International Total Survey Error Workshop 2011

ITSEW 2011 Abstracts

Session 2: Risk and paradata research	**Chair: Dave Dolson**
Aggregate and Systemic Components of Risk in Total Survey Error Models	John Eltinge
Paradata Collection Research for Social Surveys at Statistics Canada	François Laflamme
Session 3: Nonresponse and Measurement error I	**Chair: Paul Biemer**
Propensity Score Models for Nonresponse and Measurement Error.	John Dixon
How Much of Interviewer Variance is Really Nonresponse Error Variance? New Results from a National CAPI Survey in Germany	Brady West and Frauke Kreuter
Incorporating Nonresponse Propensity in a Markov Latent Class Measurement	Brian Meekins, Clyde Tucker and Paul Biemer
Session 4: Adaptive Survey Designs I	**Chair: Hélène Bérard**
An Active Collection using Intermediate Estimates to Manage Follow-Up of Non-Response and Measurement Errors	Jeannine Claveau, Serge Godbout and Claude Turmelle
Responsive Collection Design for CATI Surveys and Total Survey Error (TSE)	François Laflamme
Session 5: Adaptive Survey Designs II	**Chair: John Eltinge**
Optimizing CATI Call scheduling to minimize Data Collection Costs	Hussain Choudry, Mike Hidiroglou and François Laflamme

A Theoretical Framework for Adaptive Collection Designs	Jean-François Beaumont and David Haziza
Session 6: Estimation	**Chair: John Dixon**
Overview of error model for estimates of foreign-born immigration using citizenship and residence one year ago from the American Community Survey	Mary Mulry
Robust inference in two-phase sampling designs with application to unit nonresponse	David Haziza and Jean-François Beaumont
Estimation strategy of the National Household Survey	François Verret
Session 7: Interviewer effect and patterns for item non-response	**Chair: Brad Edwards**
Measuring Interviewer Effects on Survey Error in SHARE	Annelies Blom, Julie Korbmacher and Ulrich Krieger
Item Nonresponse in a Mail Survey of Young Adults.	Luciano Viera, S. Turner and S. Marsh
Computer Audio Recording: A Practical Technology for Managing Survey Quality	M. Rita Thissen, Hyunjoo Park and Mai Nguyen
Session 8: Health Surveys and TSE	**Chair: François Brisebois**
Proxy Pattern-Mixture Analysis of Missing Health Expenditure Variables in the Medical Expenditure Panel Survey	Robert Baskin, Samuel Zuvekas and Trena Ezzati-Rice
An Assessment of the Impact of Two Distinct Survey Design Modifications on Health Care Utilization Estimates in the Medical Expenditure Panel Survey	Steven Cohen, Trena Ezzati-Rice and Marc Zodet C

Balancing Confidentiality and Quality in Public Health Data	Lawrence Cox
Total Survey Error in Disability Assessments: Measuring Physical and Cognitive Capacity in the National Health and Aging Trends Study (NHATS)	Brad Edwards and Tamara Bruce
Session 10: Nonresponse and Measurement error II	**Chair: Brian Meekins**
Attrition and Selection of alteri Respondents in the pairfam panel	Ulrich Krieger
Nonresponse Bias Correction in Telephone Surveys Using Census Geocoding: An Evaluation of Total Error Properties	Paul Biemer and Andy Petychev
Non-Consent Error, Nonresponse Error, and Measurement Error: Total Survey Error in Linked Survey and Administrative Data	Joseph Sakshaug and Frauke Kreuter

Aggregate and Systemic Components of Risk in Total Survey Error Models

John L. Eltinge, U.S. Bureau of Labor Statistics, Eltinge.John@bls.gov

Key words: Complex and tightly coupled systems; Incomplete data; Informative missingness; Measurement error; Model identification information; Nonresponse; Operational error; Reporting error; Total statistical risk.

Extended Abstract:

1. Survey Risk, Data Quality and Total Survey Error Models

This paper suggests consideration of total survey error models within a multi-dimensional framework of risk management. For the current discussion, define *survey risk* in terms of the likelihood and impact of degradation in one or more components of survey data quality. Per Brackstone (1999, *Survey Methodology*), one may define these components of quality in relatively broad terms: accuracy, timeliness, relevance, interpretability, accessibility and coherence. In addition, within the component of "accuracy" one may consider all of the dimensions considered in total survey error (TSE) models, e.g., frame errors, sampling error, incomplete data effects, measurement error and processing effects. For any of these TSE dimensions, it can be useful to distinguish between *methodological errors* which would occur even if a given procedure were carried out exactly as specified, and *operational errors*, which occur when practical implementation deviates from the procedure specifications.

2. Aggregate and Systemic Components of Risk

2.1. Definitions and Illustrative Examples

Historically, total survey error models have focused on components of error associated with problems that arise at a relatively fine level, e.g., at the interviewer or respondent level. The impact of these errors on estimator performance can then be characterized statistically as the aggregate effects of a relatively large number of (approximately) independent random processes related to reporting errors, missing-data mechanisms or other components of a total survey error model. Thus, one may describe these as *aggregate* components of risk.

In many practical cases, however, some important sources of missing-data or measurement-error phenomena are *systemic* in nature, in the sense that a single realization of a random process can lead to serious degradation of data quality. Some relatively simple examples are as follows.

- Changes in contractual, legal or regulatory structures that have previously led to administrative records that are used for frames, weight construction or imputation
- Definitional or data-management problems associated with construction of frames from administrative records
- Errors in programming skip patterns or other features of a data collection instrument
- Errors arising in the training or management of data collection personnel
- Decisions occurring in the implementation of rules for survey data editing and imputation
- Lack of fit in implicit or explicit models used for imputation, weighting adjustments or measurement error adjustments

- Failure to identify important sources of nonsampling error in preliminary lab studies, pilot tests or field operations
- Problems in implementation of rules for edit, imputation or allocation work with previously collected data

In keeping with the definitions provided in Section 1, note that many, but not all, of these systemic errors are operational in nature.

Historically, the methodological research community has focused primary attention on the characterization, modeling and management of aggregate risk. However, anecdotal evidence from survey program managers and field operations personnel suggest that in many cases systemic risk components may also be important. To the extent that this may hold for a given survey, it is useful to explore the extent to which methodologists may be able to contribute to the management of systemic risks.

2.2 Characterization of Systemic Risks through Standard Models

Depending on the specific application, one could consider several distinct models that would account for systemic error components. For example, for a unit *j* in a coarse group i, define the mixed linear model,

$$Y_{ij} = X_{ij}\beta + Z_{ij}\gamma + u_i + e_{ij} \qquad (1)$$

where Y_{ij} is the observed outcome or survey variable, X_{ij} is a vector that defines a set of design conditions, Z_{ij} is a vector of observed paradata, u_i represents a regional or other coarse group effect, and e_{ij} represents a unit-specific error. One could then represent some systemic effects through changes in the coefficients β or γ, or through the coarse random effects u_i. For work carried out in the context of model (1), or generalized mixed linear model versions thereof, it is of special interest to identify cases in which systemic errors have an effect on the *informativeness* of the associated TSE component. For example, changes in the coefficients, or changes in the conditional moments of u_i, may affect the informativeness of the associated nonresponse indicators Y_{ij}.

In other cases, hierarchical Bayes models may be preferred. In addition, for risks associated with inadequate testing of data collection software, life-testing models may be useful. All of these models potentially involve important issues with collection of sufficient model identification information, and with limited effective sample size. In some cases, one may address these issues in part through elicitation of informative priors from experienced personnel.

Finally, some cases of systemic risk may be approximated by “normal accident” models (e.g., Perrow (1984) and others) arising from “complex and tightly coupled systems.” These models may be of special interest in one or both of the following cases.

(a) Surveys that use highly standardized procedures, and related production systems, that may not include implicit feedback loops and model checks that are effectively incorporated into surveys with a higher degree of direct review and intervention by survey personnel.

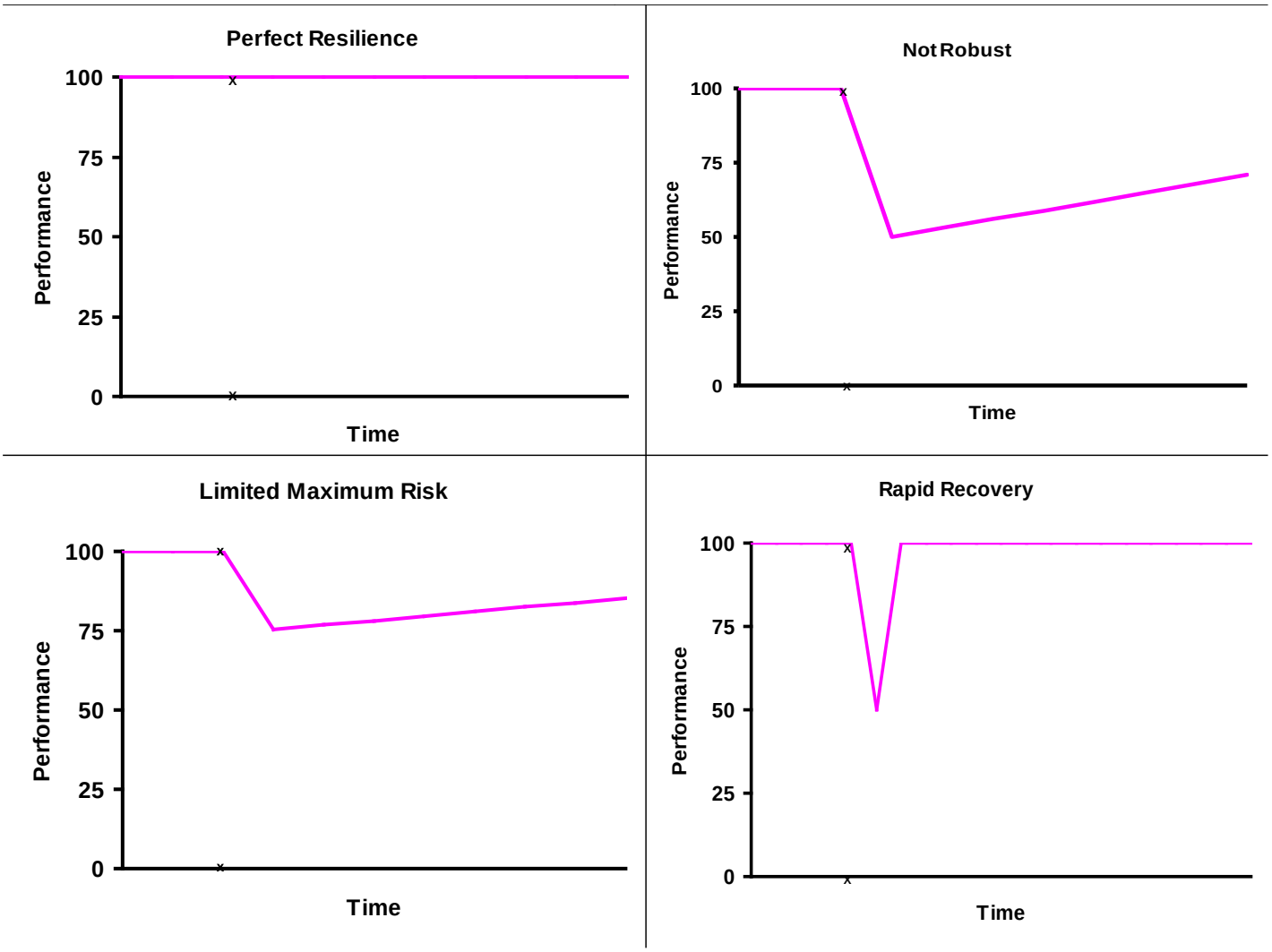
(b) Surveys that have experienced substantial resource reductions, which potentially lead to a corresponding reduction in feedback loops and buffers that had previously served to ameliorate the risks of “tight coupling” identified by Perrow (1984).

3. Impact of, and Recovery from, Systemic Errors

In practical work with systemic errors, it is important to consider population or design conditions that may affect the impact of, and recovery from, such errors when they occur. The four graphs on the next page provide simple schematic descriptions of possible outcomes. In each case, the vertical axis represents a measure or quality or efficiency, ranging from 0 (worst case) to 1 (ideal performance). The horizontal axis represents time, and the "x" mark on the horizontal axis represents the moment at which a systemic error occurs. The graph in the upper left corner represents a case of perfect resilience, in which the systemic error has no effect on estimator performance. The graph in the upper right provides a case with substantial degradation in quality, followed by a slow and partial recovery. The graph in the lower left illustrates a case with moderate degradation, again followed by slow recovery. Finally, the graph in the lower right provides a case of substantial degradation, followed by rapid recovery. This final case is somewhat analogous to cases studied in the literature on "recoverable computing" in information technology.

4. Adjusting Design Features to Account for Systemic Error Components

If we identify some prospective systemic error components in a survey, we can consider adjustments focused on one or more of three goals. First, we can try to prevent systemic errors from arising in the first place. Second, we can try to design the survey procedure to be relatively robust against systemic errors, in keeping with the first or final graphs in Section 3. Third, we can design the system to allow for timely identification of, and adjustment for, systemic errors, along the lines of classical multi-phase or responsive designs.



5. Closing Remarks

This paper has suggested that in the exploration of "survey risk" (defined in terms of the likelihood and impact of degradation in one or more dimensions of survey quality), one should consider both *aggregate* and *systemic* components of risk. We would welcome comments from the workshop participants on any aspect of the paper, and are especially interested in responses to the following.

1. Have you (or your survey organization) encountered systemic components of TSE models, or systemic components of other dimensions of survey risk?
2. If so:
   a. What are their dominant features?
   b. What models have you used to describe these systemic components, and their overall effect on data quality?

What changes in design or estimation methods have you used (or considered) to ameliorate the effects of these systemic risk factors?

Paradata Collection Research for Social Surveys at Statistics Canada

International Total Survey Error Workshop (ITSEW) Quebec, June 2011

François Laflamme, Statistics Canada

Introduction

The challenge of any statistical organization is to collect high quality data in a cost effective manner despite many influencing factors such as decreasing response rates, evolving population behaviour, increasing respondent burden, etc. Data collection is definitively a key element of the survey process because it has a direct impact on quality and it is a major component of the cost of many statistical products. Since 2006, Statistics Canada has conducted a series of paradata research to understand, assess, monitor and improve the data collection process and practices. The goals of these studies included one or more of the following six objectives:

- to learn more about the data collection survey process within and across surveys;
- to identify operational efficiency opportunities;
- to evaluate the data collection process including new initiatives ;
- to provide timely feedback and customized information for active survey management;
- to maintain and improve data quality; and ultimately
- to improve the way data collection is conducted and managed.

In order to achieve these objectives, the paradata research has been taken advantage of the Statistics Canada data collection infrastructure in which the paradata are constantly collected and stored. In particular, these researches are based on empirical paradata automatically and timely collected throughout the data collection period for CATI and CAPI social surveys. These type of researches can

- be used for operational research (i.e. essentially before and during data collection) and methodological research (i.e. mainly after data collection);
- use numerous paradata sources (e.g., transaction and contact information about each call or visit, interviewer payroll information, audit trail (key stokes), etc.);
- be often performed in conjunction with other data sources (e.g. sample design, budget information, targeted response rates, etc.);
- provide the opportunity to conduct trend analysis over time;
- assess the impact of new initiatives; and
- be generally repeated across different type of surveys to compare the results as well as to validate and generalize research conclusions.

The main objective of this paper is to provide an overview of the paradata research at Statistics Canada including some major data collection changes already implemented to take advantage of the research findings (e.g. Responsive Collection Design). In addition, future research plans and priorities that focus on the identification of viable operational strategies to improve data collection efficiency or data quality are discussed.

Paradata research

The initial paradata research essentially focussed on the descriptive analysis of the collection process and practices. In particular, this research included these basic types of investigations: contact and response rates, average number of calls to get a contact or an interview, call attempts versus time spent (how data collection time and effort is distributed throughout the collection process), best time to call, call scheduler features, sequence of calls analysis, relationship between production and cost, productivity indicators, etc. Many ad hoc studies were also conducted at that time to investigate special and emerging issues or to validate (or invalidate) anecdotal thoughts and perception about data collection process. All these investigations constituted the first but necessary step of the paradata research because they have provided a very good understanding of how data collection progresses through for both CATI and CAPI surveys.

During these paradata studies, many potential opportunities for improvements were identified. For example, historically the focus of many researches have been on the reduction of the number of calls to get a first contact but this is not where most of the data collection time and effort for both respondents and non-respondents was spent. It would be more valuable in future research to pay more attention after the first contact when interviewers are trying to get cooperation and an interview or to confirm a non-response. Other research findings have indicated that the same data collection approach does not work effectively throughout an entire data collection cycle, stressing the need to develop a more flexible and efficient data collection strategy. The collection approach should evolve through the collection period to make better use of the information available prior to the start of collection and to take advantage of the information that becomes available during data collection to adapt the data collection strategy. The research carried out also suggested that collection resources are currently not always optimally allocated with respect to the assigned workload and the corresponding expected productivity for CATI surveys.

Among the improvements identified during these studies, some have been implemented to benefit from these findings. For example, the proportion of evening shifts versus day shifts has gradually increased throughout the collection period to improve contact rate and survey productivity. As well, time slices were customized for some type of survey in order to use information available prior to and during collection, for example sample design and household socio-economic information collected from the roster. In addition, the impact of new data collection initiatives (e.g. cap on calls and time slices) was assessed and continues to be monitored, resulting in changes in the cap on calls definition for some major surveys. More recently, paradata research fully materialized with the successful implementation of a Responsive Collection Design (RCD) strategy for CATI surveys to improve the way data collection is conducted and managed. RCD is an adaptive approach to survey data collection that uses information available prior to and during data collection (e.g., sequence of calls) to adjust collection strategies to make the most efficient use of remaining available resources. As part of a RCD initiative, active management tools were improved to constantly assess the data collection process, to provide timely feedback to survey managers and to determine data collection milestones where changes to collection strategy are required.

Finally, in other to share and communicate more efficiently paradata research results and experiences, Statistics Canada has recently developed a paradata course. In conjunction with papers and presentations resulting from this research, this course aims at providing an overview of the use and practical applications of paradata to plan, manage, monitor, improve and assess the survey process.

Ongoing and future paradata research

Among the other opportunities for improvements identified, some require further investigations to better assess their feasibility and operational advantages and benefits for the data collection process. In particular, the priority of future research will be given to opportunities that could be operationally viable and lead to cost-efficiency, timeliness or quality improvements. For example, a more detailed analysis of the sequence of calls to after a first contact as well as the time spent to achieve cooperation after a first contact is required to maximize the likelihood of making fewer calls to get cooperation. Another example would be a study on the relationship between key indicators to identify more optimally and objectively when to start RCD phases. There is also a need to investigate how interviewer scheduling could be better planned and managed throughout the data collection process. The objective is to improve the distribution workload of CATI interviewers within and between surveys.

For CAPI surveys, initial research has been conducted to assess the quality of the paradata information and its limitations, and to better understand the data collection process and practices. Again, some of the potential improvement opportunities are currently investigated e.g. potential benefits of RCD to improve cost-efficiency of CAPI surveys.

Finally, one of the most important future challenges of paradata research would be to phase in, integrate and consolidate this series of opportunities by taking into account the changing data collection environment (e.g. responsive design, multi-mode surveys), evolving technologies (e.g. cell phones, internet), population behavior (e.g. unwillingness to participate to surveys) in order to maintain and/or improve the cost efficiency and quality of the data collection process in the long run.

Potential issues to be discussed

- Are there important gaps in paradata research? if so
    - Which type of research need to be done?
    - What are the research priorities?
- Sharing information (communication)
    - Paradata working group, Conferences/events (paradata sessions in many international events), International network… Is it enough?- Is it efficient?
    - Potential collaboration between organizations – can it be improved?
    - Potential collaboration between survey methodologist and data collection manager?
- What is the most efficient organizational structure for this type research?
    - To met both operational and methodological objectives

References - list of some Statistic Canada paradata research papers

An L., Laflamme G. (2010), “Coordinated Collection and the Quality of Paradata for CAPI Surveys at Statistics Canada”, paper presented at the Joint Statistical Meeting, Vancouver, Canada

Beaumont, J.F. (2005), “On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment,” Survey Methodology, 31, pp. 227-231.

Hunter, L. and Carbonneau, J.-F. (2005), “An Active Management Approach to Survey Collection”. Proceedings from the 2005 Statistics Canada International Symposium on Methodological Issues.

Laflamme, F., Maydan, M. and Miller, A. (2008), “Using Paradata to Actively Manage Data Collection”. Proceedings from the 2008 Joint Statistical Meeting.

Laflamme, F., (2008), “Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada”. American Association for Public Opinion Research 63rd Annual Conference, 2008 American Statistical Association, Proceedings from the Section on Survey Research Methods.

Laflamme, F., (2008), “Data Collection Research using Paradata at Statistics Canada”. Proceedings from the 2008 Statistics Canada International Symposium on Methodological Issues.

Laflamme, F., (2009), “Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada”. Proceedings from the 57th International Statistical Institute Conference.

Laflamme, F. and Karaganis, M. (2010), “Assessing Quality of Paradata to Better Understand the Data Collection Process for CAPI Social Surveys”, Proceedings from the European Quality Conference, Helsinki, Finland

Laflamme, F. and Karaganis, M. (2010), “Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada”, Proceedings from the European Quality Conference, Helsinki, Finland.

Lapierre, B. and S. Meyer. 2006. “Using the Audit Trail data to evaluate the quality of collection of the Canadian National Longitudinal Survey of Children and Youth.” Proceedings of the Social Statistics Section. American Statistical Association.

Mohl , C. and Laflamme, F. (2007), “Research and Responsive Design Options for Survey Data Collection at Statistics Canada”. Proceedings from the 2007 Joint Statistical Meeting.

Tabuchi, T, Laflamme, F., Phillips, O., Karaganis, M. and Villeneuve, A. (2010), “Responsive Design for the Survey of Labour and Income Dynamics”. Proceedings from the 2010 Statistics Canada International Symposium on Methodological Issues.

To come this year

Choudhry, G.H, Hidiroglou, M.A and Laflamme, F. (2011), “Optimizing CATI Call Scheduling to Minimize Data Collection Costs”, to be published in the 2011 Proceedings of the Joint Statistical Meeting.

Laflamme, F. and St-Jean, H. (2011), “Highlights and Lessons from the First Two Pilots of Responsive Collection Design for CATI Surveys”, to be published in the 2011 Proceedings of the Joint Statistical Meeting.

Laflamme, F. and St-Jean, H. (2011), “Proposed Indicators to Assess Interviewer Performance in CATI Surveys”, to be published in the 2011 Proceedings of the Joint Statistical Meeting.

Laflamme, F. (2011), “Using Paradata to Manage Responsive Collection Design for CATI surveys”, to be published.

Propensity score models for nonresponse and measurement error.

John Dixon

Propensity scores are proving very useful in studies of nonresponse error (Schouten and Leufkens, 2010 ITSEW). The current study uses nonresponse propensity models (Dixon, 2010 Nonresponse workshop) and develops propensity models for different sources of measurement error.

Many indicators of nonresponse from paradata could be reduced to a more manageable set of reasons for noncontact and refusal. Dixon (2010 Nonresponse workshop) found 2 factors related to noncontact. The first factor was related to timing concerns, the second to barriers or ability to contact the household. The factors were consistent between 3 surveys; National Health Interview Survey, Current Population Survey, and the Consumer Expenditure Quarterly Survey. He found 4 factors related to refusal, similar to Maitland et. al. (2008 Joint Statistical Meetings). The first factor was related to hostile response to the survey request, the second to timing of the survey, the third to the busyness of the potential respondent, and the last to privacy concerns.

Indicators of measurement error may similarly be able to be reduced to a more manageable set. Proxy reporting, use of records, item or section timing data, and logical edits can be used to try to separate different patterns of potential measurement error. The combined model would give an opportunity to study the common relationships between nonresponse and measurement error as well as the unique contributions to total survey error. A factor analysis of process and edit variables produced 3 factors; one related to economic variable edits, the second to personal variable edits (probably due to proxy reporting, e.g.: age), the third to process variables (e.g.: telephone interview). An additional variable was used, “the use of records”, which is thought to relate to data quality. Other studies have found little relationship between measurement error and nonresponse error. The effects are small in this study.

Table 1. Noncontact and Refusal correlated with Measurement error indicators.

Variable	fecon	fpers	fprocess	nouserec
Noncontact	0.08294	-0.00909	0.05472	0.11020
Ntiming	0.09642	0.00656	0.06171	0.03250
Nbarrier	-0.00364	-0.04230	-0.00216	0.16663
Refusal	0.21429	0.05939	0.16930	-0.04461
rhostile	0.19808	0.03860	0.13678	-0.06559
rtime	0.13350	0.06655	0.12342	-0.09529
rbusy	0.18245	0.04086	0.13556	0.15390
rprivacy	0.10792	0.04955	0.08554	-0.02740

In Table 1 the propensity scores for noncontact were related to not using records. This appeared to be due to barrier issues. Refusal propensity was related to economic edits and survey processes. The economic relationship was related to hostile refusal propensity as well

as busyness propensity. Process variables were related to all forms of refusal but less with privacy. Busyness was also related to not using records.

The relationship between the propensity scores and measurement error indices and measures from the survey give an indication of bias (Table 2). Employment had a negative relationship with barrier noncontact propensity, a positive relationship with hostile refusal propensity, and a negative relationship with busyness refusal propensity. It also had a negative relationship with not using records (i.e.: employed used records in the interview more). Earned income had a similar pattern. Total expenditures had a similar patter too, but also had a relationship with edit indicators for economic and personal variables.

Table2. Correlation of Noncontact, Refusal, and Measurement Error Indicators with Measures.

Variable	Ceemp	ernincome	totex
Noncontact	-0.04240	0.00055	-0.07002
Ntiming	0.05237	0.03150	-0.03976
Nbarrier	-0.16814	-0.07476	-0.10836
Refusal	0.11895	0.02281	0.04875
Rhostile	0.18225	0.10666	0.10781
rtime	0.10247	-0.08299	-0.03929
rbusy	-0.14175	-0.09294	-0.08660
rprivacy	0.05663	0.01477	0.04610
fecon	0.04696	0.06151	0.10524
fpers	0.03225	0.06224	0.08880
fprocess	-0.01586	-0.03572	-0.01208
nouserec	-0.33462	-0.12235	-0.17956

The nonresponse propensity scores had reversed signs which reduced the bias overall. The three measurement indicators were only related to total expenditures, but record use was the strongest indicator of bias. The edit indicators were often in opposite directions to the record use indicator, suggesting they are measuring some other type of error.

I want to expand the list of process variables in the measurement indicator models and replicate the study with another survey.

How Much of Interviewer Variance is Really Nonresponse Error Variance? New Results from a National CAPI Survey in Germany

Presentation Date and Time: Monday, June 20, 2011 (10:30 – 12:00)

Author Biographies:

Brady West is a PhD candidate who just finished his third year in the PhD program of the Michigan Program in Survey Methodology (MPSM), and also serves as a Lead Statistician at the Center for Statistical Consultation and Research (CSCAR) on the University of Michigan-Ann Arbor campus. He received an MA in Applied Statistics from the U-M Statistics Department in 2002, being recognized as an Outstanding First-year Applied Masters student, and a BS in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001. His current research interests include the implications of measurement error in auxiliary variables for survey estimation and responsive survey design, survey nonresponse, interviewer variance, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed modeling procedures (Linear Mixed Models: A Practical Guide using Statistical Software, Chapman Hall/CRC Press, 2007), and he is a co-author of a second book entitled Applied Survey Data Analysis (with Steven Heeringa and Pat Berglund), which was published by Chapman Hall in April 2010.

Frauke Kreuter is an Associate Professor in the Joint Program in Survey Methodology (JPSM) at the University of Maryland-College Park. She received a Masters Degree in Sociology from the University of Mannheim, Germany and her PhD in Survey Methodology from the University of Konstanz. Before joining the University of Maryland she held a postdoc at the UCLA Statistics Department. Her research interests include sampling and nonsampling errors in complex surveys, nonresponse, systematic measurement errors in survey response, and growth mixture modeling for non-normal outcomes. In her work at JPSM she maintains strong ties to the Federal Statistical System, and she has served in advisor roles for the National Center for Educational Statistics and the Bureau of Labor Statistics. She is currently on leave from Maryland to head the statistical methods unit at the Institute for Employment Research in Nuremberg, Germany and to teach at the Statistics Department at the University of Munich.

Research Objectives: In a recent paper, West and Olson (2010) decomposed -- for multiple respondent-based survey estimates -- interviewer variance into measurement error variance and nonresponse error variance among interviewers. Their goal was to examine what proportion

of the interviewer variance is really due to interviewers systematically varying in their success to obtain cooperation from respondents with varying characteristics, rather than variance among interviewers in systematic measurement difficulties. Unfortunately, their analysis only considered data from a CATI survey, and thus suffers from two limitations: Interviewer effects are commonly much smaller in CATI surveys, and more importantly sample units are often contacted by several CATI interviewers before a final outcome (response or final refusal) is achieved. The latter introduces difficulties in assigning nonrespondents to interviewers, and thus interviewer variance components are only estimable under strong assumptions. The paper presented here aims to replicate the analysis performed by West and Olson, using data from a national CAPI survey in Germany where CAPI interviewers were responsible for working a fixed subset of cases. Secondary aims of this research included 1) assessing evidence of interpenetrated assignments in this type of CAPI setting, 2) examining added contributions of nonresponse error variance and measurement error variance among interviewers to total interviewer variance if interpenetrated assignment is not evident, and 3) providing additional motivation for analytic development of estimators of intra-interviewer correlations that recognize both the measurement error variance and the nonresponse error variance that can be introduced by interviewers.

Research Methodologies: We analyze a unique survey data set [Waves 1 and 2 of the PASS (Labor Market and Social Security) survey in Germany], where official administrative records are available for all sampled cases on variables that were also measured in the survey. Data were collected using CAPI from a national sample of persons having received unemployment benefits. A total of 158 interviewers trained in CAPI methods and assigned to different sampling points in Germany collected the data. We apply the analysis used by West and Olson (2010) to four key variables collected in the PASS survey, using generalized linear mixed models to estimate 1) variance in the means of true values of subsample assignments among interviewers (allowing for an examination of interpenetrated assignment), 2) variance in the means of true values of *respondents* among interviewers (allowing for an examination of nonresponse error variance), and 3) variance in the means of reported values for respondents (total interviewer variance). Estimation of these components allows for decomposition of total interviewer variance into sampling variance, nonresponse error variance, and measurement error variance. In line with past findings (Schnell and Kreuter, 2005), we found interviewer effects for all four variables; however, the effects on demographic variables (age in particular) seem to be largely a result of nonresponse error variance, whereas more sensitive items show larger effects of measurement error variance.

Issues, Limitations, and Concerns for Discussion at ITSEW 2011:

- This type of analysis assumes that interviewer-level nonresponse errors and measurement errors are independent of each other. Is this a reasonable assumption? Consent requirements prevented us from fully examining this assumption in the PASS survey data set for all respondents.

- Are participants aware of any other CAPI surveys where record values are available for the entire sample, and this analysis can be replicated? Replications are certainly needed in other survey contexts.

- We once again find evidence of nonresponse error variance among interviewers in terms of age, which is consistent with West and Olson (2010). Interviewer training implications would include a need to monitor responding cases for each interviewer in terms of known frame characteristics, and compare those to nonresponding cases. If large differences are arising, interventions may be needed. Should interviewer training focus more on ensuring that respondents are a representative sample of the assigned sample, rather than focusing primarily on standardizing measurement?

- What role can multilevel modeling play in studies of this phenomenon? West and Olson (2010) and Biemer (1980) propose imputing measurement errors for nonrespondents, and this would enable estimation of nonresponse error variance, measurement error variance, *and* the correlation of the two error sources across interviewers in multilevel models. Does this sound like a reasonable idea?

REFERENCES

Biemer, Paul P. (1980). A survey error model which includes edit and imputation error. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 616-621.

Schnell, R. and Kreuter, F. (2005). Separating interviewer and sampling point effects. *Journal of Official Statistics,* 21(3), 380-410.

West, B.T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5), 1004-1026.

Incorporating Nonresponse Propensity in a Markov Latent Class Measurement

Error Model of Consumer Expenditure

Brian Meekins, Bureau of Labor Statistics

Clyde Tucker, Bureau of Labor Statistics

Paul Biemer, RTI International

Abstract

The Consumer Expenditure Interview Survey (CE) is a rotating panel survey conducted quarterly, by the Census Bureau, for the Bureau of Labor Statistics. Respondent households are asked in four consecutive quarters about their expenditure on a number of commodities in the preceding quarter. Recent research by Tucker et al (2003, 2011) was successful in predicting measurement error from item nonresponse using Markov Latent Class models with a number of different commodities. To date, however, these models have not attempted to account for unit nonresponse or show the relative size of this error compared to measurement error. This work attempts improve upon previous models by, at least partially, accounting for error from unit nonresponse through the inclusion of a response propensity "factor" in the models. In addition, the impact of the mode of survey administration on measurement error is assessed. Recent research by Safir and Goldenberg (2008) shows modest negative effects for telephone interviewing compared to in-person interviewing on key data quality measures while controlling for a number of demographics. Preliminary findings by this author (Meekins et al. 2010) finds that the telephone mode - especially when the interview is likely completed by cell phone - has a modest effect on expenditure reporting, length of interview, and use of records (receipts, bills, etc.) compared to the in-person mode. This remains true controlling for some types of panel nonresponse and key demographic variables.

Brian Meekins
Meekins.brian@bls.gov
(202) 691-7459

Clyde Tucker
nctucker@cox.net

Paul Biemer
ppb@rti.org

An Active Collection using Intermediate Estimates to Manage Follow-Up of Non-Response and Measurement Errors

Jeannine Claveau, Serge Godbout and Claude Turmelle

Statistics Canada must deliver relevant, current and highest priority information to Canadians. Not only the information must be ensured with sufficient quality, but it also needs to be produced at the lowest cost possible. Currently, the Unified Enterprise Surveys (UES) program of Statistics Canada consists of 60 annual business surveys which are integrated in terms of content, collection and data processing. For most business surveys, data collection takes approximately ten months. Statistics Canada is undertaking a general restructuring of its business statistics programs. One of its goals is to integrate a much larger array of business surveys than the UES currently does. Another goal is to let electronic data collection become the principal mode of collection for business surveys. Currently, many business surveys continue to use mail questionnaires for initial data gathering. Telephone follow-up is conducted to resolve edit problems with mailed-back questionnaires and to collect data from units who have not returned the questionnaires after a pre-specified period. Under the new approach with multi-mode collection, telephone follow-up would continue to be necessary. However, management of the collection process of the business surveys needs to be enhanced in order to make the process more efficient.

Statistics Canada is developing an Active Collection methodology to manage follow-up of non-response and measurement errors (Godbout, Beaucage and Turmelle, 2011). The objectives are to reduce non-response, lower operating costs, enhance quality assurance and improve responsiveness in order to save resources without compromising quality. To do so, intermediate estimates would be produced periodically throughout collection in order to monitor quality evolution and have an early picture of the estimates. Once an acceptable level of quality is met, estimates are deemed final and collection resources are re-allocated. If all targets are met for one specific survey, then collection is stopped for that survey. The Active Collection would use quality indicators of the intermediate estimates to allocate and prioritize follow-up activities and to determine when the collection period ends. Non-response follow-up operations will include fax and email reminders and telephone follow-ups; since the first two actions are assumed to be much cheaper than the third one, this presentation focuses on telephone follow-ups.

The proposed strategy will be based on a dynamic adaptive design (Schouten, Calinescu and Luiten, 2011; Groves and Heeringa, 2006). The basic collection strategy is the following: at the beginning, all sampled units will receive a survey questionnaire and regularly throughout collection period, non-respondents will get fax and email reminders. On the other hand, the collection strategy will be upgraded to telephone follow-ups for a subset of significant non-respondents after each intermediate estimates produced, until the quality levels are met. This subset, representative of the non-respondents, will be randomly selected using global Measure of Impact (MI) scores. Each unit will have a positive selection probability but the most significant ones in terms of MI scores above a given threshold (called influential units) will be selected with certainty for follow-up. The thresholds and the number of units to be selected will be allocated according to quality targets and collection capacity.

The MI scores will associate a score to each unit in order to give priority for collection and editing activities. The MI score of a unit for a given estimated parameter is defined as the standardized difference between the actual estimated parameter and its predicted value when the unit goes from its observed (or unobserved) values to some predicted values:

$$MI_k(\hat{\theta}) = (\tilde{\theta}_k - \hat{\theta})/\varphi$$

where $\tilde{\theta}_k$ is the estimated parameter after changing reported values and/or covariates of unit *k* respectively to $\tilde{y}_k$ and/or $\tilde{x}_k$ and φ is a scaling factor.

The set of estimated parameters includes estimated totals and quality indicators (QI). The quality indicators can be covariate-based (independent of surveys variables) such as weighted response rates and response representativeness (R-indicator) or item-based such as variance (or CV) and imputation rate of a variable of interest. The R-indicator can be used to measure and control the potential bias and it is based on response propensities.

The Active Collection methodology will include a large number of variables of interest to monitor. Monitoring all of them will be a challenge. Since not all of them are equally important, a limited number of key variables will be identified and monitored using item-based QIs and MI scores. The quality of the non-key variables will be controlled using covariate-based QIs. To estimate the response propensities that are required to compute most covariate-based QIs, auxiliary data and paradata will be used.

In order to prioritize units for telephone follow-up, one global score per unit is needed. The MI scores for each estimated parameter and quality indicator are considered local scores. The sum, the maximum and the Euclidian distance of the local scores as proposed by Hedlin (2008) could be used.

Validation of the proposed strategy will be done in the future by conducting simulations and developing prototypes.

Discussion

1) What quality indicators are appropriate to measure the risks of potential bias in the estimates?

2) What is the best way to use quality indicators (e.g. R-indicator) to monitor collection of highly skewed business surveys?

3) What are the pros and cons of using item-based or covariate-based quality indicators to monitor quality evolution?

4) The proposed approach obviously affects the response propensities throughout collection. Although we can adjust the estimator later on to take this into account, is it something we should move away from? Or should we take advantage of it?

5) In the proposed approach, are there any additional aspects that should be considered?

REFERENCES

Godbout, S., Beaucage Y., Turmelle C. (2011). Quality and Efficiency Using a Top-Down Approach in the Canadian's Integrated Business Statistics. Paper presented at the UNECE Work Session on Statistical Data Editing in Ljubljana, Slovenia, May 2011.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design For Household Surveys: Tools for Actively Controlling Survey Errors and Costs. J. R. Statist. Soc. A (2006), 169, Part 3, pp. 439–457.

Hedlin, D. (2008). Local and Global Score Functions in Selective Editing. Conference of European Statisticians, Work Session on Statistical Data Editing (Vienna, Austria, 21-23 April 2008).

Schouten, B., Calinescu, M. and Luiten, A. (2011). Optimizing Quality of Response Through Adaptive Survey Designs. To be published.

Responsive Collection Design for CATI Surveys and Total Survey Error (TSE)

International Total Survey Error Workshop (ITSEW) Quebec, June 2011

François Laflamme, Statistics Canada

1. Introduction

Responsive Collection Design (RCD) is an adaptive approach to survey data collection that uses information available prior to and during data collection to adjust the collection strategy for the remaining cases. The RCD objectives are to monitor and analyse collection progress against a pre-determined set of indicators to identify critical data collection milestones that require significant changes to collection approach and to adjust collection strategies to make the most efficient use of remaining available resources. In RCD context, control of the data collection process is not determined solely by a desire to maximize the response rate or reduce costs. Numerous other considerations come into play when determining which aspects of data collection to adjust and how to adjust them. These include quality, productivity, the response propensity of in-progress cases, the mode of collection and competition from other surveys for resources. Statistics Canada has been developing an RCD strategy and has tested it on two CATI surveys, namely, the 2009 Households and the Environment Survey (HES) and the 2010 Survey of Labour and Income Dynamics (SLID), which is a longitudinal survey.

This paper provides an overview of the RCD and discusses potential sources and types of error with regards to the RCD strategy used at Statistics Canada[1]. For more detailed information, please refer to Laflamme and Karaganis (2010), especially for the HES experience. The paper by Tabuchi et al. (2009) discussed the design for SLID but does not include results as it was written prior to the implementation of RCD[2].

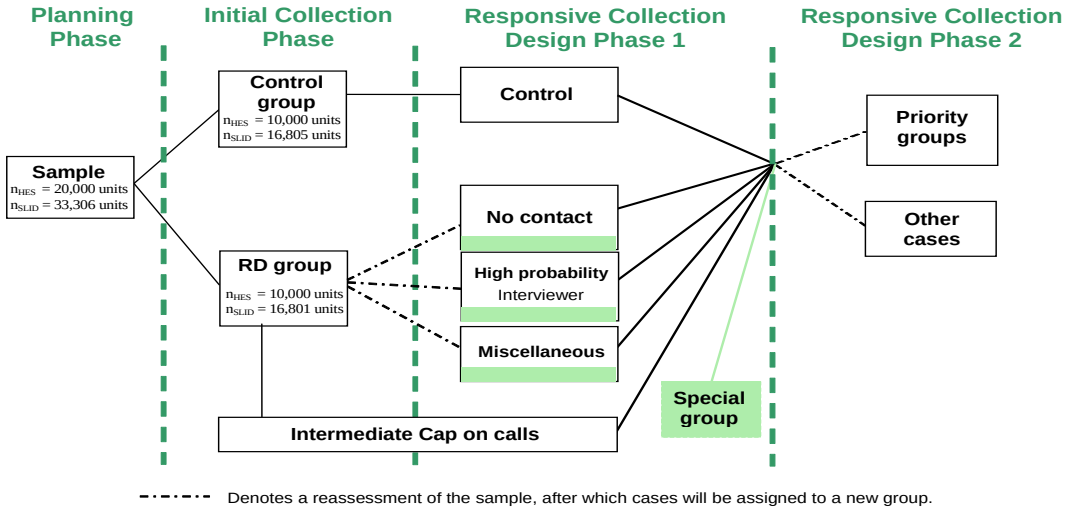
2. The Responsive Collection Design strategy

Figure 1 presents an overview of the RCD strategy for the HES and SLID. The RCD strategy used for SLID was slightly modified to take advantage of the lessons learned from HES and to respond to its specific needs.

[1] The presentation will also briefly describe the plans, tools and approach used to actively manage data collection and presents the highlights obtained along with lessons learned for the RCD surveys.

[2] Two other papers on RCD are planned in 2011.

Figure 1: RCD strategy for HES and SLID



Notes:

1. For SLID 2010, another group called “High probability - Tracing” was used during RCD phase 1.
2. The “Special group” contains cases with that need particular attention at the end of RCD phase 1.

The first phase (planning) occurs before data collection starts. During the planning phase, data collection activities and strategies are planned out, developed and tested for the other three data collection phases including the development of the response propensity model(s). The second phase (initial collection) includes the first portion of the data collection process, from the collection start date up until it is determined that RCD phase 1 needs to be initiated. An intermediate cap on calls was also introduced to avoid cases capping out before the last data collection phases. During this initial collection phase, many key indicators of the quality, productivity, cost and responding potential of in-progress cases are closely monitored to identify when the next RCD phase should be initiated. The third phase (RCD phase 1) categorizes and prioritizes in-progress cases using information available prior to the beginning of collection and paradata information accumulated during collection with the objective of improving overall response rates. During this phase, key indicators continue to be monitored. In particular, the sample representativity indicator (R-indicator)[3] provides information on the variability of response rates between domains of interest to determine when the last phase should begin. The last phase (RCD phase 2) aims at reducing the variance of response rates between the domains of interest (improving sample representativity) by targeting cases that belong to the domains with lower response rates.

[3] The R-indicator concept was first discussed by Schouten, Cobben and Bethlehem (2009).

Both HES and SLID samples were randomly divided into two equal groups based on the sample design information, the control group and responsive design group, to assess the impact of the RCD strategy[4]. The control group followed the usual collection process, and the responsive design (RD) group followed the new strategy. The two groups are combined again at the end when overall representativity of the sample is seeked.

3. Discussion about TSE in the RCD context - potential issues

- According to the current RCD strategy, it is expected to reduce the variance of the non-response adjustment factors between domains of interest. What about potential non-response bias?
- A propensity logistic model was used to evaluate a household's likelihood of being interviewed during collection and to categorize and prioritize each in-progress case. The HES and SLID response propensity model(s) were developed by regional office using three sources of information for the two pilots: sample design information, paradata from previous and current data collection cycles to identify the explanatory variables for inclusion in each model. During the RCD, these variables remained the same while the parameters of the model are re-evaluated daily using the most recent cumulated paradata at the end of each collection day.
    - Is the current RCD strategy likely to modify the response propensity of individuals sample units (grouping cases into more homogeneous groups, no incentives used, no-sub sampling of non-respondents)?
    - Propensity model is also another source of potential error. Is a special weighting methodology appropriate? If so, how should it be accounted for?
- Are there other strategic opportunities to improve the adopted RCD strategy with respect to TSE?

References

Groves, R. M. and Heeringa, S. G. (2006), Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, Journal of the Royal Statistical Society, Series A, 169, 439-457.

Laflamme, F. and Karaganis, M. (2010), Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada, presented at the European Conference on Quality in Official Statistics (Q2010), Helsinki, Finland.

Schouten, B., Cobben, F. and Bethlehem, J. (2009), "Indicators for the representativeness of survey response", Survey Methodology, 35, pp. 101-114.

[4] More recently, SLID 2011 used a full RCD approach and test a new initiative on a portion of the 2010 respondents. For those, the first call was made in the same time slice where the SLID 2010 interview was conducted.

Tabuchi, T., Laflamme, F., Phillips, O., Karaganis, M. and Villeneuve, A. (2009), Responsive Design for the Survey of Labour and Income Dynamics, Proceedings of Statistics Canada Symposium 2009, Longitudinal Surveys: from Design to Analysis.

OPTIMIZING CATI CALL SCHEDULING

Choudhry, G.H.[1], Hidiroglou, M.A., Laflamme, F.
[1]Statistical Research and Innovation Division, Statistics Canada,

16th Floor, R. H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario, K1A0T6

1. Introduction

Statistics Canada is facing increasing challenges in maintaining cost-effective data collection and obtaining high-quality outputs to meet the evolving demands for timely and accurate data from a wide range of users. Since 2006, Statistics Canada has studied paradata to evaluate its current data collection process and practices. The studies carried out so far have identified a number of options to improve the way the agency conducts and manages its surveys with respect to CATI surveys.

Some of these studies were carried out to obtain a better understanding of the relationship between interviewing efforts and the expected workload progress during the data collection cycle. These investigations suggested that the interviewer staffing level were not always well aligned with the workload sample and the expected productivity. For example, given that in-progress units are likely to be called more often during the second half of the collection period within a given day, suggests that interviewer staffing levels are greater than the sample workload in the first-half of the collection period. It has also been observed for CATI surveys that the proportion of completed questionnaires decreases rapidly over time given a fixed number of calls (Laflamme 2009). Data collection managers need to improve interviewer staffing management and planning tools to reduce some of the tension between collection productivity and costs (Couper et al., 1998), and maintain high level of data quality.

Operational constraints involving the interviewing staff have also increased collection costs and limit the capacity to optimize the interviewers' schedules. For example, rules concerning notice of shift changes for a unionized interviewing workforce need to be factored into any action plans. In addition, the overall Regional Office capacity by time slice (i.e. day and evening shifts) also needs to be considered.

The methodology and results presented in this paper only represents the first phase in the development of an optimized interviewer scheduling tool for a single CATI survey. It does not account for interviewer operational constraints such as their availability for each day of the

week, sick leave and vacation, assigned hours per quarter as well as their work shift preferences. Furthermore, the interviewer workload is not optimized over several concurrent surveys. The following problem is addresses. Given that the only constraint is that targeted response rates needs to be achieved, what is the optimal mix on the CATI interviewers by time slice (morning, afternoon, early and late evening shifts).

2. Methodology

2.1 Data Collection

Data collection for CATI surveys is conducted from six call centres managed by Regional Offices (ROs). CATI collection procedures for a given survey can vary by site depending on the mix of concurrent and large scale surveys in collection, workload and availability of interviewers. However, there is paradata standardization across the regional offices because CATI survey data are collected using Blaise. This software automatically collects paradata, and stores it in the Blaise Transaction History (BTH) file. A BTH record is automatically created each time contact with a sampled unit is closed, and this takes place whether the BTH record was opened for data collection or other purposes. The BTH record contains detailed data about each call made to contact a sampled unit during the data collection period. This includes the survey and unit identification, the date, the time the case was open, the identification of the interviewer who worked on it, the results of the call,as well as additional relevant information.

2.2 Problem definition

The data collection period for a given survey takes place over a number of continuous days, say $d = 1,\ldots,D$. Each day is split into time slices say *t=1, T*. In this paper, we assume that these time slices correspond to interviewer shifts. In other words, an interviewer shift will consist of one or more time slices. An interviewer shift represents the number of hours that an interviewer is scheduled to work within a given day. These time slices are fixed periods within a day during which CATI interviewers call the selected units. There are a total of $s = 1,\ldots,S$, with $S = DT$, time slices over the whole data collection period. Calls have two outcomes: a call results in a completed questionnaire or it does not. The observed probability of completing a questionnaire for a given time slice is the proportion of calls resulting in completed questionnaires. We also assume that a sampled unit (telephone) is called only once during any given time slice.

We used paradata from the 2010 cycle of the Survey of Labour Income and Dynamics (SLID) to optimize the probability that a call made during a time slice would result in a completed questionnaire. SLID is an annual longitudinal survey of about 34,000 sample households that uses CATI for collecting data. CATI data collection took place over twenty-eight continuous days for SLID. Each a day was divided into four time slices: 7:00 - 11:00, 11:00 - 15:00, 15:00 - 19:00, and

19:00 - 23:00, that is *T*=4. Thus, there were *S*=112 time slices over the twenty eight days of data collection period.

We observed from the paradata that the probability of completing a questionnaire deceases over time. Therefore, we developed two models: one using cumulative number of calls made up to and including the time slice and the other using the cumulative cost (time spent in minutes) up to and including the time slice. Moreover, there are time slices within the day when the probability of completing a questionnaire is higher. Therefore, we also included dummy variables to indicate period within the day in both the models.

For each time slice we defined dummy variables $z_{ts}, t = 1, 2, 3$, where z_{ts} =1 if *t=s mod 4* and 0 otherwise. It should be noted that only 3 dummy variables need to be defined because the 4th time slice (the last time slice) in each day will be the reference time slice. The optimization of the total number of calls to be made within a time slice was carried in two steps. In the first step, we predicted the probability p_s of completing a questionnaire within a time slice *s* $(s = 1, 2, \ldots, S)$ using the estimated regression model from either of the two models. In the second step, the total predicted cost, based on the number of productive calls (resulting in a completed questionnaire) and the number of non-productive calls (not resulting in a completed questionnaire), was minimized subject to the following constraints: *i.* The number of calls within each time slice was non-negative, and *ii.* The expected overall response rate was equal to some pre-specified response rate. This is an iterative procedure because the objective function (total cost) depends on the number of calls and the probability of productive call by time slice, whereas the probability of productive call in a time slice is a function of cumulative number of calls made up to and including the particular time slice.

2.3 Models for Predicting the Probability of Productive Call

2.3.1 Cumulative Number of Calls as predictor

The linear regression model is given by:

$$p_s = \alpha + \sum_{j=1}^{3} \beta_j \, z_{js} + \gamma \, \bar{C}_s + \varepsilon_s \qquad (2.1)$$

where z_{ts} $(t=1,2,3;\ s=1,\ldots,S)$ are the dummy variables defined above; $\overline{C}_s = \sum_{j=1}^{s} c_j / n$ is the average number of cumulative calls per sampled unit up to and including time slice *s*, where c_j is the total number of calls made during time slice *j* and *n* is the total number of sampled units for the regional office being analyzed. Note that the total number of cumulative calls up to and including time slice *s* per sampled unit is $C_s = n\ \overline{C}_s$.

In the case of linear regression model (3.1), the associated predicted probability for time slice *s* is given by:

$$\hat{p}_s = \hat{\alpha} + \sum_{j=1}^{3} \hat{\beta}_j\ z_{js} + \hat{\gamma}\overline{C}_s \tag{2.2}$$

Since p_s is a proportion between 0 and 1, we also used the corresponding logistic model,

$$ln\left(\frac{p_s}{1-p_s}\right) = \alpha^* + \sum_{j=1}^{3} \beta_j^*\ z_{js} + \gamma^*\ \overline{C}_s \tag{2.3}$$

The corresponding predicted probability for time slice *s* is given as:

$$\hat{p}_s^* = \frac{\exp\left(\hat{\alpha}^* + \sum_{j=1}^{3} \hat{\beta}_j^* z_{js} + \hat{\gamma}^* \overline{C}_s\right)}{1+\exp\left(\hat{\alpha} + \sum_{j=1}^{3} \hat{\beta}_j^*\ z_{js} + \hat{\gamma}^* \overline{C}_s\right)} \tag{2.4}$$

2.3.2 Cumulative Time Spent as predictor

The second model uses the observed cumulative average per unit cost (time spent) up to and including the time slice s. The auxiliary variable $\overline{X}_s$ was computed as $\sum_{j=1}^{s} x_j / n$ where x_j is the observed cost (time spent in minutes) in making calls for a given time slice *j*. The linear regression model given by (3.1) was used by substituting $\overline{C}_s$ by $\overline{X}_s$, i.e. the linear regression model is given by:

$$p_s = \alpha + \sum_{j=1}^{3} \beta_j\ z_{js} + \gamma\ \overline{\mathrm{X}}_s + \varepsilon_s \tag{2.5}$$

where $\overline{X}_s = \sum_{j=1}^{s} x_j / n$ is the average cumulative cost (time spent in minutes) per sampled unit up to and including time slice s; x_j is the time spent during time slice *j* and *n* is the total number of sampled units for the regional office being analyzed.

In the case of linear regression model (3.4), the associated predicted probability for time slice *s* is given by:

$$\tilde{p}_s = \left[\hat{\alpha} + \hat{\beta}_1 z_1^{(s)} + \hat{\beta}_2 z_2^{(s)} + \hat{\beta}_3 z_3^{(s)} + \frac{\hat{\gamma}}{n} \sum_{j=1}^{s-1} \left\{ t_1 \tilde{p}_j c_j + t_2 \left(1 - \tilde{p}_j\right) c_j \right\} + \frac{\hat{\gamma}}{n} t_2 c_s \right] / K_s$$

where $K_s = \left[1 - \frac{\hat{\gamma}}{n} \left\{ t_1 - t_2 \right\} c_s \right]$: t_1 is the unit cost (time in minutes) for completing a questionnaire (productive call) and t_2 is the unit cost (time in minutes) for not completing a questionnaire (non-productive call). The above expression for the predicted probability can be derived by setting X_s equal to $\frac{1}{n} \sum_{t=1}^{s} \left[t_1 \tilde{p}_t c_t + t_2 \left(1 - \tilde{p}_t\right) c_t \right]$.

Thus, the predicted probability for the time slice *s* can be defined in terms of cumulative per unit time spent during the previous time slices which in turn depends on the predicted probabilities for the previous (s-1) time slices.

We did not consider logistic regression model using average cumulative time spent up to and including the time lice s because the predicted probabilities for the optimization algorithm would have to be obtained numerically which would be too cumbersome.

3.3 Optimum Number of Calls by Time Slice

The cost of making CATI calls for a given time slice can be expressed as a linear combination of productive calls (a questionnaire is completed) and non-productive calls (a questionnaire is not completed). Let t_1 be the unit cost (time in minutes) when the call results in completed questionnaire (productive call) and t_2 be the unit cost (time in minutes) when a questionnaire is not completed (non-productive call). The predicted cost in making calls for a given time slice can be expressed as $\tilde{x}_j = t_1\ \tilde{p}_j\ c_j + t_2 \left(1 - \tilde{p}_j\right) c_j$, where $\tilde{p}_s$ is determined from one of the above two models. The total data collection cost is given by the function $g\left(\underset{\sim}{c}\right)$ defined as

$$g\left(\underset{\sim}{c}\right) = \sum_{j=1}^{S} \left\{ t_1 \tilde{p}_j\ c_j + t_2 \left(1 - \tilde{p}_j\right) c_j \right\}$$

The "call" vector $\underset{\sim}{\boldsymbol{c}} = \left(c_1, c_2, ..., c_S\right)$ is determined such that the function $g\left(\underset{\sim}{c}\right)$ is minimized subject to the following constraints:

i. The number of calls for each time slice is greater than or equal to zero, and

ii. The expected response rate $\sum_{j=1}^{S} \tilde{p}_j c_j / n$ is equal to a pre-specified response rate *R*.

Additional constraints, e.g. upper and lower bounds on number of calls and/or cost (time spent) by time slice, would result in decreased cost savings. It should be noted that the predicted cost for the time slice s is given as $t_1 \tilde{p}_s x_s + t_2 (1 - \tilde{p}_s) x_s$.

3. Summary Results and Conclusion

The methodology and results presented in this paper only represents the first phase in the development of an optimized interviewer scheduling tool for a single CATI survey. The results of this study indicated that the inclusion of an intercept term and continuous variable was significant for all six regional offices. Although the best times to call varied between regional offices, three patterns emerged: morning, late evening, early and late evening. Also, the number of calls within each time slice should be constant, as opposed to quite variable (present case currently). Given the re-allocation of calls via the optimization procedure, gains up to 22% could be achieved for the targeted response rates. It is important to note that these gains are over-optimistic as they do not include interviewer operational constraints such as their availability for each day of the week, sick leave and vacation, assigned hours per quarter as well as their work shift preferences. Details of these gains are given in Table 1.

Table 1: Percentage Savings in terms of Minutes

Regional Office	Linear regression	Logistic Regression
1	11.2	17.2
2	5.9	11.5
3	22.6	23.3
4	17.6	17.2
5	17.2	14.1
6	0.3	0.1
Canada	12.9	14.9

We are presently working on accommodating two surveys within this framework. Furthermore, the interviewer workload is not optimized over several concurrent surveys. Additional constraints also need to be accounted for. These include: Legal , ergonomic, and operating constraints; Minimum

and maximum number of days that interviewers work within the week; Shift duration per day (no more or less than a fixed number of hours), including starting time range of each interviewer; and reasobabel number of shifts within a day. If these addtional constaints cannot be accomodated within our optimization procedure, we can try out commercial software such as XIMES: XIMES can account for such constraints, and schedule each interviewer by time shift (Gartner, Musliu, and Slany 2001)

4. References

Couper, M. P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J. Nicholls W.L. and O'Reilly, J.M. (1998). Computer Assisted Survey Information Collection. Wiley series in survey methodology section, Chapter 15 by Edwards, Suresh and Weeks, 301-306.

Laflamme, F., (2009). Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada. Proceedings from the 57th International Statistical Institute Conference.

Slany, W., Musliu, N., Kortsarz, G. and Gärtner J.(2000). Theory and practice of shift scheduling (invited paper). RIMS Kokyuroku of the Research Institute of Mathematical Sciences, Kyoto University, 1185: 172-181, 2000.

A Theoretical Framework for Adaptive Collection Designs

Jean-François Beaumont, Statistics Canada
David Haziza, Université de Montréal
June 10, 2011

Abstract

We present a theoretical framework for adaptive collection designs in the context of computer-assisted telephone interview surveys. By adaptive collection designs, we mean any procedure of calls prioritization and resources allocation that is dynamic as data collection progresses; i.e., the procedure uses paradata or other information to adapt itself to what is observed during data collection. We focus on calls prioritization. The goal of an adaptive collection design is to increase quality for a given cost or alternatively to reduce cost for a given quality. The literature has essentially focused on finding collection designs that lead to a reduction of nonresponse bias of an estimator that is not adjusted for nonresponse. Thus, improvement of quality is associated with nonresponse bias reduction. We argue that it is not the best criterion to use as the bias that can be removed at the data collection stage of a survey through an adaptive collection design can also be removed at the estimation stage through an appropriate nonresponse weight adjustment procedure. Instead, we minimize the nonresponse variance of an estimator that is adjusted for nonresponse. We develop a procedure of calls prioritization that attempts to achieve this goal.

1. Introduction: selected literature review

The literature on adaptive collection designs, sometimes called adaptive survey designs, responsive collection designs, responsive survey designs or simply responsive designs, is fairly recent. In our context, we prefer the terms adaptive collection designs and responsive collection designs as they make it clear that we are concerned with improvements in data collection methods so that they avoid any confusion with the different notion of adaptive sampling designs, which are typically used to sample from rare populations..

Groves and Heeringa (2006) defined a responsive survey design as one that uses paradata (i.e., data about the data collection process) to guide changes in the features of data collection in order to achieve higher quality estimates per unit cost. Two examples of features of data collection are the data collection mode and the use of incentives. The implementation of responsive designs in practice requires to define quality and to determine suitable quality indicators. A cost function must also be chosen. There are two other main concepts underlying the Groves and Heeringa (2006) framework: phase and phase capacity. A phase is a period of data collection during which the same set of methods is used. The first phase is used to gather information about data collection features. In subsequent phases, features are modified (e.g., subsampling of nonrespondents, larger incentives, etc.). A given phase is continued until it reaches its phase capacity, which is typically judged by the stability of some indicator (e.g., an estimate) as the phase matures.

Schouten, Cobben and Bethlehem (2009) proposed an indicator of nonresponse bias, called R-indicator, as an alternative to response rates. An R-indicator is sometimes chosen as the quality indicator to be used in conjunction with an adaptive collection design. The proposed R-indicator is a function of estimated probabilities of response to the survey. One drawback of this indicator is that it depends on the proper choice of a nonresponse model; in particular, the proper choice of explanatory variables. For instance, if no explanatory variable is included in the nonresponse model, the indicator is equal to 1, which is the best value it can reach. Thus, a poor choice of explanatory variables may lead to an artificially large value of the indicator but does not tell anything about the actual nonresponse bias. Indeed, the nonresponse bias may vary from one variable of interest to another. Since the R-indicator is independent of any of these variables, it can only provide limited information about nonresponse bias. The authors also proposed to consider the maximal bias of an estimator that is not adjusted for nonresponse (no adjustment of design weights). This additional measure is related to the R-indicator and depends on the variable of interest. Like the R-indicator, the maximal bias depends on the proper specification of a nonresponse model. Another limitation of the maximal bias is that it is based on an estimator that is rarely used in practice: the unadjusted estimator.

Peytchev, Riley, Rosen, Murphy and Lindblad (2010) investigated an approach to reducing nonresponse bias through case prioritization. They suggested targeting individuals with lower estimated response probabilities. For instance, they could be given larger incentives or interviewers could have larger incentives for completing these cases. Their approach is basically equivalent to trying to increase the R-indicator (or achieving a more balanced sample). They also recommended using explanatory variables that are associated with the variables of interest so that the R-indicator is also indirectly associated with these variables.

Laflamme and Karaganis (2010) developed and implemented responsive collection designs for Computer-Assisted Telephone Interview (CATI) surveys at Statistics Canada. Their approach fits well into the Groves and Heeringa (2006) framework. They considered four phases: a planning phase, an initial collection phase and two responsive design phases. The planning phase is conducted before data collection starts. It consists of analyzing previous data, determining strategies, etc. The initial collection phase is used to evaluate different indicators to determine when the next phase should start. It is the first phase of the Groves and Heeringa (2006) framework. The two responsive design phases differ in the way cases are prioritized. The goal of the first responsive design phase is to improve response rates by targeting individuals with higher estimated response probabilities. This tends to increase the number of respondents, which is desirable. The goal of the second responsive design phase is to reduce the variability of response rates between domains of interest, which is essentially equivalent to increasing the R-indicator. This will likely reduce the variability of weight adjustments, which is also desirable. Note that objectives of both phases are intuitively appealing but may be contradictory in terms of cases prioritization. Laflamme and Karaganis (2010) tried to achieve a compromise between these conflicting objectives by separating data collection into two responsive design phases, each one focusing on a single objective. Our approach tries to make a compromise by using a single objective function (quality indicator).

Schouten, Calinescu and Luiten (2011) proposed an interesting theoretical framework for adaptive survey designs. It is apparently the first paper that develops some theory in this topic. The authors suggested maximizing quality for a given cost or, equivalently, minimizing cost for a given quality. The framework requires the choice of a quality indicator such as the overall response rate, the R-indicator, the maximal bias, etc. The authors do not provide any recommendation regarding the choice of an appropriate indicator and a cost function. Our approach fits into this framework in the sense that we maximize quality for a given cost. In the next section, we argue for a specific quality indicator and cost function. The mathematical details will be coming later in a longer version of this paper.

2. Our framework

We define an adaptive collection design as any procedure of calls prioritization or resources allocation that is dynamic as data collection progresses; i.e., the procedure uses paradata or other information to adapt itself to what is observed during data collection. In this paper, we focus on calls prioritization and CATI surveys. The reason for restricting to CATI surveys is that it is easier to come up with a cost function since the overall cost is highly related to the total time used to conduct data collection. We also focus on maximizing quality for a given cost.

Regarding the choice of a quality indicator, the literature has mainly focused on the nonresponse bias of an estimator that is not adjusted for nonresponse. This leads to considering the R-indicator or the maximal bias of the unadjusted estimator as a quality measure. In our view, the focus should not be placed on nonresponse bias because any bias that can be removed at the collection stage through an adaptive collection design can also be removed at the estimation stage through an appropriate nonresponse weight adjustment procedure. In other words, the information that is used at the data collection stage to reduce nonresponse bias can also be used at the estimation stage. For instance, we could consider an adaptive collection design that tries to equalize response rates between domains of interest and use the unadjusted estimator. In terms of nonresponse bias we expect this strategy to be equivalent to using an estimator that adjusts design weights by the inverse of response rates within domains of interest even though no adaptive collection design has been implemented.

Instead of considering the nonresponse bias, we suggest reducing the nonresponse variance of an estimator adjusted for nonresponse. Our adjusted estimator is obtained by multiplying design weights by the inverse of response rates within cells. This estimator is asymptotically unbiased assuming that nonresponse is uniform within cells. Our objective consists of finding the target probabilities of response to the survey (one for each cell) that minimize the nonresponse variance of the adjusted estimator subject to a fixed expected overall cost. The expected overall cost depends on the cost of conducting an interview, the cost of a missed call, the target response probabilities and the expected number of attempts made at the end of data collection for each unit. This number depends on the response probability at each attempt, the maximum number of calls allowed per unit and the target response probability. To simplify derivations, we assume that the expected number of attempts does not depend on the target response

probability. As a result, the expected overall cost becomes a linear function of the target response probabilities. The solution to the above optimization problem is similar to sample size allocation in stratified sampling. It is equivalent to maximizing the R-indicator only in a very specific scenario; i.e., when the target response probabilities are found to be all equal.

Once the target response probabilities have been determined, we must find the effort (number of attempts) needed to achieve these target probabilities. Then, our procedure consists of selecting cases to be interviewed with probability proportional to the effort. Obviously, the effort for a given unit increases with the target response probability and decreases with the response probability at each attempt. A larger response probability at each attempt indicates that this unit is easier to contact and thus requires less effort to achieve the target response probability. Note that it might be advisable to ensure that the (estimated) response probability at each attempt is not too low so as to avoid unduly large efforts for some units. It might also be advisable to ensure that a certain time has elapsed between two consecutive calls.

It could be useful to graph the minimum nonresponse variance as a function of the expected overall cost. The nonresponse variance should decrease as the cost increases. There may be a value of cost where the variance reduction is negligible when the cost is increased and it may not be justified to spend more than that value.

The solution to the above optimization problem is found before data collection starts. However, it may be a good idea to revise the solution periodically (e.g., daily) as data collection progresses. Some parameters might need to be modified (e.g., the estimated response probability at each attempt) and the remaining budget and expected overall cost need to be updated. The solution to the revised optimization problem is similar to the initial solution. The revised target response probabilities need also to account for the current number of respondents. In some cells, the current response rate might be closer to the target response probability than in other cells, which means that less effort should be spent in such cells.

3. Conclusion

The next steps of this research are:

- to perform a simulation study to evaluate the usefulness of the theory;
- to adapt the theory for practical applications; and
- to test the approach in a real production environment.

We have justified using the nonresponse variance of an estimator that is adjusted for nonresponse as a possible quality indicator to be optimized. Other indicators could possibly be useful. Some more thinking might lead to justifying alternative indicators in the context of adaptive collection designs.

Finally, if one is really interested in reducing nonresponse bias, it appears to us than subsampling of nonrespondents might be the only option. Questions such as the appropriate

subsampling rate or subsampling method need further investigations. Since there would likely be nonresponse in the subsample, our adaptive approach could be used within the subsample.

References

Groves, R.M., and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society*, *Series A*, 169, 439-457.

Laflamme, F., and Karaganis, M. (2010). Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada. In *Proceedings of the European Conference on Quality in Official Statistics*, Helsinki, Finland, May 2010.

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of Nonresponse Bias in Surveys through Case Prioritization. *Survey Research Methods*, 4, 21-29.

Schouten, B., Calinescu, M., and Luiten, A. (2011). Optimizing Quality of Response through Adaptive Survey Designs. Discussion paper, Statistics Netherlands.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Overview of error model for estimates of foreign-born immigration using citizenship and residence one year ago from the American Community Survey

Mary H. Mulry, U.S. Census Bureau[5]

Research Design

Demographic Analysis (DA) estimates of the U.S. population on April 1, 2010 included estimates of foreign-born immigration each year between 2000 and 2010 based on data from the American Community Survey (ACS). Our goal is to design methodology to describe the uncertainty in the estimates of foreign-born immigration. The main estimation method we are considering in this paper uses the responses to two questions, one that asks citizenship and another that asks residence one year ago (ROYA). We use an error model that accounts for sampling and nonsampling errors. Unfortunately, time and resource limitations prevent us from conducting studies to measure the nonsampling errors so we intend to propose reasonable estimates based on studies of nonsampling errors in ACS for other purposes or studies of nonsampling errors in other surveys.

This paper describes our strategy to use the error model in the design of a simulation to study the propagation of errors in the estimates of foreign-born immigration. The results of the simulation study will produce a sensitivity analysis that assesses the uncertainty in estimates of foreign-born immigration.

Estimator for foreign-born immigration. For a population control cell C, define:

P = size of population in cell C

F = size of the foreign-born population in cell C

Y = size of foreign-born population in cell C who resided outside the U.S. one year ago

$s_F = F/P$ = proportion of the population within cell C who are foreign-born

$r_Y = Y/F$ = proportion of the foreign-born within cell C with ROYA outside the U.S.

[5] **Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.**

If we let $\hat{P}$ be the estimate of the population size *P* for cell C from the Population Estimates Program (PEP) used as the population control, the estimator of the foreign-born immigration for cell C, $\hat{T}$, has the form

$$\hat{T} = \hat{P}\hat{s}_F\hat{r}_Y$$

where $\hat{s}_F$ and $\hat{r}_Y$ are estimators of s_F and r_Y from the ACS using survey weights before the population controls are applied.

Error model for foreign-born immigration. The estimate of the foreign-born immigration may be affected by

- errors in PEP estimates
- ACS data errors that create error in the ACS estimate of the foreign-born who resided outside the U.S. one year ago
- errors caused by an inconsistency between the ACS and the PEP in the variables used to form cells for the ratio adjustment.

Each year, the immigration component of the PEP estimates uses ACS data collected the previous year. Therefore, since the full implementation of the ACS, there has been no overlap in the ACS data used in estimating the foreign-born immigration for two consecutive years.

The error in the estimate of the foreign-born immigration $\hat{T}$ may be expressed in terms of a bias component β and a random error component ε. Then the form of the model is

$$\hat{T} = P + \beta + \varepsilon \text{ , where } E(\varepsilon) = 0 .$$

The bias β may be expressed as the sum of the bias due to inconsistency in the reporting of characteristics between PEP and ACS β_I, and the bias due to error in the data used to form the PEP estimation and the ACS estimation β_D, $\beta = \beta_I + \beta_D$. The random error ε has terms due to sampling variance ε_S and variance due to the imputation for missing data ε_M, $\varepsilon = \varepsilon_S + \varepsilon_M$.

Error from inconsistent characteristics between ACS and PEP. Sources of inconsistencies in measurement of characteristics used in population control cells between the PEP and ACS:

- Coding of responses of race and Hispanic ethnicity in the ACS
- Changes in reporting of race/Hispanic ethnicity since Census 2000, which is the base for the PEP
- Differences in characteristics used in forming ratio adjustment cells between the ACS and the data used to form updated PEP estimates during the decade. These differences could be caused by errors or changes in record keeping by the data sources.

PEP data error. Sources of potential errors in estimates of population size from the PEP include:

- Coverage error in the Census 2000 numbers, which are used as a base
- Errors in the data used to form updated population estimates during the decade.

ACS data collection errors. Potential sources of ACS data collection errors that may contribute to bias are:

- Frame coverage error
- Net error in whether foreign-born or native-born
- Error among foreign born that leads to error in whether lived outside U.S. one year ago
- Address errors
- Roster errors

To form the estimates of the terms in the data collection bias components, we need to consider the sequence of response modes used by ACS because many of the interviews of foreign-born respondents tend to occur by telephone or in person. If we let the subscripts *m, t,* and *p* represent the mail responses, the telephone responses, and the in-person responses, respectively, the estimator of bias due to data collection error is the weighted sum of estimated biases for the modes,

$$\hat{\beta}_{dc} = w_m\hat{\beta}_m + w_t\hat{\beta}_t + w_p\hat{\beta}_p$$

where the weights are the weighted proportion of the responses by each mode. When estimating the bias for proportions, the weights are calculated using the proportion of responses for the denominator.

ACS data processing errors. Potential sources of ACS data processing errors fall into three categories:

- Errors in editing (citizenship and ROYA)
- Errors in coding (ROYA only)

- Errors in keying (ROYA only)

These types of errors may occur for responses. When questions left blank are imputed, any error is considered imputation error.

Simulation and analyses. Once the nonsampling errors, their variances, and covariance matrix are estimated, the simulation will draw repeatedly and independently from their joint distribution to produce the distribution of a bias estimate. The probability distributions will be centered on the observed values adjusted for the estimated biases, and their random component will be derived from a multivariate normal specification with mean vector equal to zero and estimated covariance matrix. Simulation from this distribution will yield distributions of the estimates of immigration. Differences between the mean of the latter distribution and the original estimate indicate the estimated biases in the original estimates, and the standard deviations indicate the standard deviations of the sampling distributions.

As shown in the discussion of the joint error distribution, the probability models will be developed somewhat differently for (1) sampling error, (2) error from missing data, (3) effect of inconsistent classification, and (4) other frame coverage errors, data collection errors, and data processing errors.

Part of the research will be to specify the domains for the analysis. The following statistics will be computed from the simulated distribution: (i) estimate of bias, (ii) estimate of standard deviation (reflecting both sampling error and random nonsampling errors), and (iii) deciles of the distribution. The analyses will include a sensitivity analysis to aid in determining the most influential error sources and error types.

Questions

(1) Are there relevant studies of errors in a similar citizenship question or in a similar question regarding residence one year ago?
(2) Are there other error components that should be included?

Robust inference in two-phase sampling designs with application to unit nonresponse

David Haziza and Jean-François Beaumont

Université de Montréal and Statistics Canada

1. Introduction

Two-phase sampling is often used in surveys when the sampling frame contains little or no auxiliary information. In this case, it may be wise to first select a large sample in order to collect data on variables that are inexpensive to obtain and that are related to the variables of interest. Then, using the variables observed in the first-phase, an efficient sampling procedure can be used to select a (typically small) subsample from the first-phase sample in order to collect the variables of interest. The theory on two-phase sampling can also be helpful in the context of unit nonresponse because the set of respondents is often viewed as a second phase sample.

Influential units occur frequently in surveys, especially in the context of business surveys that collect economic variables whose distribution is highly skewed. The presence of influential units in the sample does not introduce a bias but lead generally to very unstable estimators. Methods for dealing with influential units include winzorization and M-estimation; see e.g., Beaumont and Rivest (2009) and Beaumont, Haziza and Ruiz-Gazen (2011).

In this paper, we extend the results of Beaumont, Haziza and Ruiz-Gazen (2011) for uni-phase sampling designs, who suggested constructing robust estimators of population totals using the concept of conditional bias of a unit; see Moreno-Rebollo etal. (1999) and Moreno-Rebollo et al. (2002). The conditional bias of a unit can be seen as a measure of influence that fully accounts for the sampling design.

The outline of the paper is as follows: in Section 2, we describe the theoretical framework. The concept of conditional bias for two-phase sampling designs is presented in Section 3. Then, in Section 4, we consider a general robust estimator of a population total based on the estimated conditional bias. Finally, the application to the case of unit nonresponse is discussed in Section 5.

2. Set-up

Consider a population U of size N. We are interested in estimating the population total $Y = \sum_{i \in U} y_i$ of a study variable y. We select a sample according to a two-phase sampling design: in the first phase, a sample s_1, of size n_1, is selected from U according to a given sampling design $p(s_1)$. In the second phase, s_2, of size n_2, is selected from s_1 according to $p(s_2|s_1)$. For simplicity, we

assume that $p(s_2 \mid s_1) = p(s_2)$. That is, we consider two-phase designs which satisfy the invariance property. The results can be extended to two-phase designs that do not satisfy this property.

Let I_{1i} be a sample selection indicator attached to unit i such that $I_{1i} = 1$ if unit i is selected in s_1 and $I_{1i} = 0,$ otherwise, and let $\mathbf{I}_1 = (I_1, \cdots, I_N)'$. Let I_{2i} be a sample selection indicator attached to unit i such that $I_{2i} = 1$ if unit i is selected in s_2 and $I_{2i} = 0,$ otherwise. Let $\pi_{1i} = P(I_{1i} = 1)$ and $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ denote the first-order and second-order probabilities in s_1. Similarly, let $\pi_{2i} = P(I_{2i} = 1 \mid I_{1i} = 1)$ and $\pi_{2ij} = P(I_{2i} = 1, I_{2j} = 1 \mid I_{1i} = 1, I_{1j} = 1)$ denote the first-order and second-order probabilities in s_2.

In the absence of nonsampling errors, an estimator of *Y* is the double expansion estimator

$$\hat{Y}_{DE} = \sum_{i \in U} \pi_{1i}^{-1} \pi_{2i}^{-1} y_i I_{1i} I_{2i}. \qquad \textbf{(1)}$$

To study the properties of **Error! Reference source not found.**, we express its total error as

$$\hat{Y}_{DE} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{DE} - \hat{Y}_E), \qquad \textbf{(2)}$$

where $\hat{Y}_E = \sum_{i \in U} \pi_{1i}^{-1} y_i$ denotes the expansion estimator that one would have used had the design been a single phase design. The terms $\hat{Y}_E - Y$ and $\hat{Y}_{DE} - \hat{Y}_E$ on the right hand side of **Error! Reference source not found.** denote the errors due to the first phase and second phase, respectively. Let $E_1(.)$ and $V_1(.)$ denote the expectation and variance with respect to the first phase and $E_2(. \mid \mathbf{I}_1)$ and $V_2(. \mid \mathbf{I}_1)$ denote the conditional expectation and conditional variance with respect to the second phase. Noting that $E_2(\hat{Y}_{DE} \mid \mathbf{I}_1) = \hat{Y}_E$ and $E_1(\hat{Y}_E) = Y,$ it follows from **Error! Reference source not found.** that $E_p(\hat{Y}_{DE}) \equiv E_1 E_2(\hat{Y}_{DE} \mid \mathbf{I}_1) = Y;$ that is, $\hat{Y}_{DE}$ is design-unbiased for Y. Also, using **Error! Reference source not found.**, the design variance of $\hat{Y}_{DE}$ can be expressed as

$$V_p(\hat{Y}_{DE}) = V_1 E_2(\hat{Y}_{DE} \mid \mathbf{I}_1) + E_1 V_2(\hat{Y}_{DE} \mid \mathbf{I}_1)$$

$$= \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij}^*}{\pi_i^* \pi_j^*} - 1 \right) y_i y_j, \qquad \textbf{(3)}$$

where $\pi_i^* = \pi_{1i} \pi_{2i}$ and $\pi_{ij}^* = \pi_{1ij} \pi_{2ij}$.

In the presence of influential units, the estimator **Error! Reference source not found.** remains design-unbiased. However, its design variance may be very large. In other words, including or excluding an influential unit from the calculations may have an important impact on the magnitude of the total error, $\hat{Y}_{DE} - Y$. Note that an influential unit may have a large impact on

the first phase error, $\hat{Y}_E - Y$, and/or on the second-phase error, $\hat{Y}_{DE} - \hat{Y}_E$. Next, we define a measure of influence.

3. Measuring the influence: the conditional bias

For uni-phase sampling designs, Moreno-Rebollo etal. (1999) and Moreno-Rebollo et al. (2002) introduced the concept of conditional bias attached to a unit as a measure of influence; see also Beaumont, Haziza and Ruiz-Gazen (2011). We extend this concept to the case of two-phase designs. We distinguish between three types of units: (i) the sample units, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 1$; (ii) the units selected in the first-phase sample but not in the second phase, i.e., the units for which $I_{1i} = 1$ and $I_{2i} = 0$ and (iii) the non-selected units, i.e., the units for which $I_{1i} = 0$ and $I_{2i} = 0$. It is worth noting that each type of unit may have an influence on the total error. However, only the influence of the sample units can be reduced at the estimation stage. In other words, nothing can be done for (ii) and (iii) at this stage.

The conditional bias of sampled unit i is defined as:

$$\begin{aligned} B_i^{DE}(I_{1i}=1, I_{2i}=1) &= E_1E_2(\hat{Y}_{DE} - Y \mid \mathbf{I}_1, I_{1i}=1, I_{2i}=1) \\ &= E_1(\hat{Y}_E - Y \mid I_{1i}=1) + E_1E_2(\hat{Y}_{DE} - \hat{Y}_E \mid \mathbf{I}_1, I_{1i}=1, I_{2i}=1). \end{aligned}$$

For an arbitrary two-phase design, we obtain

$$\begin{aligned} B_i^{DE}(I_{1i}=1, I_{2i}=1) &= \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1\right) y_j + \sum_{j\in U}\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}\left(\frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}} - 1\right) y_j \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}^*}{\pi_i^*\pi_j^*} - 1\right) y_j. \end{aligned} \qquad \textbf{(4)}$$

Example 1: Simple random sampling in both phases: the conditional bias **Error! Reference source not found.** reduces to

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \frac{N}{(N-1)}\left(\frac{N}{n_2} - 1\right)(y_i - \bar{Y}),$$

where $\bar{Y} = Y/N$. The previous expression suggest that a unit has a large influence if its "total weight", $\pi_i^{*-1} = Nn_2^{-1}$, is large and/or if its y-value is far from the population mean $\bar{Y}$.

Example 2: Poisson sampling at both phases: the conditional bias **Error! Reference source not found.** reduces to

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \left(\pi_i^{*-1} - 1\right) y_i.$$

Hence, a unit has a large influence if its "total weight" π_i^{*-1} is large and/or if its y-value is large.

Example 3: Arbitrary design in the first phase and Poisson sampling in the second phase: the conditional bias **Error! Reference source not found.** reduces to

$$B_i^{DE}(I_{1i}=1, I_{2i}=1) = \sum_{j\in U}\left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}}-1\right)y_j + \pi_{1i}^{-1}\left(\pi_{2i}^{-1}-1\right)y_i. \qquad \textbf{(5)}$$

The previous expression will be particularly useful in the context of unit nonresponse.

In general, note the conditional bias depends on unknown population parameters that should be estimated robustly or using an independent source of data. The resulting estimated conditional bias is denoted by $\hat{B}_i^{DE}(I_{1i}=1, I_{2i}=1)$. Also, it follows from **Error! Reference source not found.** and **Error! Reference source not found.** that the design variance of $\hat{Y}_{DE}$ can be expressed as

$$V_p(\hat{Y}_{DE}) = \sum_{i\in U} B_i^{DE}(I_{1i}=1, I_{2i}=1)y_i,$$

illustrating that the conditional bias of a unit can be interpreted as its contribution to the design variance of $\hat{Y}_{DE}$.

4. Robustifying the double expansion estimator

Following Beaumont, Haziza and Ruiz-Gazen (2011), we consider the following robust version of $\hat{Y}_{DE}$:

$$\hat{Y}_{DE}^R = \hat{Y}_{DE} - \sum_{i\in s_2}\left\{\hat{B}_i^{DE}(I_{1i}=1, I_{2i}=1)\right\} + \sum_{i\in s_2}\psi\left\{\hat{B}_i^{DE}(I_{1i}=1, I_{2i}=1); c\right\}, \qquad \textbf{(6)}$$

where $\psi(.)$ is a function, which role consists of curbing the impact of influential units and c is a tuning constant whose value must be determined. A popular ψ -function is the so-called Huber function given by

$$\psi(t) = \begin{cases} c & \text{if } t > c \\ t & \text{if } |t| \le c \\ -c & \text{if } t < -c \end{cases}$$

When $\pi_{2i}=1$ for all $i\in s_2$ (i.e., case of a single phase sampling design), the robust estimator **Error! Reference source not found.** reduces to that proposed by Beaumont, Haziza and Ruiz-Gazen (2011).

5. Application to unit nonresponse

In this section, we consider the problem of robust estimation in the context of unit nonresponse. In this context, s_1 denotes the sample selected from the population, whereas s_2 denotes the random set of respondents, I_{1i} and I_{2i} denotes respectively the sample selection indicator and the response indicator attached to unit i. Also, π_{1i} and π_{2i} denotes respectively the inclusion probability in the sample and the response probability for unit i. We assume that the units respond independently of one another; that is $\pi_{2ij} = \pi_{2i}\pi_{2j}$. This situation is identical to that of Example 3 (see Section 3), except that the π_{2i}'s are now unknown. If the π_{2i}'s were known, a propensity score adjusted (PSA) estimator would be given by (1) and the conditional bias of a responding unit would be given **Error! Reference source not found.**. In practice, the response probabilities π_{2i} are unknown and must be estimated. We assume that they can be parametrically modeled by:

$$\pi_{2i} = m(\mathbf{x}_i, \boldsymbol{\alpha}), \qquad \textbf{(7)}$$

where $m(.)$ is a known function, $\mathbf{x}$ is a vector of auxiliary variables available for all the sampled units (respondents and nonrespondents) and $\boldsymbol{\alpha}$ is a vector of unknown parameters. A special case of **Error! Reference source not found.** is the logistic regression model, which is frequently used in practice. Based on **Error! Reference source not found.**, an estimator of π_{2i} is given by $\hat{\pi}_{2i} = m(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, where $\hat{\boldsymbol{\alpha}}$ denotes a suitable estimator of $\boldsymbol{\alpha}$. A PSA estimator of Y is thus given by

$$\hat{Y}_{PSA} = \sum_{i \in U} \frac{1}{\pi_{1i}\hat{\pi}_{2i}} y_i I_{1i} I_{2i}. \qquad \textbf{(8)}$$

The total error of $\hat{Y}_{PSA}$ can be expressed as

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_{PSA} - \hat{Y}_E). \qquad \textbf{(9)}$$

The terms $\hat{Y}_E - Y$ and $\hat{Y}_{PSA} - \hat{Y}_E$ in **Error! Reference source not found.** denote the sampling error and the nonresponse error, respectively. Using a first-order Taylor expansion, we have

$$\hat{Y}_{PSA} - \hat{Y}_L = O_p\left(\frac{N}{n}\right), \qquad \textbf{(10)}$$

where

$$\hat{Y}_L = \frac{1}{N}\sum_{i \in s_1} \pi_{1i}^{-1} \left\{ \pi_{1i}\pi_{2i}\mathbf{h}_i'\hat{\boldsymbol{\gamma}} + \frac{I_{2i}}{\pi_{2i}}\left(y_i - \pi_{1i}\pi_{2i}\mathbf{h}_i'\hat{\boldsymbol{\gamma}}\right) \right\}$$

with $\mathbf{h}_i = \partial\{logit(\pi_{2i})\}/\partial\boldsymbol{\alpha}$ and

$$\hat{\boldsymbol{\gamma}} = \left\{ \sum_{i \in s_1} \pi_{2i}(1-\pi_{2i}) \mathbf{h}_i \mathbf{h}_i' \right\}^{-1} \sum_{i \in s_1} \pi_{1i}^{-1}(1-\pi_{2i}) \mathbf{h}_i y_i ;$$

see Kim and Kim (2007). Using **Error! Reference source not found.** in **Error! Reference source not found.**, we obtain

$$\hat{Y}_{PSA} - Y = (\hat{Y}_E - Y) + (\hat{Y}_L - \hat{Y}_E) + O_p\left(\frac{N}{n}\right). \qquad \textbf{(11)}$$

Ignoring the higher order terms in **Error! Reference source not found.**, the conditional bias of responding unit *i* can be approximated by

$$B_i^{PSA}(I_{1i}=1, I_{2i}=1) = E_1 E_2(\hat{Y}_{PSA} - Y \mid \mathbf{I}_1, I_{1i}=1, I_{2i}=1)$$
$$\approx E_1(\hat{Y}_E - Y \mid I_{1i}=1) + E_1 E_2(\hat{Y}_L - \hat{Y}_E \mid \mathbf{I}_1, I_{1i}=1, I_{2i}=1)$$

After some tedious but relatively straightforward algebra, we obtain

$$B_i^{PSA}(I_{1i}=1, I_{2i}=1) \approx \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j - \pi_{1i}^{-1}(\pi_{2i}^{-1} - 1)\left(y_i - \mathbf{c}_i' \right)$$
$$- \mathbf{c}_i' \mathbf{T}^{-1} \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right)(1-\pi_{2j})\left(y_j - \mathbf{c}_j' \right),$$

where $\mathbf{c}_i = \pi_{1i}\pi_{2i}\mathbf{h}_i$ and

$$\boldsymbol{\gamma} = \mathbf{T}^{-1} \sum_{i \in U} (1-\pi_{2i}) \mathbf{h}_i y_i$$

with $\mathbf{T} = \sum_{i \in U} \pi_{1i}\pi_{2i}(1-\pi_{2i}) \mathbf{h}_i \mathbf{h}_i'$. Finally, a robust version of $\hat{Y}_{PSA}$ is given by

$$\hat{Y}_{PSA}^R = \hat{Y}_{PSA} - \sum_{i \in s_2} \hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1) + \sum_{i \in s_2} \psi\left\{ \hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1); c \right\},$$

where $\hat{B}_i^{PSA}(I_{1i}=1, I_{2i}=1)$ is a suitable estimator of $B_i^{PSA}(I_{1i}=1, I_{2i}=1)$.

REFERENCES

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2011). A unified approach to robust estimation in finite population sampling. *In revision for Biometrika*.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers with survey data}. Handbook of Statistics, Volume 29, Chapter 11, Sample Surveys: Theory Methods and Inference, Editors: C.R. Rao and D. Pfeffermann, 247-279.

Kim, J.K. and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501-514.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923-968.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: estimating the conditional bias. *Metrika,* 55, 209-214.

The estimation strategy of the National Household Survey

Summary for the 2011 International Total Survey Error Workshop

François Verret

Senior Methodologist

Social Survey Methods Division

Statistics Canada

Coauthored by Mike Bankier, Wesley Benjamin and Lisa Hayden

1. Introduction

In the 2006 Canadian Census a random sample of 20% of all households received a long form questionnaire while the other 80% received a short form questionnaire, each form being mandatory. In the 2011 Census, every household will receive a mandatory short form. In place of the mandatory long form, about 30% of the households will also be asked to complete the National Household Survey (NHS) voluntarily about a month after the Census. As was the case for the 2006 long form, the information collected in the NHS will provide crucial data for planning, delivering and supporting federal, provincial/territorial and local government programs directed at target populations. The Census short form collects data on demography, dwelling type, family structure and language, whereas the NHS long form collects data on topics such as education, ethnicity, income, immigration, labor and mobility. The sample size of the NHS is 4.5 million households and the expected number of responding households is 2.6 million. The smallest geographical domains targeted by the NHS (and the Census) are called dissemination areas and have an average population size of approximately 300 households.

2. Errors in the NHS

The 2006 Census short and long forms shared the same general infrastructure, reference date and target population. This is still the case for the 2011 Census and NHS (the NHS target population is in a fact a subset of the Census target population). Moreover, there is a direct correspondence between the 2011 Census and NHS dwellings at all times during the survey process (sampling, collection, estimation, etc.). Hence, both surveys share many similarities with regards to non-sampling errors such as coverage, measurement and processing errors. Census coverage is measured by the Reverse Record Check Study, the Census Overcoverage

Study and the Dwelling Classification Survey. The measurement and processing errors are reduced, to some extent, by the 2011 Census and NHS coding process, the Response Integration and Verification Task and by edit and imputation using the Canadian Census Edit and Imputation System (CANCEIS).

However, the non-sampling error that retains the most attention in these surveys, at least in the media, is non-response error in the NHS. A significant total non-response rate is expected in the survey because of its voluntary nature. To minimise its effect, a non-response follow-up (NRFU) is to be done for a subsample of 1.1 million non-responding households (expected sub-sampling fraction of 41%). NRFU sampling and corresponding estimation methodologies were originally developed by Hansen and Hurwitz (1946). Multi-phase sampling theory is used to decompose the sampling and response random mechanisms (see Figure 1):

- a first phase sample *s* is first selected from the population *U*;
- response s_r and non-response s_{nr} is observed in the sample;
- an NRFU subsample is selected from the set of non-responding households s_{nr};
- and finally response $NRFU_r$ and non-response $NRFU_{nr}$ is observed in the NRFU subsample.

Hansen and Hurwitz's original setting is in fact simpler because they assume full-response to the NRFU. This assumption is the key to producing estimates free of non-response bias. It does not hold in the NHS and unbiased estimation in the survey is not possible without further assumptions on the last non-response mechanism.

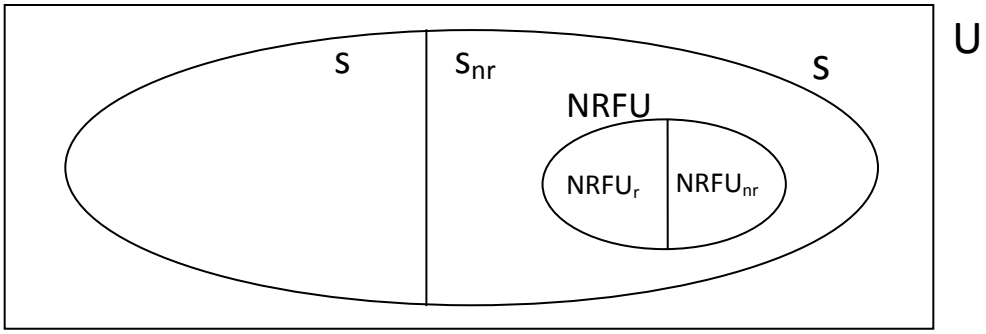


Figure 1: Schematic representation of sampling and non-response in the National Household Survey

Although there will not be full response to the NHS subsample, selecting this subsample and targeting the collection efforts on it will have the effect of reducing the overall non-response bias. This could be seen as a reallocation of resources to transform non-response error into sub-sampling error. The non-response rate to the first phase sample (before selection of the NRFU) and to the NRFU are expected to be of 37% and 78%. The resulting overall weighted

total response rate is expected to be 86% (57% unweighted). The remaining non-response bias will most likely be greater than it was in 2006 with the mandatory long form sample because the response rate was 94% at the time.

3. Handling non-response error

Handling this total non-response at the estimation stage is very challenging given its extent and the very ambitious objectives of the survey. The estimation method chosen to adjust for non-response has to minimize non-response bias because the estimates will be subject to intense scrutiny by data users. It also has to rely on as few bias assumptions as possible. It has to be simple enough to be explained easily to data users and to make sure it is implemented correctly and within the short production timeframe.

All the information from the census short forms will be copied to the NHS, providing crucial auxiliary information to minimize the effect of non-response. This also changes to some extent the total non-response into item non-response. Furthermore, tax data will be linked to the NHS sample with the sole purpose of doing a better non-response adjustment. At the end of the estimation process, calibration of the survey weights to known census counts will be done at a geographical level called Weighting Area (WA) and on 60 calibration totals in each WA. WA are made of dissemination areas and are not yet defined for 2011. In the 2006 Census there were approximately 6,600 WA in Canada with an average population size of around 1,900 households per WA. A single weight variable is to be produced for each household for operational simplicity and to avoid inconsistency between estimates. Calibration should help to effectively reduce the effect of non-response bias. Depending at what level it is done, it could also reduce sampling and sub-sampling variances. Another reason calibration could be important is that not achieving agreement between the NHS estimates and the census totals might demonstrate to the data users that correcting for biases present in the sample has not been successful.

A first possible estimation methodology to correct for non-response is to reweight the respondents to take into account the survey design and the non-response mechanisms. Weighting is the traditional way of dealing with total non-response in household surveys. It is also the estimation method taken by Hansen and Hurwitz under their setting. It is straightforward to weight for both sampling mechanisms because they are known by design. Moreover, it is sufficient to condition on response to the first phase for the first of the two response mechanisms to get the Hansen and Hurwitz estimator. However, to perform weighting for the second response mechanism you must make some extra assumptions. For example, you can estimate the probability of response to the NRFU using logistic regression and the

scores method and assume non-response occured through Poisson sampling with these predicted probabilities.

Because non-response to the NHS can also be viewed as item non-response, another way to handle the problem is through imputation. Nearest-neighbor imputation is used in the Census and the NHS to handle partial non-response using CANCEIS and could be used in the NHS to handle total non-response as well. Using this approach, the answers of non-respondents to either phases of sampling could be imputed to produce a full rectangular file somewhat similar to that produced for the 2006 Census. This will be called massive imputation in the presentation. Alternatively, imputation could be done for non-respondents to the NRFU only, while weighting would take care of the remaining sampling and non-response phases as in Hansen and Hurwitz's setting.

The pure weighting approach eliminates the non-respondents' micro-level information and does not create synthetic information, while massive imputation does the reverse and NRFU imputation both discards and creates information. Furthermore, more theoretical work has been done to support the weighting approach and massive imputation could ignore the Hansen and Hurwitz multi-phase sampling theory depending on the way it is done. However, an advantage of imputation over weighting is using CANCEIS is somewhat easier to implement than weighting in the NHS because it is the traditional way of dealing with partial non-response in the census and it is already part of the complicated estimation process. Another advantage of imputation is that keeping more micro-level census information also gives more room to calibrate to census totals.

4. Simulation set-up

A simulation study to compare the ability to correct for non-response bias of these three estimation methodologies was done using data from the 2006 Census 20% sample. Due to time constraints, the simulation was restricted to the Census Metropolitan Areas (CMA) of Montreal, Toronto, Winnipeg and Vancouver. It aimed at replicating the expected sampling and non-response conditions of the NHS. Simulating the first phase sampling of the NHS would have required selecting 30% of the population, but the 2006 Census had a 20% sample only. However, this should not matter in the comparisons since the primary goal of the simulation is to measure non-response bias. Non-response prior to the selection of the NRFU sample was simulated by blanking out the non short form data for the 63% of the 2006 Census long form households that responded the latest. From these non-responding households, a 41% stratified random sample was selected. The population was stratified by collection units[6] and allocated

[6] **There are approximately 50,000 collection units in Canada.**

proportionally to the number of non-responding households in the strata. Finally, the 78% of the non-responding households to the first phase sample who responded first in 2006 had their non short form responses restored (but only if they were in the NRFU subsample). This was done independently by CMA so that the sampling and non-response rates were fixed in each studied CMA.

To quantify the non-response and imputation errors, the estimators obtained with the above methods are compared to two estimators free of the non-response error. The first of these estimators is the one obtained with the full 2006 first phase sample. The second is the Hansen and Hurwitz estimator obtained when full response to the NRFU is observed. These comparisons are done for 84 (non short form) NHS characteristics. The ability of each method to allow calibration to the census constraints will also be presented. Results will be given as part of the presentation.

5. References

Hansen, M. H., Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Measuring Interviewer Effects on Survey Error in SHARE.

Annelies Blom[1], Julie Korbmacher[2] and Ulrich Krieger[2]

[1] Survex - Survey Methods Consulting, Mannheim, Germany (ablom@staff.uni-mannheim.de)

[2] Mannheim Research Institute for the Economics of Aging, University of Mannheim, Germany

In all interviewer-mediated surveys interviewers play a crucial role during the entire data collection process. They make contact with gain cooperation from the sample unit, ask survey questions, conduct measurement, record answers and measures, and maintain respondents' motivation throughout the interview (Schaeffer et al. 2010). To reduce variation in the data collection process, surveys are conducted through standardized interviews. But even in highly standardised surveys, interviewer effects can be found regarding different aspects of the data collection process. This holds for the unit nonresponse process, where some interviewers are more successful at obtaining contact and cooperation than others, as well as for item nonresponse and other measurement errors.

To learn more about these interviewer effects one needs detailed information about the interviewer which is not available in the majority of surveys. For that reason we designed an interviewer survey which was implemented in the German *Survey of Health, Ageing and Retirement in Europe* (SHARE). The survey is based on a conceptual framework which identifies four sets of interviewer characteristics that might explain interviewers' differential appearance and actions:

- General interviewers' attitudes that might shape the way interviewers approach sample units and ask their respondents for sensitive information.
- Interviewers' behaviour and hypothetical behaviour when faced with survey requests or similar measurements.

- Interviewers' experience with measurements, for example, experience with conducting specific surveys or the collection of specific measurements like biomarkers or consent to record linkage.
- Interviewers' expectations about the unit response, consent and item response rates they will achieve on a given survey.

Table 1: Conceptual framework of interviewer questionnaire

	Unit nonresponse	Unit nonresponse (incentives)	Consent to biomarker collection	Consent to record linkage	Item nonresponse (income)
General attitudes	Q3: reasons for being an interviewer Q5: how to achieve response Q6, Q11, Q12: trust, data protection concerns		Q6, Q11, Q12: trust, data protection concerns	Q6, Q11, Q12: trust, data protection concerns	Q6, Q11, Q12: trust, data protection concerns
Own behavior	Q8, Q9: own survey participation Q27: use of internet social networks/online banking	Q10: incentives received	Q22: consent to biomarkers, hypothetical Q24: blood donation	Q13: data disclosure, hypothetical Q14, Q16: data linkage, hypothetical Q17:“Kontenklärung” Q27: use of internet social networks/online banking	Q27: use of internet social networks/online banking Q34: income response
Experience with measurements	Q4: conducting standardized interviews Q18: SHARE experience	Q18: SHARE experience	Q23: bloodspots		Q18: SHARE experience
Expectations	Q19: effect of incentives on unit response	Q19: effect of incentives on unit response	Q21: consent to biomarker	Q15: consent to data linkage	Q20: income response

All German SHARE interviewers were asked to fill in a paper-and-pencil questionnaire during the interviewer training prior to the fourth wave of SHARE. We received a response rate of 83% which is a good starting point for our analysis.

This paper builds upon interviewer effects found in SHARE and presents the conceptual framework of the new interviewer questionnaire developed to explain these effects. In the future we will link this data directly with the survey data each interviewer collected so that we can compare interviewers' characteristics with their outcome in SHARE. Since SHARE data collection started only some weeks ago, the SHARE data is not yet available.

However, first analyses of the interviewer survey show interesting associations between the response and consent rates the interviewer expect to achieve in SHARE and other dimensions in the interviewer characteristics collected.

Item Nonresponse in a Mail Survey of Young Adults

Authors:

1) Luciano Viera, Jr. (lviera@forsmarshgroup.com; 1-703-696-9439)
2) Scott R. Turner; (sturner@forsmarshgroup.com)
3) Sean M. Marsh (smarsh@forsmarshgroup.com)

Affiliation (all authors): Fors Marsh Group

Abstract:

Item nonresponse occurs when a respondent provides answers to some questionnaire items but fails to do so for others and it may occur because of apathy, confusion, or a desire to protect one's privacy (Wolfe, Converse, Airen & Bodenhorn, 2009). Item nonresponse threatens the quality of survey metrics, particularly when the nature of missing data is not random. Previous research has shown that item nonresponse has been linked to features of the questionnaire design, such as requesting personal/sensitive information and item formatting (Brener, Kann, & McManus, 2003; Gruskin, Geiger, Gordon, & Ackerson, 2001; Healey, 2007; Johanson, Gips, & Rich, 1993; Messmer & Seymour, 1982; Smyth, Dillman, Christian, & Stern, 2006; Wolfe, 2003). However, this existing work is still far too generalized to explain many unique, situation-specific cases of item nonresponse.

This paper presents the results of item nonresponse analyses conducted on data collected from a semi-annual national mail survey tracking the future career plans of young adults. In previous years, this survey was conducted using an interviewer-administered, random-digit-dial (RDD) telephone methodology but was switched to a mail-based methodology primarily to combat the trend of declining coverage of cell phone households, particularly among young adults (Blumberg & Luke, 2010). One of the key metrics tracked in this survey is general military propensity, defined as the proportion of youth responding that they will definitely or probably be serving in the U.S. Military in the next few years. In addition, the survey also tracks Service-specific propensity, defined as the proportion of youth responding that they will definitely or probably be serving in each of the twelve (12) U.S. Military Services in the next few years. In the paper-based questionnaire, respondents are first presented with the general military propensity item and then asked the Service-specific propensity items presented in a matrix-style format.

Because nonresponse for the Service-specific items were higher than found in the previous RDD administrations, the existing data were examined to determine whether there was a pattern to the missing data and how refusals should be handled (e.g., include/exclude from denominator, imputation, etc.). Also, the Service-specific propensity questionnaire item was modified in the next administration to include a note at the end of the item reminding youth to provide a response for each of the Services. Therefore, analyses were also performed on subsequent data to determine whether this item revision had a meaningful impact on item nonresponse.

Overall, results show that the pattern of nonresponse in the Service-specific propensity items appear to be nonrandom as item refusals are related to a lack of propensity for the Military in general or a preference for a specific Service(s). Moreover, the modification made to Service-specific propensity item significantly reduced item nonresponse among the Service-specific propensity items. Implications for existing survey practice and directions for future research will be discussed.

Computer Audio Recording: A Practical Technology for Managing Survey Quality

M. Rita Thissen, Hyunjoo Park and Mai Nguyen, RTI International

Intro

Of the many potential contributors to total error in interviewer-administered surveys, several types of structural and functional errors lend themselves to straightforward management through the use of audio recording technology. Among these, computer audio-recorded interviewing (CARI) may assist with addressing

- Structural errors which happen systematically, including faulty question wording, translation concerns, or logical faults within an electronic instrument
- Functional errors which happen in the process of survey operation, including data-collection mode effects, inconsistency of interviewers' questionnaire presentation, failure to gather complete information during open-ended response, data entry mistakes, classification errors, and item non-response

Each of these types of error is amenable to detection, quantification, intervention and control through astute use of audio recording, perhaps aided by screen-image capture during the interview. Subsequent playback of the audio recordings with associated response data allows behavior coding, quality assurance coding, authenticity evaluation, assessment of interviewer performance and input into data editing, providing paradata on which to base tailored interventions to reduce total error.

The CARI Interactive Data Access System is software for review and coding of CARI recordings at the US Census Bureau (Thissen et al, 2010). With development begun in 2009, the web-based coding system offers a multi-functional interface for behavior coding, data quality review, authenticity review and interviewer performance management, which allows survey managers to detect various types of error and to improve quality in both field and telephone surveys. Our presentation describes system design as well as user feedback from a field test of the software.

Survey Error

Theoretically, total survey error consists of all variable errors and all biases, that is, the sum of variance and squared bias. As part of variance and bias, there are two types of errors: non-observational errors and observational errors. Non-observational errors arise because measurements were not taken for part of the population. For example, sampling error is the most familiar type of non-observational error, and this errors stems from measuring only a subset of the population, not the entire population. Observational errors arise because the answers of respondents deviate from their true values on the measure (Groves, 1989).

In this presentation, we address observational error only, and how to detect it, reduce it and control it via CARI technology.

Generally speaking, survey error is the difference between the truth and the survey's results. Over the process of a survey, there are many ways that survey statistics based on respondents' answers depart from the truth. Using the simplified summary of survey process steps shown in Figure 1, we examine the ways in which error introduced at a particular step may be mitigated through the use of CARI.

1. Define research objectives
2. Choose methods of data collection
3. Detailed planning, including quality assurance plan
4. Construct and pretest questionnaire
5. Collect data
6. Analyze and report

Figure 1. Survey process steps

Structural Errors

CARI can help reduce structural errors those that take place systematically due to faults embedded in the questionnaire specification or in its computerization process. Such systematic errors may be limited to a subpopulation of respondents, as mentioned before with respect to wording. The phrase "drinking fountain" is not used in Milwaukee, for example; interviewers in that city assessing child vocabulary as part of a national survey objected when they were required to mark a response incorrect if a child identified a picture as a "bubbler," the common term in that region. This type of confusion may be preserved for review through CARI technology.

Similarly, translation errors may be addressable through CARI. In one survey, "For whom did you work?" was translated to "What's the name of your company or employer?" in Korean, an attempt to offer a natural-sounding translation. However, this caused respondent confusion about whether to write the restaurant name or the owner of the restaurant, if the person was working at a restaurant.

Generally, software testers are able to confirm instrumentation logic through exercise of all paths through each branch point, referred to as a gateway question. However, when gateway logic becomes extremely complex, and especially when response options are dynamically determined by earlier responses, some untested paths may remain. In order to use CARI as a way to confirm logic, the gateway question needs to be recorded along with one or two subsequent items, allowing reviewers to determine whether the follow-up questions were appropriately presented; recording a time-slice may be useful in this situation, capturing the gateway question and whatever follows it.

Functional Errors

CARI can also help reduce functional errors, those due to human errors such as respondent behaviors, Interviewer behaviors, and interactions between those individuals. These errors may include data-collection mode effects, inconsistency of interviewers' questionnaire presentation, failure to gather complete information during open-ended response, data entry mistakes, classification errors, and item non-response.

Mode effects typically are discovered during data analysis, when substantial numbers of responses are available from each of the survey modes. However, on an informal basis, CARI review can provide advance warning of certain types of mode effects, such as respondent reluctance to provide highly personal information during in-person interviews, or interviewers hurrying through questions on the telephone to prevent the respondent from breaking off the interview. More formal comparison of CATI and CAPI data can take place after data collection has progressed far enough to provide enough data for credible analysis.

Innocuous-seeming inconsistencies in presentation of the survey questions may elicit remarkably different responses. For example, when the actual questionnaire text ""Where do you get most of your news about current events in this country – from the radio, the newspapers, TV, or talking to people?" was shortened by the interviewer to "Where do you get most of your news about current events?", respondents who used different sources of information for international events provided divergent answers. Similarly, the omission of the single word "usually" changes the meaning of "By what form of transportation did you usually travel to work?" from a long-term description to one interpreted as referring to the present day only.

When problems are known to exist with data entry by interviewers of lengthy open-ended responses, CARI may be a means not only of detecting the problems but of solving them as well. In one study, audio recordings simply replaced interviewer data entry efforts, after it was determined that the quality of the field data entry was unacceptably low due to difficulty capturing the response (Edwards et al, 2010).

Items with awkward, lengthy or complex wording, especially among the response options, may not work well with some respondents, whether from cognitive fatigue, impatience, or failure to understand distinctions among the many choices. Error rates are high in this situation, in part

due to incomplete or misclassified responses and in part from mistakes or short-cutting by interviewers trying to record complex answers (Mitchell et al, 2008).

Even the thorny issue of item non-response can be examined through CARI, if recordings of refusals and "don't know" answers are collected for review and coding. In situations where the respondent asks for clarification or gives any indication of reluctance to answer, that information may be useful in training interviewers.

CARI System at the US Census Bureau

A full system for employing CARI is a set of software tools consisting of two main components: (1) Recording software to collect audio recordings and screen images of the interviewers' and respondents' vocal exchanges during an interview; (2) CARI Monitoring System to allow researchers, supervisors and QA staff to listen to the recordings at a later time for evaluation of respondent behavior, questionnaire design, interview authenticity, data quality assurance and interviewer's performance.

The US Census Bureau has gradually implemented CARI in a series of research and development projects begun in 1999 and continuing at least through 2012.

In the current implementation at the Census Bureau, audio recordings and screen images are captured by Blaise software on the laptop of a field interviewer or the desktop of a telephone interviewer. Regardless of where the data originates, the files are transmitted at least daily to a centrally located master control system (MCS) and then loaded into the CARI Interactive Data Access System. A demonstration and detailed discussion of the monitoring system will be available during the poster session of this workshop.

Feedback from CARI System Users

The American Community Survey (ACS) Content Test was conducted by the US Census Bureau in late 2010, starting as a mail survey, and then followed by CATI and CAPI interviewing. Data collection in CATI and CAPI modes was conducted from October-November 2010, and CARI recordings were collected in both modes (Pascale, 2011). This survey was the

first to use the CARI Interactive Data Access System, and only the behavior coding module of the system was used.

Features that the CARI System users found most valuable included the following

- Flexibility in defining what behaviors to code
- Image display offering the exact display of wording (including fills) as well as the actual data entry value
- Real time monitoring of coding quality through inter-rater reliability tests
- Data available for extraction at any time

In general the field test demonstrated the value of the approach, offered a proof of concept for operational use in behavior coding, and allowed researchers unfamiliar with the approach to gain hands-on experience. Results of the actual research will be tabulated, analyzed and serve to improve the questionnaire design of the full ACS in the future.

At ITSEW, we will present additional details of the use and evaluation CARI System by the ACS Office of the Census Bureau.

Acknowledgements

The authors would like to express thanks to members of the Census Bureau CARI team, especially Sherry Thorpe and Joanne Pascale, and to the members of the RTI CARI team, especially Carl Fisher and Chris Siege.

This work was conducted under Census Bureau contract 50-YABC-2-66053 Task Order 16. Opinions expressed are those of the authors, not necessarily those of the Census Bureau or the Census Bureau's CARI team.

References

Edwards, Brad, W. Hicks and Michelle Carlson (2010). Old Dogs New Tricks: CARI and CASI Innovations to Reduce Measurement Error and Nonresponse, Presented at the International Total Survey Error Workshop, Stowe, VT

Groves, Robert M. (1989) *Survey Errors and Survey Costs*, Wiley & Sons, NY, pp. 8-12

Mitchell, S.B., K.M. Fahrney, M.M. Strobl, M.R. Thissen, M.T. Nguyen, B.S. Bibb, and W.I. Stephenson (2008). “Using Computer Audio-Recorded Interviewing to Assess Interviewer Coding Error,” Presented at the 63rd Conference of the American Association for Public Opinion Research (AAPOR). New Orleans, LA

Pascale, Joanne (2011) Using Behavior Coding to Evaluate Questionnaires, Presented at FedCASIC Workshop, Washington, DC

Thissen, M.R., Thorpe, S., Peytcheva, E., Barnes, C., Park, H., & Fisher, C. (2010). Improving Data Quality Through Audio and Image Recording. Poster presented at International Total Survey Error Workshop, Stowe, VT

Contact Information

M. Rita Thissen

Research Computing Division

rthissen@rti.org

+1 919 485 7728

Hyunjoo Park

Survey Research Division

mpark@rti.org

+1 301 468 8287

Mai Nguyen

Research Computing Division

mnguyen@rti.org

+1 919 541 8757

RTI International

PO Box 12194

Research Triangle Park, NC 27709

Proxy Pattern-Mixture Analysis of Missing Health Expenditure Variables in the Medical Expenditure Panel Survey

Robert M. Baskin, Samuel H. Zuvekas, and Trena M. Ezzati-Rice

Agency for Healthcare Research and Quality (AHRQ)

540 Gaither Road, Rockville, MD 20850 USA

Robert.baskin@ahrq.gov (Primary Contact Author, 301-427-1669)

Trena.ezzati-rice@ahrq.gov (Coauthor)

Samuel.Zuvekas@ahrq.gov (Coauthor)

Proxy Pattern-Mixture Analysis (PPMA), developed by Andridge and Little, is a method for assessing non-response bias in the mean of a survey variable Y subject to non-response, when there is a set of covariates observed for both the respondents and non-respondents. The covariates are reduced to a single proxy variable X that has the highest correlation with Y (the predicted value from a linear regression). This relation is estimated from a regression analysis of the respondents and is used to measure the impact of non-response. The impact of the non-response then depends on the response rate, the strength of the proxy variable in predicting Y, and the difference in proxy means for respondents and non-respondents. The PPMA is a likelihood based method that proceeds by factoring the likelihood of the variable Y, covariate X, Missingness indicator M, and parameter set (θ,π) as

$f(Y,X,M \mid \theta,\pi) = f(Y,X \mid M,\theta) f(M \mid \pi)$ but assumes that $f(M \mid Y,X,\theta,\pi)$ is a monotonic function of $\lambda Y + X$ for some unknown and non-estimable index λ. Note that this is an assumption on the form of missingness with λ=0 corresponding to the usual missing-at-random assumption (MAR) and λ>0 a form of missing-not-at-random (MNAR). Since λ is not known in practice the non-response bias must be estimated over a range of λ values to get an idea of the magnitude.

However, it is claimed by Andridge and Little, supported by simulation, that the method of fraction of missing information (FMI), obtained through multiple imputation as the between variance over the total variance, can be used under the PPMA model even if the missingness mechanism is MNAR and not MAR. In this paper, explicit estimates under PPMA models as

well as the FMI are used to assess and compare the impact of missingness among selected variable level statistics, in particular health care expenditure variables, in the Medical Expenditure Panel Survey (MEPS). The MEPS is a nationally representative survey of the U.S. civilian noninstitutionalized population. The set of households selected for each year's panel of the MEPS is a subsample of households participating in the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. The FMI is applied to the item missingness in MEPS and this provides the richest set of covariates for determining the proxy model. MEPS uses imputation for item non-response on the expenditure variables and has traditionally used weighted sequential hotdeck as the imputation method. However, imputation for expenditures in some types of medical events have recently used a predictive mean matching approach that allow greatly improved regression models with an expanded number of covariates as well as continuous covariates. These regression models can use log or square-root transformed expenditures for prediction of closest match without the ensuing problems of back-transforming the predictions. PPMA and FMI are used to evaluate the level of bias in the mean estimate under the predictive mean matching imputation.

There are four types of models used in the evaluation which were used in the internal testing of the imputation methodology. The basic model includes the predictors used in the weighted sequential hotdeck without any transformation of the expenditure variable. There is an expanded model that uses all of the predictors from the weighted sequential hotdeck as well as indicators of chronic conditions (e.g. diabetes or asthma) and GPCI codes (geographic payment code indicators from Medicare). The expanded model was run with non-transformed payments, log transformed payments, and square root of payments. For the event type inpatient hospital events and for the event type office based physician visits the R-squared values are given in the following table:

Model Type	Hospital Inpatient Stay Expenditures	Office-Based Physician Visit Expenditures
Basic (no transform)	0.54	0.61
Expanded (no transform)	0.56	0.62
Expanded (log transform)	0.61	0.20
Expanded (square root)	0.60	0.66

Using the PPMA model, under the assumption that the missingness is a monotonic function of $\lambda Y + X$, the non-response bias is estimated, based on maximum likelihood estimation, to be $\overline{Y} - \overline{Y}_{respondents} = \frac{\lambda + \rho}{\lambda\rho + 1}(\overline{X}_{all} - \overline{X}_{respondents})$. For the two types of events in the previous table

and using the ρ from square root transformation the estimated bias for three values of λ (0, 1, ∞)is given by:

	Hospital Inpatient Stays	Office-Based Physician Visits
λ=0	0.13%	0.01%
λ=1	0.15%	0.13%
λ=∞	2.5%	2.9%

Multiple imputation with predictive mean matching was carried out using the mice package in R. Under multiple imputation the FMI is the between imputation variance divided by the total imputation variance. The unadjusted FMI for hospital inpatient stays is 17% and adjusting the multiple imputation for unequal sampling weights gives an FMI of 11%.

Question for Discussion: Starting with the statement that PPMA using maximum likelihood underestimates the bias due to non-response in the mean of responders and FMI possibly overestimates the variance due to imputation in the imputed mean, how is this result interpreted?

An Assessment of the Impact of Two Distinct Survey Design Modifications on Health Care Utilization Estimates in the Medical Expenditure Panel Survey

Steven B. Cohen, Trena M. Ezzati-Rice, Marc Zodet, Center for Financing, Access and Cost Trends (CFACT), Agency for Healthcare Research and Quality

National health care utilization estimates for the overall population and specific population subgroups are critical to policymakers and others concerned with access to medical care and the system's use, cost and sources of payment for that care. The Medical Expenditure Panel Survey (MEPS) is one of the core health care surveys in the United States that serves as a primary source for these essential national health insurance, health care utilization and expenditure estimates. The survey is designed to provide annual national estimates of the health care use, medical expenditures, sources of payment and insurance coverage for the U.S. civilian non-institutionalized population. More specifically, the MEPS permits national estimates of annual health care utilization patterns for the U.S. civilian non-institutionalized population, further distinguished by the following types of services: office based care, outpatient visits, inpatient hospitalization stays, emergency room visits, dental visits, prescribed medicine purchases, home health care visits and purchases of other medical equipment and supplies. The longitudinal design of the MEPS permits the derivation of both annual health care utilization estimates and estimates that cover two consecutive calendar years. The survey is also characterized by an integrated survey design linked to the National Health Interview Survey (NHIS), which facilitates the derivation of national estimates for extended periods of time and yields enhancements to the conduct of longitudinal analyses.

Design modifications to ongoing national survey efforts are often implemented to improve the quality of one or more survey features and outputs. Sample design alterations have the capacity to improve the precision of survey estimates at reduced costs, while modifications to the survey instrument and editing procedures may yield visible improvements in the quality and reliability of resultant estimates. Enhancements to the information technology (IT) and data processing components of the survey enterprise offer the promise of net gains in timeliness, efficiency and accuracy. In addition, revisions to the post-survey estimation procedures, which permit greater specificity in the application of nonresponse, post-stratification and raking adjustments, may also result in visible improvements to the accuracy of survey estimates.

In 2007, the MEPS experienced two dominant survey design modifications: (1) a new sample design attributable to the sample redesign of the National Health Interview Survey, which serves as the sample frame for the MEPS and (2) an upgrade to the Computer Assisted Personal Interview (CAPI) platform for the survey instrument, moving from a DOS to a Windows based environment. The change in the NHIS sample design offered the following potential improvements to the design of the MEPS: (1) improved coverage of the population based on more current address listings on the sample frame; (2) greater capacity to oversample minorities (Asians, in addition to Hispanics and African-Americans) based on their targeted oversampling in the NHIS; and (3) improved precision in survey estimates based on the greater dispersion of the NHIS sample across the nation. Similarly, the upgrade to a Windows-based platform for the

MEPS CAPI was implemented based on the following expectations: (1) greater flexibility in the selection of new laptops with enhanced memory and processing speed for use by the survey interviewers; (2) greater flexibility in operationalizing survey instrument design modifications; and (3) enhanced capacity to implement "real-time" data editing quality control checks in the interview administration.

This study examines the impact of these recent MEPS design modifications on resultant national estimates of health care utilization. To ensure a comprehensive investigation, this research effort examines several dimensions of the potential impact of the MEPS survey design modifications. The overlapping panel design of the MEPS survey and its longitudinal features are particularly well suited to inform these analyses. The first arm of the study examines the alignment of MEPS health care utilization estimates across panels within calendar year, controlling for design features. This is supplemented by a model-based analysis of the impact of design modifications on MEPS utilization estimates, controlling for pre-dispositional factors associated with health care utilization. Furthermore, the linkage of the MEPS to the NHIS permits a related set of analyses to discern the impact of the MEPS sample redesign initiated in 2007 and associated survey attrition on national estimates. Using prior year NHIS data in concert with the restricted sample of MEPS respondents, attention is also given to an evaluation of the alignment of NHIS health care utilization estimates derived using the MEPS estimation weights, relative to those obtained from the full sample NHIS. The paper concludes with a discussion of the strategies that have been implemented and those under consideration that may yield additional improvements in the accuracy of critical policy relevant survey estimates obtained from the MEPS.

The first arm of the study examined the alignment in MEPS health care utilization estimates across panels, controlling for design features. Based on the findings from the panel specific comparisons of the calendar year MEPS utilization estimates expressed as both population means and population totals, there was some evidence of differentials in estimates attributable to the joint effects of the 2007 MEPS survey redesign. With some exceptions, when differentials were detected, the estimates generated by the new panel experiencing the MEPS survey redesign in 2007 were consistently lower. Alternatively, when focusing attention on comparing the calendar year health care utilization related estimates expressed as totals across panels for years prior to the MEPS survey redesign (2002-2006), other than for prescribed medicines no statistically significant differentials in estimates were detected both overall and when distinguished separately by age for children and adults. By pooling the two panel specific estimates in MEPS, any extant effects attributable to the survey redesign are mitigated.

These descriptive analyses were supplemented by a model-based analysis of the impact of recent MEPS design modifications on healthcare utilization estimates, controlling for pre-dispositional factors associated with healthcare use. When testing for the joint influence of 2007 MEPS survey design modifications on health care utilization estimates, the results of the regression analyses varied by the type of health care service measure under consideration. While no significant effect for MEPS Panel classification was detected in distinguishing the level of health care service utilization for annual estimates of outpatient visits, ER visits, inpatient stays or prescribed medicine purchases, a panel effect was detected for the predictions of annual office based visits and dental visits. When significant differentials were operational in 2007, higher model based utilization estimates were associated with the older panel. The results provided

important data to illustrate the level of impact of the recent MEPS design modifications have had on resultant calendar year health care utilization estimates and related model based studies.

The final series of analyses attempted to isolate the effects of MEPS sample design modifications and adjustments for survey attrition on utilization estimates from those attributable to the CAPI design modifications introduced in 2007. The survey integration between the MEPS and the NHIS facilitated this type of investigation. For each of the individual MEPS Panels operational in 2007, national estimates of prior year NHIS health care utilization related measures were derived from the MEPS first part of year and annual responding samples, and compared with those obtained from the full NHIS in 2006. A review of the results of the MEPS and NHIS generated estimates based on the same NHIS measures revealed only modest differences in estimates. Taken in concert, these findings are indicative of the level of stability in utilization related estimates attributable to sample design modifications and the effectiveness of the MEPS nonresponse adjustments.

To the extent the source(s) of the observed differences in estimates can be attributed to specific survey design differentials, estimation strategies could be developed to bridge the redesign-based estimates with those of the original design for analyses of trends over time. In this study, a set of options are presented for consideration when attempting to aligning redesign-based estimates with the original design to enhance analyses of trends over time. One of these approaches employs direct standardization technique, where the overall estimated utilization totals derived from the new panel are "aligned" to converge with the national utilization estimates derived from the old panel via adjustments to the survey estimation weights. The analysis of the impact this bridging strategy on the overall MEPS estimates of utilization totals by event type for calendar year 2007 revealed a modest, but non-significant increase in the total number of office based provider visits.

Since 2008, each of the overlapping panels in MEPS has been operating under the same CAPI platform and sample design. Future research efforts will be directed to determining whether the alignment of panel specific estimates return to their historic patterns of concordance. Additional studies will focus on a comparison of estimates of transitions in health care utilization patterns over a two year period, examining the estimates of transitions in health care use observed between NHIS and MEPS in relation to those obtained entirely from MEPS for the same time period, prior to and following the MEPS survey redesign modifications. Both the linked design of the MEPS and its overlapping panel design features will help facilitate these benchmarking efforts. Additional attention is also being given to a review of the modifications in the CAPI programming specifications and data editing rules associated with the health care utilization measures that characterize the MEPS survey redesign. Findings from these additional and on-going investigations may result in future enhancements to the MEPS survey design, operational and estimation procedures that yield can additional gains in accuracy for the critical policy relevant survey estimates from MEPS.

BALANCING CONFIDENTIALITY AND QUALITY IN PUBLIC HEALTH DATA

Lawrence H. Cox, Ph.D.
National Institute of Statistical Sciences
cox@niss.org

Preserving the confidentiality of data pertaining to individuals released in aggregate form is one component of data quality. Conversely, perturbation-based statistical disclosure limitation (SDL) methods such as rounding increase nonsampling error of aggregate data and suppression-based SDL methods degrade its utility and usability. Balancing quality and confidentiality is a core responsibility of national statistical offices (NSOs) and other organizations that release statistical data for public use.

In the United States, health care providers, including physicians, clinics and hospitals, and coroners are required to report certain health and mortality encounters and accompanying demographic information at the person level to local public health departments. Encounters include *reported cases* of communicable diseases such as measles, HIV/AIDS, and tuberculosis, and *death by cause* such as by influenza, malignancies, cardiovascular disease, pertussis, accident or homicide. In addition to name and address, demographic characteristics such as age, gender, race and ethnicity are recorded whenever available. Public health agencies at the city, county, state and national level aggregate, share and release these data for public health and research purposes. Government statistical reports in tabular form include "reported cases by race/ethnicity" or "reported cases by age group" within individual states and counties, and "death by cause" or "death by cause and gender (or age)", also often at both the state and county or local (city/postal code) level. In addition, the release of customized tabulations—often high dimensional—may be available through online statistical query systems.

Individual encounter health data are extremely confidential and, being derived from standardized, required reports, are quite accurate. Unfortunately, as we demonstrate, disclosure limitation policies and procedures applied to these data are often substandard and ineffective in comparison to proven, effective disclosure limitation methodologies commonplace official statistics literature and practice. We present examples of substandard practice that we have anonymized to disguise the releasing agency and to protect data on individuals. Using these failed examples as a basis, we illustrate methods and output tables properly protected and suitable for public release. Examples are drawn from U.S. sources. We relate this situation to the quality and usability of tabular public health data, to issues related to creating generalized methodologies and software development, and to the issue of *transparency* of statistical disclosure limitation procedures. Transparency can reduce total survey error but also can increase disclosure risk.

In the simplest terms, there are two aspects to confidentiality protection. The first is to determine and specify in quantitative terms the occurrence and extent of disclosure (*disclosure definition*). For tabular data, disclosure is typically synonymous with the presence or derivation of small cell counts: a cell count of 1 may identify an individual precisely and a count of 2 may

identify that individual to a second individual with identical characteristics. Consequently, *threshold rules* identify occurrence of disclosure with a cell value that is less than a predetermined threshold, such as 3, 5 or, less often, larger values such as 25. (We note that there are theoretical issues with threshold rules beyond the scope here.) The second aspect of confidentiality protection is to determine and apply a masking method that assures that cell values below the threshold cannot be determined or reliably inferred (*disclosure limitation*). Post-tabulation SDL methods include rounding, perturbation, suppression and an imputation-based method called controlled tabular adjustment. Pre-tabulation swapping of underlying microdata values has also been done. Unfortunately, the landscape for confidentiality protection in public health data is extremely uneven for both disclosure definition and disclosure limitation.

Disclosure rules in current use for public health data presented as frequency tables include:

- Threshold rules of 3, 5, 16 applied to all counts
- Threshold rules applied only to highest resolution cells
- Threshold rules applied only to certain outcomes (morbidity) but not others (mortality)
- Threshold rules applied only to counts for geographic areas below a population threshold
    - Thresholds vary considerably: 100,000 or 25,000 or smaller
    - Thresholds often do not account for geographic overlaps below threshold
- Threshold rules without accounting for concentrated aggregates (one cell equal to a total)
- Apply a threshold rule at one geographic level (state) but no rule at a higher level (national)

Disclosure limitation methods in current use for public health data presented as frequency tables are extremely limited and include:

- Do nothing—disclosure is laid bare
- Suppress only the small counts--often these can be reconstructed fully or in part

Suppressing only small counts often leads to disclosure even for simple two-way public health tables. When multi-dimensional data or multi-views of the data are released, disclosure is typically massive and complete—revealing numerous specific details on certain individuals. Suppression also degrades the usability of data by analysts who are unprepared or uninterested in reconstructing suppressed information. Perturbative methods such as rounding or controlled tabular adjustment would provide better protection and restore data usability. There are, however, theoretical and computational issues surrounding SDL for multi-dimensional tables that have yet to be sorted out. Software purporting to “solve the problem” at this time does not and cannot do so thoroughly and should be used with caution.

From the quality perspective, there are legitimate differences between public health and socioeconomic data analysis settings that need to be addressed. Prominent among these are zero cells. Zero cells are generally of minor importance in socioeconomic studies and can create a nuisance analytically, e.g., for log-linear models, but are very important in the public

health arena. The absence of occurrence of a disease in an area—particularly over an extended period of time—is an important public health finding upon which, e.g., disease mapping relies critically. Consequently, methods such as swapping or rounding that may convert a zero value to a nonzero one, or conversely, can be problematic. For all the reasons considered here, SDL for public health data needs to be looked at carefully from the bottom up and any resulting rules, procedures and software should be vetted and communicated in a transparent manner for the benefit of data subject and data user alike.

Total Survey Error in Disability Assessments: Measuring Physical and Cognitive Capacity in the National Health and Aging Trends Study (NHATS)

Brad Edwards and Tamara Bruce, Westat

The National Health and Aging Trends Study (NHATS) is a new U.S. survey beginning baseline data collection in 2011. Its primary focus is disability among older adults. The NHATS design draws on a comprehensive framework for conceptualizing disability – one that distinguishes among activities, physical and cognitive capacity, and accommodations made to bridge gaps between capacity and demands of tasks or activities. Data collection will be annual with continuing follow-up of participants and replenishment of the sample at regular intervals. This paper will focus on one key component of the NHATS framework --measures of physical and cognitive capacity – which are newer to national surveys and complex to administer. We begin with a brief description of the NHATS conceptual framework and the role of capacity measures. Innovations and rationale for including measures of capacity (both self-report and performance) are described and data from a validation study conducted in Spring 2010 are presented (including completion rates, results from behavior coding, and test-retest results). The effect of these measures on TSE will be reviewed. Finally, some lessons in implementation are discussed. Additional details follow.

- The NHATS framework is a blend of Nagi's widely-used model (Nagi, 1965) and the more recent language and perspective on disability from the World Health Organization's International Classification of Functioning, Health, and Disability (Freedman 2009). Measures in many surveys do not distinguish between capacity – the building blocks for activities that reflect what a person can/could do—and activities – what is actually done. Both self-report and performance measures of capacity are included in NHATS to capture physical and cognitive capacity. Assessments will be done on an annual basis to parallel collection of other components of the disability protocol.

- Where possible, measures were selected to capture a full spectrum of capacity, from high functioning to low functioning. Self-report capacity measures make use of nested high and low functioning items, an approach that was supported in the validation study (for example, only 2% of persons who were unable to bend over (low functioning) reported being able to kneel down and get back up (high functioning)). An innovation in cognitive capacity in the NHATS is inclusion of the computerized Stroop test, a measure of executive function that assesses reaction time and accuracy under conditions of interference (naming colors of symbols and colors of incongruent words – for example, the word blue shown in green). Results from the validation study indicate widespread acceptance of the computerized version, in addition 299 of 326 subjects completed the test.

- CARI and behavior coding of the interviewer-respondent interaction allow us to estimate the TSE associated with the self-report items and the cognitive tests administered in CAPI. Comparisons between the self reports and the physical assessments are useful in establishing validity.

- Implementation of performance tests of capacity in a large national survey of several thousand people is challenging. An important consideration in interpreting the results of these tests has to do with those who do not perform them. Understanding the difference among persons excluded for specific reasons, persons who do not do the tests because of concerns about safety, and persons who do not do the tests for other nonhealth or safety reasons (e.g. insufficient room to conduct the walking test) is critical. In the validation study, completion of physical performance tests was high: 89% for walking speed, 84% for chair stands, 90% for grip strength, 98% for peak air flow, 91% for balance tests. NHATS is using a combination of CAPI and a specially developed booklet to administer the physical performance tests. Cognitive tests which are conducted using CAPI pose other challenges. Lessons from the validation study that led to dropping a physical and cognitive performance test and changing the design of a filter question will be discussed.

Attrition and Selection of alteri Respondents in the pairfam[7] panel

Ulrich Krieger, SHARE MEA University of Mannheim[8]

Outline

The German family panel study is an annually conducted panel survey of individuals. In addition to the main respondents, their partners, and since wave two their parents and their children are also approached for an interview (alteri respondents). Selection and nonresponse errors of the alteri respondents will be addressed. Goal of the analysis is to gain insight on which alteri we keep and which ones we loose over the course of the panel.

For the purposes of this presentation I will concentrate on partner respondents as these are the only alteri respondents which have been observed for two waves.

Partner Survey

The partner survey within the family panel study is a 20 page PAPI questionnaire. It consists of key demographics and mirrors instruments from the main respondent interview.

Partnership is defined very broadly as those that the main respondent regards as a relationship. Thus partners are contacted inside the respondents household if they cohabitate either by the interviewer or by proxy of the main respondent or via mail if they do not cohabitate.

Sampling of the partner survey depends on the relationship status. If he or she stays in the relationship from wave one to wave two, the Alteri stays in the survey. If a relationship splits, the former partner is no longer followed in the panel.

The main respondent also serves as a gate keeper in the contact process of the alteri as alteri are only approached if she has given explicit consent to do so. The design for main respondents is monotone from wave one to wave two, while that of partners is not. They are identified to be either the same as in wave one or a different partner. If a relationship stays intact over the waves thus the same person is approached for an interview in wave 1 and wave 2.

Two processes have to be examined: The selection process of the partners in wave two and the nonresponse error / attrition of partners in wave two.

[7] ***This paper uses data from the German Family Panel (pairfam), coordinated by Josef Brüderl, Johannes Huinink, Bernhard Nauck, and Sabine Walper. Pairfam is funded as a long-term project by the German Research Foundation (DFG).***

[8] **I acknowledge the help I received by my the pairfam data team at the university of Mannheim, especially Volker Ludwig and Klaus Pforr.**

Data

Results from the partner study are shown in table 1

Wave 2	W1	W2
Anchor interview	12402	9069
with partners	7234	5408
same partner as in W1	-	4273
Consent to partner interview	5281	3882
Consent in both waves	-	3009
returned partner questionnaire	3743	2688
returned partner quest. in both waves	-	2081

Selection

Selection in wave two is again a two stage process. There is the the stability of the relationship. There may be main respondents who omit partnerships to ease the interview and prevent any contacting attempts whatsoever but this argument will not be followed here as we can think of no way to check for these instances.

Besides the relationship stability the main respondent can choose not to consent to the partner interview. He will be prompted for consent twice during the interview but can simply decline the request. On this step we do expect effects of relationship quality or relationship stressors as reported by the main respondent. Those with problematic relationships may not want to have their partners talk about it in an interview.

As a side note: Main respondents also have to provide a postal address in case they do not cohabitate or pass the questionnaire on to the partner in case they cohabitate and the partner is not present during the CAPI Interview of the main respondent. These aspects are not followed in this presentation.

Nonresponse

After selection into the partner sample partners have to deal with the interview request. Besides individual factors the survey mode plays a role in the response process. cohabitating parters did have higher response propensities than those living apart from the main respondent. Analyses on wave one showed that if the interviewer is more involved in administering and collection of the partner questionnaire response rates will be higher than for those who use the return envelope for their questionnaire.

Relationship quality will also have an effect for partners as they themselves might not want to report on problematic relationships.

Also prior interview experience from wave one should have a positive effect on cooperation rates.

Model

I am very unsure how to model the impact of all selection and attrition processes. Below a simple logit model on cooperation in wave two is show. Here only main respondents keeping the same partner over the two waves are included. dummy variables for consent in wave 1 and wave two as well as cooperation in wave two are included. Besides these information, some demographic characteristics and relationship quality measures are included in the model.

I included the model here just to illustrate that there is a strong influence of past selection and cooperation decisions.

Table 2: Logistic regression on cooperation in Wave 2 partner survey.

Variables	Model 1 OR/se	Model 2 OR/se	Model 3 OR/se
W1 Part cooperation	11.85***	11.24***	11.26***
	(1.34)	(1.28)	(1.28)
W2 Main Resp: Consent	0.91	0.94	0.95
	(0.13)	(0.13)	(0.13)
W2 Survey handed out	4.76***	4.81***	4.85***
	(0.45)	(0.46)	(0.46)
Part. fulltime empl.		1.02	1.01
		(0.11)	(0.11)
Part. parttime empl.		1.44*	1.42*
		(0.25)	(0.25)
Part. self empl.		0.64*	0.64*
		(0.13)	(0.13)
Part. Years of educ.		1.07***	1.07***
		(0.02)	(0.02)
Partner female?		1.20	1.18
		(0.12)	(0.12)
Partner born in Germany?		1.43**	1.40**
		(0.17)	(0.17)
Main: Satisf. Relationship			1.06**
			(0.02)
W2 Main: Satisf. Relationship			0.94**
			(0.02)
N	3748	3748	3748
Pseudo-R^2	0.31	0.32	0.32
BIC	3538	3534	3538

* p<0.05, ** p<0.01, *** p<0.001

Odds Ratios shown, standard Errors in Brackets

Issues

I am quite unsure how to proceed with this research question not only in terms of modelling but also conceprionally. Helpful comments are welcome.

Nonresponse Bias Correction in Telephone Surveys Using Census Geocoding: an Evaluation of Error Properties

Paul P. Biemer and Andy Peytchev

RTI International

The threat of nonresponse bias has been increasing with the precipitous decline of survey response rates, particularly in random-digit-dialed (RDD) telephone surveys. Often, researchers have only geographic information for RDD nonresponse cases. To compensate for nonresponse in landline samples, census demographic information can be appended at varying levels of geographic aggregation for both respondents and nonrespondents. The effectiveness of this approach depends on the error properties of the census geocoding (CG) process; however, to date, this process has never been thoroughly evaluated. In extreme situations, errors in the CG process can do more harm than good for survey estimates. If components of the process can be identified as more susceptible to error, then improvements can be made to enhance the process and its ability to reduce nonresponse bias. Using parameters from an RDD survey, we imposed nonresponse on a face-to-face survey with a much higher response rate in order to evaluate the error in the CG process. This approach provides a gold standard for respondents and nonrespondents, as well as for listed and unlisted telephone numbers. Preliminary findings show that although for some variables incorrect matches of unlisted numbers contribute the most to bias in the CG process, surprisingly, bias is also relatively large for correctly matched telephone numbers. This paper concludes with practical suggestions for the use of the CG method with surveys in general and with RDD surveys in particular.

Non-Consent Error, Nonresponse Error, and Measurement Error: Total Survey Error in Linked Survey and Administrative Data

Joe Sakshaug[9] and Frauke Kreuter[10]

Background

Linking survey and administrative records offers many potential benefits to survey researchers, including more research opportunities for data users and improved utility of the survey data, shorter interviews and fewer burdens for respondents, and an overall reduction in survey costs (Calderwood, 2009). A necessary prerequisite to directly linking survey and administrative records is obtaining informed consent. Studies show that linkage consent is not universal and consent rates vary widely across studies (see Kho et al., 2009, for a review). In addition, these studies find systematic differences between consenters and non-consenters based on common survey variables. A limitation of these studies is that administrative variables are not incorporated into their bias assessments. Administrative outcomes are highly important to researchers analyzing linked survey data, but no study has assessed their susceptibility to linkage consent biases. Examining differences between consenters and non-consenters based on administrative outcomes is complicated by the fact that administrative records are typically unavailable for the non-consenting portion of the respondent sample, which limits the estimation of self-selection biases.

Another important research gap is the assessment of non-consent error relative to other traditional forms of survey error. Several studies have assessed the joint impact of multiple sources of error within a single survey, with a particular focus on nonresponse and measurement errors, but have not compared them against non-consent errors. A critical question is whether the effort of getting consent (and the possible consequences of doing so, in terms of introducing non-consent bias) pays off in terms of improved data quality over asking respondents to self-report their administrative information during the survey interview (and possibly introducing measurement error)? The presumption is certainly yes, but this has not been tested.

Research questions

1. How large are non-consent biases for key administrative variables in linked survey data sets? Is there variation in consent bias across variables? How do non-consent biases compare to nonresponse biases?

2. What is the relative contribution of non-consent and measurement error bias to the overall error? Do non-consent biases for administrative variables outweigh measurement error biases for survey variables? Specifically, is it better, from a total survey error perspective, to obtain linkage consent over asking respondents to self-report their administrative information?

Data and Methods

[9] Program in Survey Methodology, Institute for Social Research, University of Michigan.
[10] Joint Program in Survey Methodology, University of Maryland; Institute for Employment Research, Nuremberg, Germany; Department of Statistics, Ludwig-Maximilians University of Munich.

In this analysis of non-consent, nonresponse, and measurement error bias, we utilize data from the first wave of the German Panel Study "Labour Market and Social Security" (PASS). In addition to survey data for responding units, this data set has rich supplementary administrative data for both respondents and nonrespondents, including respondents who did not consent to link their survey responses to their administrative records. This permits the joint estimation of non-consent, nonresponse, and measurement error for a given set of variables.

PASS is a relatively new dual-frame mixed-mode (CATI and CAPI) household panel survey for labor market, welfare state, and poverty research in Germany, conducted annually since 2006. The gross sample of PASS includes a total of 49,052 households. About half (n = 23,735) of those households were sampled from the Federal Employment Agency's (FEA) register of benefit recipients (benefit sample). The other half (n = 25,316) were selected from a commercial database of residential addresses (population sample). Both samples were selected within the same geographic clusters. The overall response rate in the first wave of the PASS survey was 26.7 percent (28.7 percent for the recipient sample and 24.7 percent for the population sample; RR1). All analyses presented in this paper are based exclusively on the CATI recipient sample. This is the subsample for which we can draw on supplementary administrative data for the assessment of non-consent and nonresponse biases, and the validation of survey responses.

All PASS respondents were asked for permission to link their survey record to their Integrated Employment Biographies (IEB) record. The IEB file is provided by the Research Data Center of the FEA and contains detailed employment and benefit data. PASS shows a fairly high consent rate – approximately 80 percent of respondents gave consent to data linkage, but the risk of bias is still present. For this assessment of non-consent bias, nonresponse bias, and measurement error bias, we do not link the survey data to the administrative data. Instead we received permission to link paradata (contact protocols and disposition codes) and the linkage consent indicator to the administrative data.

Findings

1. Non-consent bias is present for some variables
2. Overall, non-consent biases are small
3. Nonresponse and measurement error biases tend to be larger than non-consent biases, i.e., data linkage makes sense from a total survey perspective

Limitations

1. PASS response rate is low (26.7%)
2. Special population (benefit recipients)
3. Quality of administrative data is unknown

Discussion items

1. Interviewer effects
2. Non-consent bias trends over time
3. Other surveys/populations
4. Exact linkage vs. statistical matching/imputation
5. Mechanisms of linkage consent
6. Gain/loss framing (Tourangeau and Ye, 2009)

7. Confidentiality assurance (Singer, Hippler, and Schwarz, 1992)
8. Effect of consent on satisficing and response accuracy
9. Placement of consent request

Table 1. Percentage/Mean in Each Subgroup (and Standard Errors), According to Administrative Data and Survey Data

	Administrative Data				Survey Reports	
	Sample (n = 17,167)	Contacts (n = 10,717)	Respondents (n = 4,513)	Consenters (n = 3,538)	All Respondents (n = 4,513)	Consenters (n = 3,538)
Variable						
Age	39.5 (0.1)	40.3 (0.1)	39.5 (0.2)	39.3 (0.2)	39.5 (0.2)	39.2 (0.2)
Foreign	16.5 (0.4)	13.6 (0.5)	11.0 (0.8)	10.0 (0.7)	8.5 (0.6)	7.6 (0.6)
UB II	80.2 (0.3)	80.8 (0.4)	83.4 (0.6)	83.1 (0.6)	75.9 (0.7)	76.0 (0.8)
Disability	4.9 (1.7)	5.4 (0.2)	5.3 (0.3)	5.3 (0.4)	11.3 (0.5)	11.2 (0.5)
Employed	29.3 (0.4)	30.4 (0.5)	30.3 (0.8)	30.6 (0.8)	29.3 (0.8)	30.0 (0.8)
Income	799.9 (5011; 11.2)	788.8 (3234; 14.0)	728.5 (1352; 21.3)	730.2 (1070; 24.8)	1130.9 (1352; 29.7)	1124.7 (1070; 32.9)

Note: Parenthetical entries for the first five variables are standard errors; for the last variable (income), which was only collected for employed respondents, the parenthetical entries are sample sizes followed by the standard errors.

Table 2. Nonresponse, Non-Consent, and Measurement Error Bias Estimates (and Standard Errors), by Survey Statistic

	Nonresponse Bias				Measurement Bias	
	Noncontact	**Refusal**	**Total Nonresponse**	**Non-Consent**	**All Respondents**	**Consenters**
Variable						
Age	0.8 (0.1)***	-0.7 (0.1)***	4.6 (0.2)†	-0.3 (0.1)* †	-0.4 (1.4)	0.03 (0.02)
Foreign	-3.0 (0.2)***	-2.6 (0.3)***	-5.6 (0.4)*** †	-0.9 (0.2)*** †‡	-2.5 (0.3)*** ‡	-2.5 (0.3)*** ‡
UB II	0.6 (0.2)*	2.6 (0.5)***	3.2 (0.5)*** †	-0.3 (0.3) †‡	-7.5 (0.4)*** ‡	-7.1 (0.1)*** ‡
Disability	0.5 (0.1)***	-0.1 (0.3)	0.4 (0.3) †	0.01 (0.2) †‡	6.1 (0.4)*** ‡	6.0 (0.5)*** ‡
Employed	1.0 (0.3)***	-0.1 (0.5)	1.0 (0.5)	0.3 (0.4)	-1.0 (0.6)	-0.6 (0.6)
Income	-11.1 (8.6)	-60.3 (15.5)***	-71.4 (15.6)*** †	1.7 (9.5) †‡	402.4 (28.4)*** ‡	394.5 (31.4)*** ‡

Notes: Noncontact bias is computed as the difference between the contacted and full sample estimates in Table 2; refusal bias is the difference between the respondents and full sample estimates; and so on.

* < 0.05; ** 0.001<p<0.01; *** p < 0.001

† indicates that the difference between non-consent bias and total nonresponse bias is significantly significant, p < 0.05

‡ indicates that the difference between non-consent bias and measurement error bias is statistically significant, p < 0.05