare of families and children, and the status of progress in education. Evaluation and improvements to existing methods of model-fitting and diagnosis are important in order to make the most of the data that are collected in these surveys and to avoid conclusions that may be misleading or erroneous. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Analysis of Survey Data Using Imputation for Nonrespondents

0705033

Jun Shao

University of Wisconsin-Madison

Total Award Duration: 36 months

Amount: \$216,389

Imputation is a popular technique in handling nonresponse in surveys. This project focuses on the development of imputation methods that produce approximately unbiased and efficient survey estimators when imputed values are treated as observed data and standard methods are applied to compute the survey estimators. Various imputation methods will be studied, such as the nearest neighbor imputation, kernel nonparametric regression imputation, empirical likelihood, and techniques of handling measurement error. Emphasis will be placed on the study of multivariate survey variables and/or multivariate covariates, and problems with nonresponse in not only the main survey variables but also the covariates. For each imputation method, variance estimation that takes nonresponse and imputation into account will be studied, using a direct derivation approach or a replication method (such as the jackknife, the balanced half samples, the random groups, and the bootstrap) that contains a re-imputation component to assess the variability caused by imputation. In particular, some shortcut replication methods that reduce the amount of computation will be investigated.

Many statistics and government agencies collect data through surveys. Most surveys have nonresponse. Item nonresponse occurs when some sampled units cooperate in the survey but fail to provide answers to some questions. Imputation techniques, which insert values for nonrespondents, are commonly used compensation procedures for item nonresponse. In some cases, when auxiliary information is properly used, imputation increases statistical accuracy. An essential requirement for an imputation method is that one can obtain unbiased (or approximately unbiased) survey estimators and their variability estimators by treating the imputed values as observed data and using the standard estimation formulas designed for the case of no nonresponse. This requires developments on imputation methodology and statistical analysis procedures to take nonresponse and imputation into account. Since most of the proposed research topics are motivated by problems in survey agencies such as the Census Bureau, the Bureau of Labor Statistics, Westat, and Statistics Canada, results obtained from the proposed research will have significant impacts on the imputation and variance estimation methodology for these survey agencies. The research is supported by the Methodology, Measurement, and Statistics Program, the Statistics and Probability Program, and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

FY 2005

A Nonparametric Approach to Population Size Estimation for Multiple System Capture-Recapture Surveys

0518904

Song X. Chen

Iowa State University

Total Award Duration: 26 months

Amount: \$299,496

In surveys aimed at estimating the size of a population, there often are individuals in the target population who are unaccounted for in the survey. If such undercounts are not adjusted for, a misrepresentation of the population results. This has led to the development of multiple system surveys, often called capture-recapture surveys. Estimation for such surveys requires a number of restrictive assumptions on unobserved characteristics of the population being studied, and it is of interest to develop estimation and inference methods that are less sensitive to those assumptions. This project intends to relax several of these fundamental assumptions with a unified nonparametric framework based on the kernel regression estimator. A novel nonparametric estimator of the capture probability will be developed that admits both continuous and categorical variables, and can be used in an unequal-probability survey context. This estimator then will be applied in several areas of population size estimation, including in the adjustment of existing population estimators, the correction for frame errors, and in model and assumption checking.

Estimating the magnitude of a population is an important goal of many surveys, including national censuses, surveys of special populations, and many wildlife studies. A large and important application is the U.S. Decennial Census, which is designed to provide accurate counts on the total and subtotal of various groups of the U.S. population. Population size estimates from such censuses and studies are all subject to potential undercount and, more rarely, overcount. For example, analysis of U.S. Census data shows that certain sections of the U.S. population are much harder to enumerate than the rest of the population. The goal of this research is to produce robust and efficient population size estimators that adjust for the presence of undercounts and other types of survey errors, and are appropriate for a wide range of survey settings. By combining methods from several areas of statistics, this research ultimately will contribute to the development of new statistical methodology that is directly relevant to on-going work at federal statistical agencies. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Animated Agents in Self-Administered Surveys

0551300

Frederick G. Conrad

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$189,998

0551294

Michael Schober

New School University

Total Award Duration: 36 months

Amount: \$190,002

This project examines the impact of animated or conversational agent technology on web-based surveys. Animated agents are anthropomorphic software objects in the user interface that, in their most advanced implementations, produce gestures, display facial expressions, and move their eyes in coordination with their speech. They have been shown to improve user performance in some task domains (e.g. tutoring). When might this technology help or hurt survey data quality and respondents' satisfaction? How sophisticated must the agents be in order to provide benefit--or harm? In a web survey, an agent might motivate respondents to participate and complete the task, much like human interviewers do, and help respondents understand the survey questions as intended while allowing respondents to participate at their convenience, as in ordinary web surveys. On the other hand, the presence of an agent might discourage honest responding to questions about sensitive topics, much as human interviewers have been shown to do. In this project a series of laboratory experiments examine more and less human-like agents that ask questions about respondents' sensitive and non-sensitive behaviors. The studies contrast data quality and user satisfaction in non-agent web surveys to those with interface agents that vary in their dialogue capability, the degree to which they provide visual and spoken cues about their internal states, and the degree of intentionality in their speech. The agents are simulated with software that converts a video image of a live interviewer into an animation in real time; respondents thus believe they are interacting with a computer-generated agent even though there is actually a human behind the "agent." In the experiments about non-sensitive behaviors, respondents answer on the basis of fictional scenarios so that the accuracy of their answers can be determined. One possible outcome is that agents with greater dialogue capability will promote interactions that lead to more accurate understanding and thus more accurate answers. In the experiments about sensitive behaviors, respondents answer about their own lives; more reports of sensitive behaviors indicate greater respondent candor. One possible outcome is that agents with more movement (lips, eyes, and eyebrows) will lead respondents to feel less private and therefore to answer less candidly than with agents whose movement is limited. Respondents' satisfaction is measured with a post-interview questionnaire; the impact of different agent features on how respondents communicate with the survey system is measured by detailed turn-by-turn coding of all dialogue.

The practical impact of this work will be more informed decisions by survey researchers in adopting agent technology. Knowing when agents help and what features help the most can focus decisions about what interfaces to develop and which ones not to develop. For example, if

dialogue capability is more important to data quality and user satisfaction than other agent features, this could focus future development efforts on conversational competence of agents more than on visual realism. The theoretical impact of the proposed work will be in two areas. First, it will deepen our understanding of how verbal and non-verbal communication are interconnected, for example how verbal interaction is affected by the fidelity of the agents' facial display. Second, by comparing human-computer and human-human interaction the project will advance knowledge of how attributions of intentionality and human agency affect interaction more generally. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)
[Additional Information](#)

Measuring Spatial Segregation

0520400

Sean Reardon

Stanford University

Total Award Duration: 24 months

Amount: \$135,338

0520405

Stephen A. Matthews

Pennsylvania State Univ University Park

Total Award Duration: 24 months

Amount: \$102,140

The study of the causes, patterns, and consequences of racial and socioeconomic residential segregation requires the careful measurement of segregation patterns. This, in turn, requires that measures of segregation incorporate an understanding of spatial proximity/distance, something that is now possible due to the increasing availability, sophistication, and ease-of-use of desktop geographical information system (GIS) software. This project will develop and refine a new approach to measuring spatial (race/ethnic) segregation that addresses known flaws in other measures. The approach is based on the understanding that a segregation index is a measure of the extent to which the local environments of individuals differ in their racial or socioeconomic composition (or, more generally, on any population trait). This approach is operationalized by assuming each individual inhabits a 'local environment' whose population is made up of the spatially-weighted average of the populations at each point in the region of interest. Given a particular spatial weighting function, segregation is measured by computing the spatially-weighted racial (or socioeconomic) composition of the local environment of each location (or person) in the study region and then comparing the average compositions of the local environments of members of each group. This approach has a number of features that make it well suited to measuring spatial segregation. In particular, measures derived from this approach 1) are independent of choices of tract boundaries; 2) are sensitive to segregation patterns at any scale; 3) measure both spatial exposure and spatial evenness; 4) can be computed using any theory-based definition of spatial proximity and distance; 5) measure segregation among multiple racial/ethnic groups; and 6) are readily adaptable to the measurement of income segregation. This project will develop, evaluate, and refine a set of measures of segregation that a) are computable from available census and geospatial data, and b) enable researchers to measure segregation based on theory-driven definitions of social proximity and distance. In addition, the project will develop software tools, provide on-line training materials, conduct workshops, and publish descriptive analyses of segregation patterns and trends in order to enable the research community to use these measures.

Residential segregation by race and income remains a stubborn feature of U.S. society, and a growing body of scholarship shows that segregation is associated with negative outcomes for families, youth, and children living in isolated poor and minority neighborhoods. Income segregation, which results in the concentration of poverty, appears to have particularly negative effects for children in disadvantaged neighborhoods, including lower rates of high school completion and higher rates of teen pregnancy. Consequently, the study of racial and

socioeconomic residential segregation is an important area of scholarship with significant implications for social policy. This project will produce a) technical knowledge regarding the measurement of segregation; b) user-friendly software tools and training materials to enable other researchers to use the newly-developed methods of measuring segregation; and c) detailed descriptive data on patterns and trends of racial and socioeconomic segregation in U.S. metropolitan areas. These tools and descriptive results will enable researchers to better understand the causes, patterns, and consequences of residential segregation.

[Additional Information](#)
[Additional Information](#)

FY 2004

Extending Locally Dependent Item Response Models for Analyzing Psychological and Social Surveys

0417349

Eddie Ip

Wake Forest University School of Medicine

Total Award Duration: 24 months

Amount: \$70,000

This research concerns the development of flexible statistical and psychometric methods for analyzing item-response data from educational tests and social surveys. The project contains two main components. The first component is the methodological development of item response models and extensions, specifically the locally dependent hybrid kernel models for dichotomous and polytomous responses. These methods are applicable to items that do not function independently after conditioning on subject effect. Examples of locally dependent items include items that have a common reading stem (as in a reading comprehension test), or items that survey related quantities such as the frequency and the intensity of a feeling (as in a psychological test). The project is expected to advance standard item response models in several ways: (1) the handling of dependency within item clusters with separable submodels, (2) the incorporation of multiple scales, (3) the accommodation of item- and person-specific covariates, and (4) the exploitation of rating scales of items. The second component of this project addresses applications. Three data sets from three different areas - education, psychology, and health-related social study - have been identified, and each will be analyzed using the locally dependent models.

Item response models are increasingly used in calibrating scientific instruments used for measuring human traits and behavior. Recent examples include large-scale educational assessment and health-related quality-of-life research. This research supplements the most commonly used item response model - the unidimensional model - with flexible ways of dealing with potential minor deviations from the unidimensionality assumption. The benefits of the project include providing more flexible analytic methods that are compatible with the standard item-response models, which means that interpretability is retained, and adding novel features such as the handling of a wide range of response data (e.g., data with covariates or data that can be formulated as a rating scale). Given the increasing interest in the applications of refined item-response models to newer and often more complex tests and surveys, this project is expected to have an impact on improving both the quality of instruments and the subsequent analysis of gathered responses. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

Imputation for Survey Data with Ignorable or Nonignorable Nonresponse

0404535

Shao Jun

University of Wisconsin - Madison

Total Award Duration: 36 months

Amount: \$111,000

The proposed research focuses on imputation and variance estimation after imputation for survey data with nonresponse. The investigator will study different models that relate auxiliary variables and the variable to be imputed (e.g., parametric, non-parametric, and semi-parametric models); different response mechanisms (ignorable or non-ignorable); various imputation techniques (e.g., regression, nearest neighbor, and random imputation); different types of estimators (e.g., sample mean and sample quantiles); and different types of data (e.g., cross-sectional, clustered, or longitudinal data). The investigator will also study a pseudo empirical likelihood imputation method that provides more efficient survey estimators than other imputation methods. For each imputation method, variance estimation that takes nonresponse and imputation into account will be studied, using a direct derivation approach or a replication method (such as the jackknife, the balanced half samples, and the bootstrap) that contains a re-imputation component to assess the variability caused by imputation.

Many statistics and government agencies collect data through surveys. Most surveys have nonresponse. Item nonresponse occurs when some sampled units cooperate in the survey but fail to provide answers to some questions. Imputation techniques, which insert values for nonrespondents, are commonly used compensation procedures for item nonresponse. In some cases, when auxiliary information is properly used, imputation increases statistical accuracy. An essential requirement for an imputation method is that one can obtain unbiased (or approximately unbiased) survey estimators and their variability estimators by treating the imputed values as observed data and using the standard estimation formulas designed for the case of no nonresponse. This requires developments on imputation methodology and statistical analysis procedures to take nonresponse and imputation into account. Since most of the proposed research topics are motivated by problems in survey agencies such as the Census Bureau, the Bureau of Labor Statistics, Westat, and Statistics Canada, results obtained from the proposed research will have significant impacts on the imputation and variance estimation methodology for these survey agencies.

[Additional Information](#)

Estimation and Testing for Associations with Multiple-Response Categorical Variables from Complex Surveys

0418688

Thomas M. Loughin

Kansas State University

Total Award Duration: 24 months

Amount: \$53,170

0418632

Christopher R. Bilder

University of Nebraska-Lincoln

Total Award Duration: 24 months

Amount: \$ 42,594

When two or more categorical variables are measured, questions naturally arise regarding the associations among them. Well-established methods, such as Pearson chi-square tests for independence and loglinear models, have been developed to assess the association structure between "single-response" categorical variables. When one of these categorical variables arises from a survey question which asks respondents to "choose all that apply," the analysis is not as straightforward because survey respondents may respond positively to more than one item from the list and the responses are likely to be correlated, creating a "multiple-response" categorical variable. Furthermore, when the survey data arises from a complex survey design, there currently are no statistical analysis methods available to analyze association structures involving multiple-response categorical variables. This research project will develop a new set of statistical analysis procedures for testing and estimating associations and modeling multiple-response categorical variables arising through complex survey sampling. The research will build upon recently developed methods for multiple-response categorical variables in the simple random sample case. Rao-Scott adjustments, common in the analysis of associations among ordinary single-response categorical variables from complex survey sampling, will be extended to develop Pearson-type tests of associations involving multiple-response categorical variables. Odds-ratio-based measures of association and corresponding linearization-based standard errors will be derived to measure level of association. Marginal generalized loglinear models will be developed that allow the association structure to be described in terms of main effects and interactions due to the factors represented by the multiple-response categorical variables. Model-based tests for goodness-of-fit and estimates of odds ratios will be developed using asymptotic techniques. Adequacy of all methods developed will be examined by means of simulation.

Society is inundated with surveys, many of which include questions that invite respondents to "choose all that apply" from a series of items. This research will provide survey analysts with an essential set of statistical analysis tools for analyzing data from questions of this type, for which there currently is no good alternative. It also will lay the groundwork for future research including extensions to more varied data structures and the handling of missing data. Because surveys are such an integral part of our society's information-gathering and exchange system, the impact can be expected to be far-reaching, affecting areas such as public health, political science, criminology, sociology, demography, business, and technology. Any institution that uses statistically-designed surveys and includes "choose all that apply" questions stands to benefit

from the tools provided by this research. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#) [Additional Information](#)

Adaptive Sampling Designs in Network and Spatial Settings

0406229

James Rosenberger

Pennsylvania State Univ University Park

Total Award Duration: 36 months

Amount: \$300,000

The purpose of this research project is to develop new adaptive sampling designs and inference methods for sampling in network and spatially structured populations. Adaptive sampling designs are those in which the procedure for selecting the sample can depend on values of variables of interest observed during the survey. In spatial settings, that can mean adaptively adding new units to the sample in the vicinity of high or otherwise interesting observed values. In network or graph settings, links can be adaptively followed from interesting sample nodes to add new nodes to the sample. A variety of new sampling procedures, together with design and model based estimation methods, will be investigated in the study. A new, flexible and versatile class of adaptive designs, termed "active set adaptive sampling," was found during the preliminary work toward this project. Designs in this class have certain advantages over adaptive cluster sampling and some of the traditional network sampling designs in being more flexible, allowing for control of total sample size and not requiring complete inclusion of connected components. Design-unbiased estimates are possible with some of these designs, providing inferences that are robust against assumptions about the population. These designs lend themselves toward model-based inferences as well and can be used in some situations to help ensure that the assumptions for the model-based inferences are met. This project will advance the theory and methodology of adaptive sampling and in particular will fully investigate and develop several categories of new adaptive sampling designs within this class and develop and evaluate design and model based inference methods for use with adaptive designs of all types.

With adaptive sampling designs, the study design can change in response to the values and patterns observed during the study. For example, in a study of an at-risk hidden human population, social links from particularly high-risk individuals can be followed to add more individuals to the sample; in a survey of an unevenly distributed natural resource, new observations may be adaptively made in neighborhoods of high observed abundance. In previous work it has been established that in many situations the theoretically optimal sampling strategy is an adaptive one. Specific adaptive designs, such as the adaptive cluster sampling designs developed in a previous project, have been shown to give substantial gains in precision or efficiency over conventional strategies for certain types of populations, in particular rare, clustered ones. The results of the proposed research will provide research tools for other scientific fields, including the biological, environmental, health, and social sciences. Each of these fields has to deal with populations that are difficult to sample by conventional means because of their unpredictably uneven spatial and network structures. The sampling methods resulting from this project have applications to many situations of importance to society, including studies of hidden populations such as those at risk for HIV/AIDS, environmental assessment and monitoring, biological surveys, natural resources explorations and inventories, Internet surveys, rapid response to natural and induced health threats, studies in human social behavior, and archaeological studies. [Additional Information](#)

Model-based Properties of Replication Variance Estimators for Sample Surveys
0416662

Richard Valliant

University of Michigan Ann Arbor

Total Award Duration: 12 months

Amount: \$85,000

Replication variance estimation in surveys of finite populations is a standard tool of survey statisticians and researchers. Two of the most common methods used are the jackknife and balanced repeated replication (BRR). There is a substantial amount of theory available for the replication methods when they are implemented in standard ways described in textbooks and journal articles. In practice, however, these methods are operationalized in ways that often do not fit the standard theoretical requirements. In the jackknife, for example, the basic approach is to delete one first-stage sample unit, compute an estimate based on the remaining sample units, cycle through all first-stage sample units, and compute a variance among the resulting set of estimates. In practice, groups of units are formed by combining units within or across strata. Entire groups are then dropped-out in order to compute a jackknife variance estimate. This project will evaluate methods used in practice and investigate potential improvements to current methods using the model-based approach to finite population sampling. In particular, this work will study weight adjustments in the grouped jackknife and, to a lesser extent, partially balanced BRR and model-based properties of these two methods of variance estimation. The general regression estimator of population totals and other nonlinear estimators will be emphasized.

Theory for these grouped methods is limited, and it is unclear that the methods used in practice always have good theoretical properties. The ramifications of poor implementation of replicate variance estimation can be important because of the way that data bases are constructed. Weights for the full sample and for subsamples (or replicates) are created by the database constructor who appends the weights to each record in the database. Users are then instructed to use those weights to compute variances for all statistics, regardless of how complex. If a poor set of replicate weights is created, this affects all analysts. The research is intended to provide guidance on how to implement these grouped methods in surveys, particularly ones concerning the economy, health status of the population, and other applications in the social sciences. This research is supported by the Methodology, Measurement, and Statistics Program, the Statistics and Probability Program, and a consortium of federal statistical agencies as part of a joint activity to support research on survey and statistical methodology.

[Additional Information](#)

FY 2003

Topics in Small Area Estimation

0317589

Malay Ghosh

University of Florida

Total Award Duration: 36 months

Amount: \$203,643

0318184

Tapabrata Maiti

Iowa State University

Total Award Duration: 36 months

Amount: \$160,033

The term "small area" or "local area" usually refers to a small geographic area, such as a county, municipality, a census tract, or a school district. It can also refer to socio-demographic domains, such as a specific age-sex-race group within a large geographic area. Small area estimation has become a topic of growing importance in recent years because of the need for reliable small area estimates by many agencies, both public and private, for making useful policy decisions. This project is aimed at addressing several important aspects of small area estimation. One basic question to address in this context is how to use the survey weights (usually inverses of the selection probabilities of the different units in the population) in conjunction with models to arrive at meaningful small area estimators. While many exclusive model-based small area estimators have been proposed, design-assisted model-based small area estimators have been very sparse. The goal is to obtain such estimators for a very general class of distributions. The method will be used to find the proportion and the number of poor school-age children in different counties of the United States. This is a very important problem for many Federal agencies, especially for the Bureau of the Census. Another aspect of this research is to obtain small area estimates by combining results from two or more surveys designed to estimate the same quantity of interest. A typical application of this procedure consists of combining data based on the Current Population Survey (CPS) and the newly introduced American Community Survey (ACS) of the Bureau of the Census. The ACS is intended to replace the decennial census long form in the year 2010. Finally, Bayesian methods will be developed for detecting outliers in finite population sampling, especially in the context of small area estimation.

The broader impact of this research is that it aims to achieve an interface between survey methodology and survey practice. As an immediate example, the research findings have direct bearing on small area income and poverty estimation as well as small area estimation by combining estimates from two surveys such as the CPS and the ACS. The findings also should be of interest to staff at the National Center for Health Statistics who are interested in estimating the proportion of uninsured people of different ethnicities, proportion of people under Medicaid, and so on.

[Additional Information](#); [Additional Information](#)

FY 2002

Identifying Causal Mechanisms Underlying Nonignorable Unit Nonresponse Through Refusals to Surveys

0207435

Robert M. Groves

Mick P. Couper

Eleanor Singer

Stanley Presser

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$348,122

This project is a set of randomized experiments aimed at systematically both producing and eliminating unit nonresponse error in survey estimates. Specifically, three attributes of the survey request (topic interest, survey sponsor, and monetary incentives) will be experimentally manipulated in concert with choice of target population. Sampling frames containing persons with known characteristics (e.g., occupational groups, interest groups, groups of consumers of specific products or services) will be used. Randomly identified subsamples of these groups will be asked to participate in self-administered surveys on topics of relevance to the frame and topics irrelevant to the frame. Crossed with this factor, the sponsorship of the survey will be experimentally varied, with one sponsor relevant to the frame and one irrelevant to the frame. Finally, the use of a monetary incentive will be crossed with both of the other factors to measure the effects of extrinsic benefits of participation. The key hypothesis is that topic interest and sponsorship act to produce nonignorable nonresponse when the surveys contain items relevant to the frame, and that monetary incentives act to reduce the magnitude of nonresponse error by bringing into the respondent pool sample persons with low topic interest and minimal affect toward the sponsor. The effects of sponsor affect and topic interest are expected to be additive; monetary incentives are expected to counteract the nonignorability influences of both factors.

The practical importance of this work to statistics based on surveys is: a) to help agencies conducting surveys anticipate when different sponsors may obtain different results; b) to provide evidence about potentially harmful effects on nonresponse error of interviewers' emphasizing single purposes of a survey; c) to produce evidence regarding the ameliorating effects on nonresponse error of monetary incentives; and d) to test a conceptual structure that will help survey sponsors anticipate when nonresponse rates will affect error and when they will not. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

Testing for Marginal Independence Between Two or More Multiple-Response Categorical Variables

0233321

Thomas M. Loughin

Kansas State University

Total Award Duration: 12 months

Amount: \$28,065

0207212

Christopher R Bilder

Oklahoma State University

Total Award Duration: 12 months

Amount: \$51,812

Questions that ask respondents to "choose all that apply" from a set of items occur frequently in surveys. Categorical variables that summarize this type of survey data are called multiple response (or pick any/c) categorical variables. It is often of interest to test for independence between two categorical variables. When categorical variables can have multiple responses, traditional Pearson chi-square tests for independence should not be used because of the within-subject dependence among responses. This research will provide methods to test for independence between two or more multiple-response categorical variables. A modified version of the Pearson statistic will perform the test, and bootstrap procedures will provide approximate sampling distributions. First and second-order corrections will allow for chi-square distribution approximations to the sampling distribution. Generalized log linear models and multivariate binomial logit-normal models will provide a model-based approach for the test of independence.

Many survey questions are asked in a multiple-response manner. Examples include: "What types of cars do you own?" and "For what criminal offenses have you been arrested?" Other questions naturally fall into a multiple-response format, but some researchers avoid asking them in this format due to statistical analysis problems. For example, most survey questions dealing with ethnicity allow respondents to make only one choice, which is entirely inappropriate in today's highly multicultural population. Other researchers may analyze multiple-response questions as if they came from single-response categorical variables, which can lead to very conservative tests of independence. This research will allow researchers to incorporate these types of questions into surveys and use statistically correct methods of analysis. The research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

[Additional Information](#)

Theory and Methods for Nonparametric Survey Regression Estimation

0204642

Jean D. Opsomer

Iowa State University

Total Award Duration: 24 months

Amount: \$65,038

0204531

F. Jay Breidt

Colorado State University

Total Award Duration: 20 months

Amount: \$71,166

This research project develops new methods for the efficient use of auxiliary information in complex surveys, based on nonparametric regression techniques. Current practice relies on parametric regression techniques, which have good efficiency if the regression model is well specified, and which have a number of appealing operational features. The nonparametric techniques share these operational features, lose little efficiency when the parametric specification is correct, and gain efficiency when the parametric specification is incorrect. The project increases the scope of applicability of the nonparametric regression estimation approach, by considering complex survey designs, varying types of auxiliary information, and alternative smoothing techniques. Specifically, the project investigates multi-stage surveys with cluster or element-level auxiliary information; multivariate auxiliary information; and alternative smoothing techniques. Parametric and nonparametric techniques are blended using semiparametric additive models to provide a flexible tool for use in complex surveys.

Large-scale surveys are used to collect data in a wide range of fields, from studies of human populations to inventories of natural resources. Information external to the survey, such as administrative records or remote sensing, is often available. This research project makes it possible to incorporate auxiliary information easily and effectively into survey estimates, by using nonparametric regression methods. Nonparametric regression, sometimes referred to as smoothing, is widely used in other areas of statistics, but its use in survey estimation has been limited so far. The investigators show that incorporating auxiliary information into survey estimation through nonparametric regression can improve the precision of the surveys, often at reduced costs.

[Additional Information](#)

[Additional Information](#)

A Comparison of RDD and Cellular Telephone Surveys

0207843

Charlotte Steeh

Iowa State University

Total Award Duration: 12 months

Amount: \$176,296

Since very little research has examined the impact on surveys of the exploding growth of wireless communication devices, this study will assess the extent to which these devices are likely to change telephone surveys. This issue will be addressed by comparing the results of two national surveys, one using the usual list-assisted RDD sample and the other employing a sample of mobile telephone numbers. The questionnaire, which will be identical in both surveys, will include substantive items on important policy issues as well as inquiries about mobile telephone ownership and use. The analyses will search for significant differences between surveys along four dimensions-coverage, nonresponse, data quality, and relationships among variables. Methodological factors, such as the number of attempts, the percentage of sample numbers whose status as working or nonworking is indeterminate, and the effects of caller-id, will also be compared. The basic hypothesis underlying all analyses is that there are major differences between the two modes along each dimension.

Given that wireless communication devices will only become more widely used and more sophisticated in the very near future, it is necessary to determine how they might enrich and supplement the survey process. This research will provide initial evidence. Contact through mobile telephones promises to make hard-to-reach respondents more accessible and to give voice to groups either poorly represented or not represented at all in current surveys. By gauging the reactions of respondents to survey contacts via a cellular telephone, the project also will provide practical guidance on incorporating wireless devices into the survey process. Since methodologists now predict that future surveys will be multi-modal, mixing present and future wireless communication devices with fixed line telephones and the web, this study will describe both the opportunities and the pitfalls involved. The end result will be an expanded definition of surveys and better and more valid data upon which to base public policy. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

FY 2001

Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

0106914

Roderick J. Little

Trivellore E Raghunathan

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$355,280

Surveys with complex probability sampling designs involving clustering and stratification are a major source of empirical data for governmental and scientific use. The standard approach to survey inference is design-based, with statistical inferences being based on the sampling distribution with population values treated as fixed. This approach has been a powerful force for the development of objective statistical analysis of large surveys, with reliable operating characteristics under weak assumptions about the population. However, the design-based paradigm is too limited to handle: (i) the increased availability of data from a variety of sources, such as surveys, censuses and administrative records; (ii) increasing demands for analyses that go beyond simple descriptive information such as means and totals for large domains; (iii) the development and analysis of data masked to preserve confidentiality; and (iv) the analysis of data subject to unit and item nonresponse. These questions can be addressed by a model-based Bayesian approach, with models that capture the relevant features of the population under study and take into account important features of the sample design, and non-informative priors that limit subjectivity in the analysis.

Bayesian methods are enjoying a resurgence in statistics, with the development of computational tools that make them practically feasible. However, the application of Bayesian methods to sample surveys remains very limited. The goal of this research is to develop useful, practical Bayesian methods for sample survey inference that have good design-based properties. The dissemination of public use data files is crucial to the research community in order to conduct research that forms the basis for rational policy decisions. This research will develop methods for disseminating detailed micro-data files that greatly reduce the risk of disclosure of the identity of respondents to a data intruder. Methods will be based on multiple imputation of key variables, an approach that allows for valid statistical inferences using existing software and limits the degree of information loss. The methods will be tested on large government surveys collected by the National Center for Health Statistics and other federal agencies. This research also will develop Bayesian methods for three topics in the statistical analysis of complex surveys: (i) the analysis of surveys where the sampled units have differential probabilities of inclusion; (ii) the handling of unit and item nonresponse in surveys; and (iii) the analysis of samples collected using rotating panel designs. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

Small Area and Longitudinal Estimation using Information from Multiple Surveys

0105852

Sharon Lohr

Arizona State University

Total Award Duration: 36 months

Amount: \$150,064

National surveys such as the Current Population Survey (CPS) or the National Crime Victimization Survey give accurate estimates of poverty or criminal victimization at the national level. These surveys do not, however, contain sufficient sample sizes to give reliable estimates by themselves for "small areas" such as counties or minority groups, or to provide detailed information about events such as domestic violence that affect only a small part of the population. Current methods for estimating poverty in small areas incorporate auxiliary information from administrative sources such as tax records through regression. This approach assumes that the administrative data are without errors; it also does not incorporate information from other surveys or make use of longitudinal information.

This project focuses on combining information from multiple surveys, with possibly different sampling designs, to improve estimation in small areas. Thus, for estimating poverty in small areas, the CPS can be used in conjunction with the American Community Survey and other sources. Multivariate multi-level models that allow for missing data will be developed that make use of the correlations in the different surveys to increase the precision of the small area estimates, and theoretical properties of the estimates will be derived. Modifications of the multivariate models will allow information from longitudinal surveys to be combined as well, and allow longitudinal analyses from a panel survey to be supplemented by information from related cross-sectional surveys. The investigator will also develop methods for outlier detection and robust estimation using the multiple data sources.

Increasing amounts of information are available from surveys and other sources, and there is increasing demand from federal and local governments and from social scientists for estimates in small areas. The models developed in this research will combine information from different surveys and use longitudinal and spatial aspects of the surveys to improve the accuracy of small area estimates, with no additional data collection cost. The theoretical results derived in this research will also be useful for other areas of application such as genetics and quality improvement. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

Robust Small Area Estimation Based on a Survey Weighted MCMC Solution for the Generalized Linear Mixed Model

0106978

Ralph E. Folsom

Avinash Singh

Research Triangle Institute

Total Award Duration: 12 months

Amount: \$124,823

Folsom et al. (1999) developed a survey weighted hierarchical Bayes (SWHB) estimation methodology for fitting unit-level generalized linear mixed models and applied it to the National Household Survey on Drug Abuse (NHSDA). The SWHB solution for the logistic mixed model is robust against model misspecification because the small area estimates (SAEs) for any large sample areas are close to their robust design based analogs. It also assures that national aggregates of the SAEs are design consistent and, therefore, approximately self-calibrated to the robust design based national estimates. The use of unit level models also assures internal consistency of SAEs for different levels of aggregation even when different predictors are used at those levels. However, the Folsom et al. solution assumed that the survey design could be treated as noninformative after inclusion of certain covariates; i.e., the superpopulation model was assumed to hold for the sampled units. In the interest of robustness against model misspecifications, it is desirable to remove this assumption. The first goal of this research project is to improve the uncertainty measures of the SWHB solution by taking full account of the survey design effects. The second project goal is to improve the robustness properties of the enhanced solution by assuring exact calibration of the aggregated SAEs to the design consistent national survey estimate. To achieve these goals an approximate Gaussian likelihood is assumed for the joint sampling distribution of the input vector of survey weighted fixed and random effect estimating functions. In this approximate Gaussian likelihood, a design consistent variance-covariance matrix for the vector of estimating functions will be used to fully account for survey design. The second project goal is achieved by employing a 'calibrated' Markov Chain Monte Carlo (MCMC) algorithm with a Metropolitan Hastings step that exactly benchmarks the SAEs to the robust design based national estimates. Simulated data with fixed and random predictors that are not included in the analysis model will be used to compare the robustness of the calibrated and uncalibrated solutions against model misspecification. Also, the improved SWHB solution will be contrasted with other solutions on one or more large survey data sets, e.g., NHSDA, NHIS, BRFSS.

In spite of the wealth of information that is available at the national level, Federal, State and local agencies concerned with program planning face difficulties because of the lack of specific information at the local level. Typically, information is desired for States and for substate planning regions or counties. In principle, surveys that provide national statistics could be expanded so that the needed State and sub-state data were collected; however, government agencies seldom have the economic and infrastructure resources needed to collect this volume of data via a direct survey approach. Fortunately, new advances in statistics and increases in computing power offer a viable, affordable alternative to the prohibitively expensive direct survey approach and now permit the production of valid and reliable estimates for small areas. The goal of this project is to promote wider acceptance of model based SAEs for official

statistics by improving the uncertainty measures, by providing robustness against model misspecification, and by assuring the internal consistency of SAEs for different aggregation levels. This research is supported by the Bureau of the Census under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

Visual and Interactive Issues in the Design of Web Surveys

0106222

Roger Tourangeau

Reginald Baker

Mick Couper

Frederick Conrad

University of Michigan Ann Arbor

Total Award Duration: 36 months

Amount: \$366,794

The rapid growth of the Internet as a vehicle for surveys raises important questions about this new method for data collection. Web surveys are the latest in a series of methods in which computers administer questions directly to respondents. Like many of the earlier methods, Web surveys present questions visually and offer some capability for interacting with respondents. This project encompasses a series of experiments that examine the implications of the visual and interactive character of Web surveys. One set of experiments examines how respondents interpret visual cues in Web questionnaires. These studies test the general proposition that incidental visual features of the questions (e.g., the spacing of response options) can give rise to unintended inferences about their meaning. The studies examine specific hypotheses about the heuristics respondents use in interpreting visual features of questions. These hypotheses state that respondents expect response categories to proceed in a logical progression from left to right; that they expect spatially isolated options also to differ conceptually from the other options; and that they assume that items that look alike must somehow be conceptually related. Two additional experiments examine the effects of including images to supplement the text of a question. Images are necessarily concrete, and one experiment tests whether this leads respondents to interpret the question more narrowly than they otherwise would have. Another experiment tests whether items depicted in an image may serve as standards of comparison for respondents' judgments. The final series of studies examines when respondents take advantage of interactive features of a questionnaire. These experiments test three general hypotheses: Respondents are more likely to utilize information available to them interactively when 1) the information is easy to obtain, 2) it is clearly helpful, and 3) they are highly motivated to seek help.

Collectively, the results are likely to lead to practical guidelines for the design of Web surveys. For example, the initial experiments should help settle some important practical questions-Does it matter whether the response categories are arrayed vertically or horizontally? Does the spacing of the response options affect the answers? Does the use of color in response scales influence the distribution of answers? The final experiments could yield better methods for getting respondents to take advantage of features that might improve their answers. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey and Statistical Methodology Funding Opportunity.

[Additional Information](#)

FY 2000

Bayesian and Likelihood Based Multilevel Models for Small Area Estimation

9911485

Malay Ghosh

University of Florida

Total Award Duration: 36 months

Amount: \$182,058

9911466

Tapabrata Maiti

University of Nebraska-Lincoln

Total Award Duration: 36 months

Amount: \$45,069

This research project focuses on Bayesian and likelihood based multilevel models for small area estimation. These methods will be compared and contrasted against some of the existing methods, such as the pseudo maximum likelihood, penalized quasilielihood, etc. Some of the novel features of this research will be the use of stratum varying regression coefficients, new priors for the variance-covariance matrix rather than the standard Wishart prior, development of small area estimation models allowing measurement errors for covariates, use of hierarchical likelihood in the context of small area estimation, and the use of survey weights for small area estimation. One of the major applications of this project will be the estimation of income and poverty for states and counties, and possibly even for lower levels of geography such as census tracts and school districts (when data become available) between decennial censuses. However, the methods are fairly general, and can be applied to other studies as well. Among others, these methods will be applied to study youth unemployment for small areas based on the Scottish School Leavers Survey, effectiveness of schools and student character in an education survey conducted by the Inner London Education Authority, and a British Social Attitudes Survey.

The terms "small area" or "local area" are commonly used to denote a small geographical area, such as a county, a municipality, or a census division. They may also describe a "small domain;" that is, a small subpopulation such as a specific age-sex-race group of people within a large geographical area. In these days, there is a global need for reliable small area statistics both from the private and public sectors. There are increasing government concerns with issues of distribution, equity, and disparity. For example, there may exist geographical subgroups within a given population that are handicapped in many respects, and need definite upgrading. Before taking remedial action, there is a need to identify such regions, and accordingly, one must have statistical data at the relevant geographical levels. Small area statistics also are needed in the apportionment of government funds, and in regional and city planning. In addition, there are demands from the private sector since the policy-making of many businesses and industries relies on local socio-economic conditions. Thus, small area estimation techniques have global applicability, and are useful for diverse applications.

[Additional Information](#)

[Additional Information](#)

FY 1999

Developing and Testing a Computer Tool that Critiques Survey Questions

9977969

Arthur C. Graesser

University of Memphis

Total Award Duration: 36 months

Amount: \$205,990

The validity and reliability of answers to questions on a survey critically depend on whether the respondents understand the meaning of the questions. This project develops and tests a computer tool that assists survey designers in improving the comprehensibility of questions. The computer tool will have particular modules that diagnose each question in a survey on various levels of language, discourse, and world knowledge. For example, the critique identifies questions with low frequency words, vague or ambiguous terms, unclear relative terms, complex syntax, high working memory load, misleading presuppositions, and content that appears to be unrelated to the survey context. The computer tool will incorporate empirical findings and computational architectures in the fields of cognitive science, artificial intelligence, computational linguistics, discourse processing, and psychology. Some of these modules are so complex, technical, or subtle that they are invisible to the unassisted human eye, including experts in survey methodology, questionnaire design, and computational linguistics. This motivates the need for a computer tool to assist the research methodologist in revising questions and in learning about the complex mechanisms that underlie each component.

The computer tool will be useful to the extent that it provides an accurate and reliable diagnosis of problematic questions. The project will therefore evaluate the performance of the computer tool on several measures. Each module determines whether or not a particular question has a problem (e.g., unfamiliar technical term, working memory overload). These decisions will be compared with the decisions of experts. Other performance measures are needed because trained expert judges may miss subtle computational mechanisms. These other measures will assess whether the computer output can predict the behavior of respondents when they answer the questions: (a) behaviors of respondents that indicate they are having difficulty comprehending the question in a conversational interview (such as clarification questions of respondents) and (b) test-retest reliability of answers to questions when respondents answer a question on multiple occasions. Performance measures also will be compared for original questions, questions revised by survey methodologists who do not use the computer tool, and questions revised by survey methodologists who have had the benefit of using the tool. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey Methodology Funding Opportunity.

[Additional Information](#)

Small-Area Estimation - A Growing Problem for the Next Millennium

9978145

Patha Lahiri

University of Nebraska-Lincoln
Total Award Duration: 24 months
Amount: \$73,310

9978101

Jiming Jiang

Case Western Reserve University
Total Award Duration: 24 months
Amount: \$54,134

Large scale sample surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas. However, for effective planning of health, social, and other services, there is a growing demand to produce similar estimates for smaller geographic areas and subpopulations, commonly referred to as small-areas (or small-domains). The accuracy of small-area statistics is especially crucial when data are used to apportion government funds among various groups.

This project focuses on development of new robust small-area estimation methods and the associated model diagnostics. The estimation methods will be developed under general multi-level models which will be useful in solving a variety of small-area estimation problems. To address an important and yet largely neglected aspect of model validation and model selection associated with multi-level models, a test using a sample splitting technique is proposed. Splitting the sample into an estimation set and a validation set can also be used for assessing the actual power of the model. This area of research will continue to grow as social scientists find the need to use complex multi-level models to solve their problems.

The research is an outgrowth of the investigators' experiences with small-area estimation problems encountered by various federal, state, and private agencies. Importantly, this project will address a crucial practical problem underlying the work of many governmental and private institutions throughout the world. Further, this research on small-area estimation also will contribute significantly to the literature on survey sampling, generalized linear mixed models, empirical best prediction theory, linear empirical Bayes, variance component estimation, resampling methods, model diagnostics, higher order asymptotics, and statistical computing. Because of the interests of different types of researchers (e.g., survey samplers, main stream statisticians, social scientists), small-area estimation will remain one of the most intriguing problems in survey sampling as we advance into the next millennium.

[Additional Information](#)

[Additional Information](#)

The Cognitive Basis of Seam Effects in Panel Surveys

9907414

Lance J. Rips

Northwestern University

Total Award Duration: 36 months

Amount: \$209,654

This project investigates a type of error called the "seam effect" that occurs in national surveys and that affects the quality of their data. Some panel surveys, such as the Survey of Income and Program Participation and the Consumer Expenditure Survey, interview respondents three or four times a year; however, questions on these surveys ask for information about each of the preceding months. For example, a respondent might be interviewed in April and asked during that interview to provide information about his or her expenditures for each of the months of January, February, and March. The same respondent might be interviewed in July for expenditures during April, May, and June. Previous analyses of the data from these surveys show that month-to-month changes in respondents' answers are much greater when the data come from successive interviews than when they come from the same interview. In the example just mentioned, changes in the level of expenditures would be greater between March and April (data gathered from separate interviews) than between other adjacent months (data gathered from the same interview). Prior studies strongly suggest that these differences are not due to true changes between March and April, but are due to response error. The purpose of this project is to develop a model of this effect that can help predict its severity and that will aid in eliminating it or adjusting for it statistically.

The studies in the project investigate the seam effect using a procedure in which respondents answer questions about information supplied in the experiments themselves. In this way, the experiments control variables that might alter the size of the effect, and they monitor the respondents' accuracy. The strategy in these experiments is to vary separately factors that might affect respondents' memory for earlier information (e.g., the importance or salience of that information) and factors that might affect respondents' willingness to estimate or to guess at an answer. Because the first set of factors may have more impact on later parts of the response period and the second set of factors more impact on earlier parts, their combined influence can increase or decrease the size of the seam effect. These experiments test this hypothesis. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey Methodology Funding Opportunity.

[Additional Information](#)

Cognitive Issues in the Design of Web Surveys

9907395

Roger Tourangeau

Robert Tortora

Gallup Organization

Total Award Duration: 12 months

Amount: \$114,932

9910882

Mick P Couper

University of Michigan

Total Award Duration: 12 months

Amount: \$30,313

The development of new methods for collecting survey data, including Web surveys, may be ushering in a golden age for self-administered surveys. The new methods of data collection appear to offer the power and complexity of computerization combined with the privacy of self-administration. At the same time, because they do not require an interviewer, they may reduce other types of survey error and could dramatically lower the costs of conducting surveys. Still, there is mounting evidence that different methods of self administration can produce different results; these differences across methods of self administration seem to reflect apparently incidental features of the interface between the respondent and the electronic questionnaire. This collaborative research with Couper (SES-9910882) tests a theory to explain these effects of the interface. The key concept in the theory is that of social presence. To the extent that the method of data collection, its setting, or the interface gives the respondent a sense of interacting with another person, it will trigger motivations similar to those triggered by an interviewer. These motivations include the desire to avoid embarrassing oneself or giving offense to someone else, as well as enhanced motivation to complete the interview. The Web offer ample resources for attracting the interest of the respondent (color, animated images), but even apparently innocuous characteristics of the interface can create a sense of social presence, producing social desirability and related response effects.

This experiment will attempt to identify the features of the interface with a Web survey that create a virtual social presence. A sample of respondents will be recruited by telephone to complete a Web survey. The experiment will vary whether or not the electronic questioner is identified by name ("Hi! I'm John") and whether or not it offers explicit reminders of prior answers. The results of this first experiment may suggest follow-up experiments to be carried out if the budget permits. The main hypothesis to be tested in any follow-up studies is that the more the interface creates a sense of social presence, the more respondents will act as if they are interacting with another human being. This research is supported by the Methodology, Measurement, and Statistics Program and a consortium of federal statistical agencies under the Research on Survey Methodology Funding Opportunity.

[Additional Information](#)

[Additional Information](#)