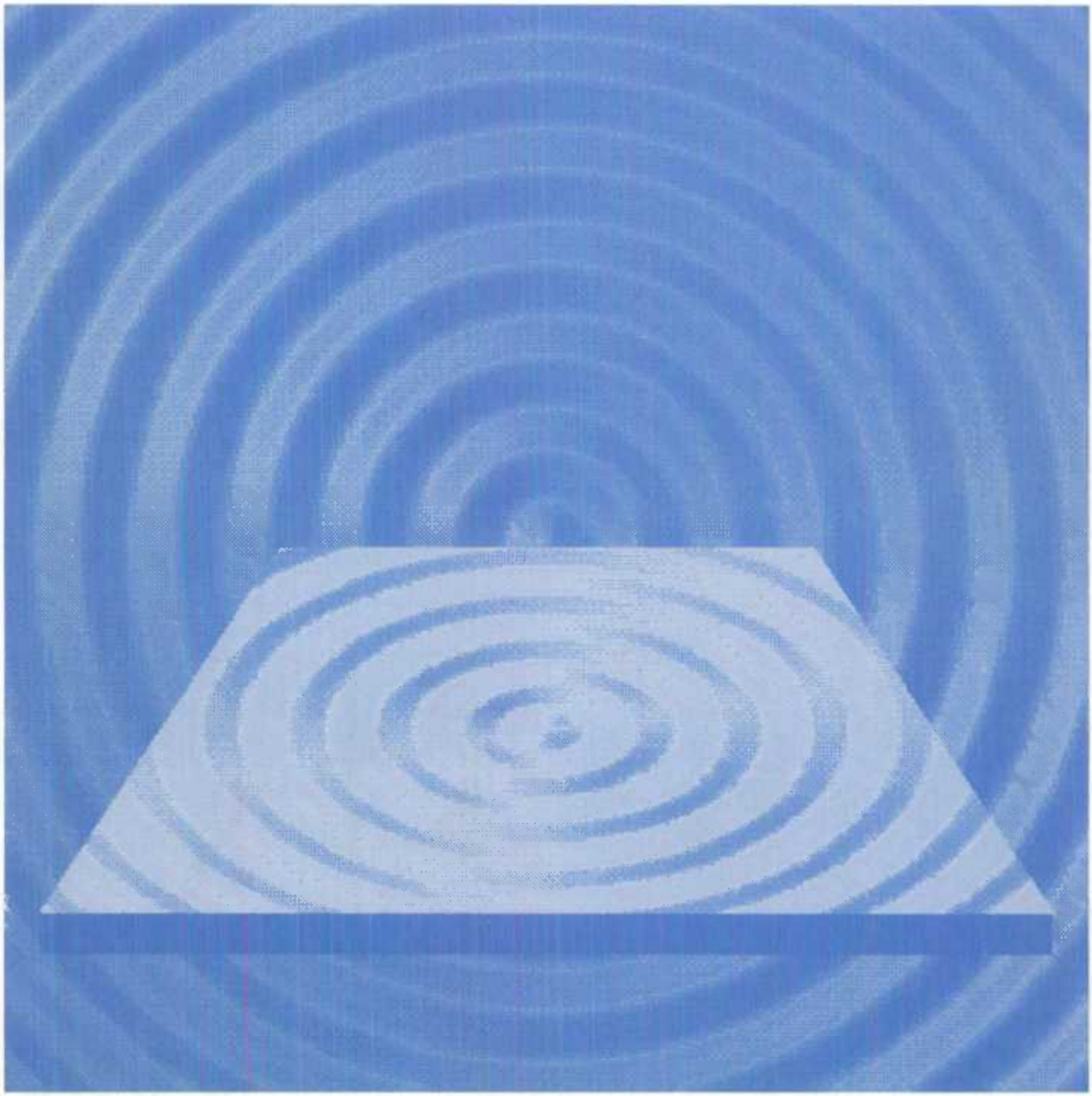
Catalogue no. 11-522-XPE

SYMPOSIUM 96

Nonsampling Errors

PROCEEDINGS



Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this publication and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-8615) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialling area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 11-522-XPE, is published in a **paper version** for $55.00 in Canada. Outside Canada the cost is US$55.00.

Please send orders to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, K1A 0T6 or by dialling **(613) 951-7277** or **1 800 700-1033**, by fax **(613) 951-1584** or **1 800 889-9734** or by Internet: order@statcan.ca. For change of address, please provide both old and new addresses. Statistics Canada publications may also be purchased from authorized agents, bookstores and local Statistics Canada offices.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Methodology Branch

SYMPOSIUM 96

Nonsampling Errors

PROCEEDINGS

Published by authority of the Minister responsible for Statistics Canada

September 1997

Catalogue no. 11-522-XPE

Frequency: Occasional

ISSN 0-660-17062-0

Ottawa

La version française de cette publication est disponible sur demande (nº 11-522-XPF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing co-operation involving Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Canadian Cataloguing in Publication Data

Symposium 96, Nonsampling errors (1996 : Ottawa, Ont.)
Symposium 96, Nonsampling errors : proceedings

Issued also in French under title: Symposium 96, Erreurs non dues à l'échantillonnage : recueil.
ISBN 0-660-17062-0
CS11-522-XPE

1. Statistics -- Congresses. 2. Statistics -- Methodology -- Congresses. 3. Statistical services -- Quality control -- Congresses. 4. Error analysis (Mathematics) -- Congresses. I. Statistique Canada. Methodology Branch. II. Title.

HA12 S92 1997 001.4'22
C97-988022-X

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences - Permanence of Paper for Printed Library Materials, ANSI Z39, 48 - 1984

PREFACE

Symposium 96 was the thirteenth in the series of international symposia on methodological issues sponsored by Statistics Canada. Each year the symposium focuses on a particular theme. This year, the theme was on nonsampling errors.

The 1996 symposium attracted over 300 people who met over three days in the Simon Goldberg Conference Centre in Ottawa to listen to experts from various statistical agencies, government agencies as well as representatives from the private industries. A total of 29 papers were presented by the invited speakers. Aside from translation and formatting, the papers submitted by the authors have been reproduced in these proceedings.

The organizers of Symposium 96 would like to acknowledge the contributions of many people involved in the preparation of this volume and those who assisted them during the symposium in November. The committee would especially like to thank Sophie Arsenault, Sophie Dionne, Guylaine Dubreuil, Hew Gough, Guido, Josée Morel, Peggy O'Neill, Micheline Sabourin and Michelle Simard for the many hours of preparing material and making arrangements for Symposium 96.

Naturally, the presenters at Symposium 96 deserve thanks for taking the time to put their ideas into written form. Publication of these proceedings also involved the efforts of many others. Processing of the manuscript was expertly handled by Lynn Savage. Proofreading was done by a number of Statistics Canada methodologists: René Boyer, Danielle Lebrasseur, Harold Mantel, Carole Morin, Claude Poirier, Michelle Simard, Jean-Louis Tambay, Charles Tardif and Alain Théberge.

Statistics Canada's fourteenth annual symposium will be held November 5 to 7, 1997 in the Ottawa region. The theme will be: New directions in surveys and censuses.

Symposium 96 Organizing Committee

Ann Brown	Jane Burgess
Johane Dufour	Éric Rancourt

STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data
1985 - Small Area Statistics
1986 - Missing Data in Surveys
1987 - Statistical Uses of Administrative Data
1988 - The Impact of High Technology on Survey Taking
1989 - Analysis of Data in Time
1990 - Measurement and Improvement of Data Quality
1991 - Spatial Issues in Statistics
1992 - Design and Analysis of Longitudinal Surveys
1993 - International Conference on Establishment Surveys
1994 - Re-engineering for Statistical Agencies
1995 - From Data to Information - Methods and Systems
1996 - Nonsampling Errors

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

Use the order form on this page to order additional copies of the proceedings of Symposium 96: Nonsampling Errors. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 96 PROCEEDINGS
STATISTICS CANADA
BUSINESS SURVEY METHODS DIVISION
R.H. COATS BUILDING, 3rd FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada" - Indicate on cheque or money order: Symposium 96 - Proceedings).

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	____ @ $10
1987 -	Les utilisations statistiques des données administratives - FRENCH	____ @ $10
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	____ @ $12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	____ @ $10
1989 -	Analysis of Data in Time - BILINGUAL	____ @ $15
1990 -	Measurement and Improvement of Data Quality - ENGLISH	____ @ $18
1990 -	Mesure et amélioration de la qualité des données - FRENCH	____ @ $18
1991 -	Spatial Issues in Statistics - ENGLISH	____ @ $20
1991 -	Questions spatiales liées aux statistiques - FRENCH	____ @ $20
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH	____ @ $22
1992 -	Conception et analyse des enquêtes longitudinales - FRENCH	____ @ $22
1993 -	International Conference on Establishment Surveys - ENGLISH (available in English only, published in U.S.A.)	____ @ $58
1994 -	Re-engineering for Statistical Agencies - ENGLISH	____ @ $53
1994 -	Restructuration pour les organismes de statistique - FRENCH	____ @ $53
1995 -	From Data to Information - Methods and Systems - ENGLISH	____ @ $53
1995 -	Des données à l'information - Méthodes et systèmes - FRENCH	____ @ $53
1996 -	Nonsampling Errors - ENGLISH	____ @ $55
1996 -	Erreurs non dues à l'échantillonnage - FRENCH	____ @ $55

PLEASE ADD THE GOODS AND SERVICES TAX (7%)
(Residents of Canada only) $____

TOTAL AMOUNT OF ORDER $____

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME _____

ADDRESS _____

_____

CITY _____ PROV/STATE _____ COUNTRY _____

POSTAL CODE _____ TELEPHONE _____ FAX _____

For more information please contact John Kovar, Telephone (613) 951-8615, Facsimile (613) 951-1462, E-mail kovar@statcan.ca

NONSAMPLING ERRORS

TABLE OF CONTENTS[1]

---

[1]In cases of joint authorship, the name of the presenter is shown **boldface**.

OPENING REMARKS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

OPENING REMARKS

G. J. Brackstone[1]

Good morning everyone. On behalf of Statistics Canada, may I offer you a very warm welcome to the 13th of our series of Methodology symposia. I hope that nobody is superstitious since we are also the 13th. You may be interested to know that we have people from many different countries: the United States, the United Kingdom, Sweden, Finland, the Netherlands and the Republic of the Philippines in addition to our usual strong contingent of Canadians.

To put this event in context, let me say a few words about the history of these symposia. The series started in 1984 and for the first four years, we alternated between a cozy little symposium in this room and a larger symposium in more spacious facilities downtown. In 1985 for example, we held an international symposium on Small area statistics in the congress center and that resulted in the book Small area statistics published by J. Wiley. In 1987, we met also downtown in the railway station for our symposium on statistical uses of administrative data. And then, for the next five years we were back here again in this room covering such subjects as the use of technologies in surveys, time series analysis, spatial data, data quality and longitudinal surveys. The 1993 annual symposium, which was held in conjunction with the International Conference on Institutions in Buffalo, gave rise to the publication of yet another book by J. Wiley. In the two years following, we returned here to discuss topics that went somewhat beyond methodology, namely, "Re-engineering for Statistical Agencies" and, last year, "From Data to Information".

This year's theme, "Nonsampling errors", lies at the very heart of survey methodology. To be sure, this is not a new topic. As early as 25 years ago, people were arguing that more attention should be paid to nonsampling error, and that we were too concerned with sampling error while the real problem lay elsewhere.

Have we made progress in our study of nonsampling errors? There is certainly a growing body of literature on nonsampling errors. Thousands of papers and reports assessing various kinds of nonsampling errors in particular surveys or censuses have been written. But how many surveys have really been designed with a thorough balancing of the likely errors from all sources? How many surveys have measures of quality that reflect more than just sampling error and maybe simple response error? If nonsampling errors are so important, aren't we misleading users by quoting sampling error alone? And have we made progress between applications and case studies in individual surveys towards a general framework or theory for the design and implementation of surveys in the face of nonsampling errors or for the measurement of overall data quality in surveys?

I think clearly the overall answer is: yes we have made considerable progress and I know we are going to learn some of that progress in the next few days. Let me say a few words about the program for the Symposium. First about its title. I think the expression nonsampling errors has two problems in my view. First by encompassing all sources of errors other than sampling, it includes a hodge podge of different error-producing mechanisms, each perhaps needing its own theoretical foundation. Unifying these

[1]Gordon J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Branch, Statistics Canada, 26-J, R.H. Coats Building, Ottawa, Ontario, K1A 0T6.

disparate error sources into a single framework is not a straightforward job. I hope as we go through the various sessions of this symposium, which tends to be organized around individual sources of errors, that we can keep in mind the need to integrate these error sources in total survey design and total quality measurement.

A second problem with the expression nonsampling errors is that it uses that favorite term of statisticians - error - with its negative connotation of mistakes. There is a sense in which sampling errors, which are clearly introduced consciously and by design, are not the result of mistakes; while nonsampling errors are seen as the result of mistakes: something in the survey was not done as it should have been. I think this is a misleading distinction. Normally, we can no more afford to avoid nonsampling errors by obtaining 100% response or by achieving zero coding error than we can afford to avoid sampling error by surveying the whole population. Trade-offs are required in survey designs to achieve an optimum quality level, taking account of all error sources, sampling and nonsampling.

The real distinguishing feature between sampling and nonsampling errors is our knowledge of and control of the error-generating mechanism. In sampling, we design and control the error-generating mechanism ourselves, while nonsampling errors are hidden from us and we struggle to understand them. That's what makes nonsampling error problems tougher to deal with and more tempting to put aside in favor of the more tractable and elegant results of sampling theory. For many of the individual sources of errors, there has been considerable progress over the years towards a better understanding of the mechanisms that generate such errors, and therefore in developing the design options for eliminating or controlling them.

That leaves the challenge of balancing the various sources of errors - sampling and nonsampling - against the costs and other constraints of government survey design. That leaves us to the topic of total survey design, about which I hope we will hear something during this symposium.

In my opening remarks to these symposia, I usually try to explain the timeliness of the selected topic. Although nonsampling error appears to be perenially interesting and worthwhile, I would like to give you a few reasons why we have chosen it for the theme of this year's symposium. First, the reputation of a statistical agency depends on the quality of the data it publishes. Its credibility is rarely challenged on account of sampling. Rather, the issue is almost always totally unrelated to sampling. Second, statistical agencies are conducting surveys on increasingly delicate and difficult topics. The primary obstacles facing such surveys are usually problems of response and non-response, not sampling. Third, longitudinal surveys which have become increasingly popular present a new set of nonsampling challenges. Finally, I think with a greater analytical use and re-use of data comes more confrontation of data with other sources and a good understanding of the full quality profile of data sets becomes crucial. These are just a few reasons for pursuing this topic.

Now, like other statistical agencies, Statistics Canada faces many challenging problems in this domain and it's good to see that so many people have come here to share their experience and expertise in helping us to resolve these problems.

I do not want to commit this symposium's first nonsampling error by exceeding my time, so let me re-iterate my welcome to all. Thank you for coming to participate in this symposium and let's look forward to an interesting and rewarding three days. I say every year that we, at Statistics Canada, benefit immensely from the presentations and exchanges that take place during these symposia. I hope that everyone will equally feel that they, and their organization have benefited from the time spent here.

OPENING KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

NONSAMPLING ERROR IN SURVEYS: THE JOURNEY TOWARD RELEVANCE IN PRACTICE

Robert M. Groves[1]

ABSTRACT

This paper offers an historical review of the conceptual structure of nonsampling errors. This review makes the following observations: a) as with all of science, progress is denoted by greater understanding of constituent components of phenomena (e.g., we now have theories about the comprehension step of respondents related to measurement errors), b) progress in inventing practical measures of nonsampling errors for routine survey use is impeded by clashes between viewpoints of experimental and observational studies (e.g., we lack a family of estimators incorporating various nonsampling error), and c) advances in understanding nonsampling errors will require use of model-assisted estimation tools now foreign to most survey analysts (e.g., most nonsampling error parameter estimates require some model of the behavioral phenomenon creating the error). The paper ends with speculations about which components of nonsampling error might yield themselves to reduction or measurement by current and future discoveries.

KEY WORDS: Nonsampling error; Measurement error models; Nonresponse adjustment.

1. OVERVIEW OF REMARKS

In my attendance at many research conferences I have encountered at least three different kinds of keynote addresses. The first attempts to provide an overview of the field that is being covered in the conference. For this conference, with its diverse topic areas, such would be a large task. The second type of keynote presents to the audience a research problem in a completely different field, but one that exhibits many of the puzzles of the field it studies. The implicit hope with this strategy is to stimulate new work, taking new perspectives. The third strategy presents a narrower development in the field of interest, attempting to prompt the audience to pursue implications of the development in their own research.

I will describe a set of experimental results in the social sciences obtained over the last few years, which together have implications for the process by which survey statisticians might approach the specification of models describing nonsampling errors. In this sense, my remarks may fit the second category of keynote addresses. In choosing this route, I hope also to communicate that the field of nonsampling errors is unlikely to be understood by a single theoretical perspective. It now seems clear that the principles underlying coverage, nonresponse, interviewer effects, question effects, mode effects, and respondent impacts on measurement error will come from a variety of social science theories. In that sense, the term nonsampling errors is increasingly too vague to communicate the sources or impacts of these errors. Indeed, the success of recent years in understanding these errors has come from dissecting the survey process into its constituents parts, as a way of isolating components that manifest different causes.

We must first note that sampling and nonsampling error research have different intellectual roots. Sampling statistics in contrast to nonsampling error research focuses much more on variance than bias. It exploits deduction from a well-elaborated base of probability theory versus hypothesis generation and experimentation. Its history is found more prominently in government statistical agencies versus the academic sector.

Theories in nonsampling error must explain human behaviors that produce the error. These theories are dominantly found in the social sciences. As such they do not have explicit links to survey estimation. This has produced a

[1] Robert M. Groves, University of Michigan and Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD 20742.

mismatch between the findings in experimental studies of nonsampling errors and their use in survey design and estimation. Advances will come when the links are made.

2. HISTORY OF RESEARCH PRODUCTS

Figure 1 provides a quick overview of the timing and product of research on sampling and nonsampling errors. For each of the major errors it lists early journal papers in the field and the first book-length treatment of the field from both statistics and the social sciences. The figure shows that many of the early work occurred in the 1930's and 1940's. Noncoverage and sampling are areas that appear to be the domain of statistics exclusively. In contrast, nonresponse, interviewer effects, question effects, and mode effects have garnered the attention of both perspectives. In general the contributions from statistics has been to promote models that describe the nature of variance properties of traditional estimators in the presence of the particular error source.

Figure 1. Early Journal Papers and Book-Length Treatments in Sub-Fields of Surveys

Noncoverage
- 1936 Stephan paper on frame problems

Sampling
- 1934 Neyman paper on allocation in stratified sampling
- 1950 Deming, *Some Theory of Sampling*

Nonresponse
- 1944 Hilgard and Payne on noncontacts
- 1946 Hansen and Hurwitz on nonresponse
- 1983 Madow *et al.*, *Incomplete Data in Sample Surveys*
- 1987 Goyder, *The Silent Minority*

Interviewer Effects
- 1929 Rice, effect of interviewer attitudes on responses
- 1946 Mahalanobis interpenetration paper
- 1954 Hyman, *Interviewing in Social Research*

Question Effects
- 1941 Rugg, experiment on wording effects
- 1951 Hansen *et al.*, response error paper
- 1951 Payne, *The Art of Asking Questions*

Mode Effects
- 1952 Larson mode effect experiment
- 1979 Groves and Kahn, *Surveys by Telephone*

Over the past twenty years research methods in nonsampling errors have undergone important changes. They have moved from observational studies to experimental designs. The motivation of the research has moved from interest in finding differences to interest in testing theories underlying the differences. In this process it has become clear that diverse theories are useful for different nonsampling errors.

Yet there remain great contrasts between the theories that underlie sampling errors and those of nonsampling errors. The theories underlying sampling error identify the circumstances under which error values are controlled to specified level. The survey designer is assured that sampling error will achieve the specified level if assumed circumstances pertain. In short the theories are directly linked to measures of error. In contrast the role of theory in nonsampling error is the identification of cause of behaviors. Many theories do not yield mathematical specifications. Hence, they identify sets of circumstances under which errors can be reduced, but not the quantitative levels to which they can be reduced. Some theories yield model-based measures of errors, but most

theories provide no links to measures of error.

Why is measurability important? Design improvements in surveys inherently involve cost-error tradeoff decisions. Since cost is eminently measurable and understood (even by MP's and congressmen), errors in surveys tend to be given less attention unless they can be measured quantitatively. The framing of design decisions is quantitative in nature, and errors that are quantified receive more attention than those not quantified. Nonsampling errors need statistical measures to get respect at the design stage.

3. CLASSES OF NONSAMPLING ERROR THEORIES THAT COULD BE INTEGRATED INTO ESTIMATION

Advances in the role of nonsampling errors in the design and analysis of survey data will require, I believe, advances in statistical models that describe their impact on survey estimates. These models cannot be useful, however, unless they incorporate the theories that describe the human behaviors that produce the errors. That is, social science theories must guide the model specification, but statistics must integrate the models into the estimation and inference process of surveys.

At this time in nonsampling error research, it appears possible to consider alteration of classical nonsampling error models to incorporate some findings from social science research into nonresponse and measurement errors. These include a) theories about direction of bias, b) theories about the causes of bias, for use in existing bias adjustment, c) theories about bias in components of error usually studied as variance properties, and d) theories about response bias *and* variance.

4. THEORIES ABOUT THE DIRECTION OF BIAS - FREQUENCY OF BEHAVIORS

It was common as early as the mid-50's to note survey errors of reporting numbers of doctor visits, shopping episodes, books read, and television programs watched. Sometimes these errors were biases, consistent overreporting, but sometimes for some items underreporting. By late 1980's we had learned much more about the causes of this phenomenon. We understand much more about the impact of closed questions on comprehension, about the tendency to use estimation with frequent, regular events, threatening overreports, and about the tendency to use counting for rare or irregular events, threatening underreports of the phenomenon in question.

For example, Table 1 shows results from Schwarz *et al.* (1985) based on a split sample experiment imbedded in a survey. Both randomly identified half-samples were asked about the frequency of their television watching using a closed question. One half-sample was given six response categories ranging from less than one half hour to over two and a half hours, with the middle category being one to one and a half hours. The second half sample was given six categories ranging from less than two and a half hours to over four and a half hours, with the middle category being three to three and one half hours. As Table 1 displays, over 84% of the sample given the low frequency scale but 62% of the higher frequency scale cases claimed less than two and hours of television watching. That is, it appears that provision of response categories with higher frequencies induces reporting of more television watching.

Table 1. Reported Number of Hours of TV Consumption by High and Low Frequency Response Categories (from Schwarz *et al.* (1985))

Low Frequency Alternatives	Percentage	High Frequency Alternatives	Percentage
< 0.5 hr	7%		
0.5 - 1.0 hr	18		
1.0 - 1.5 hr	26	<2.5 hr	62%
1.5 - 2.0 hr	15		
2.0 - 2.5 hr	18		
		2.5 - 3.0 hr	23
		3.0 - 3.5 hr	8
> 2.5 hr	16	3.5 - 4.0 hr	5
		4.0 - 4.5 hr	2
		> 4.5 hr	0
Total	100%		100%

This experiment and others like it lend support to the theory that the response formation process in survey reports of frequencies of behavior may not involve deliberate enumeration of events to provide the desired report. Instead, in many circumstances the respondent uses the response categories to make judgements. One common heuristic in valuing the response categories is the assumption that the middle category describes the central tendency of the population. With this assumption the respondents are able to report that they are above, at, or below the average in choosing the response category. They provide such a judgement as their response. Thus, the value assigned to the middle category affects their response.

This obviously has practical conclusions for questionnaire designers. For example, one might use these findings to avoid closed questions because of the implied population distribution in the response categories. But the research has also shown that with open questions about number of times, some respondents will estimate, some count (without revealing their response strategy). When respondents are asked to provide a total number of events for a rare or sporadic behavior, they will tend to count the eligible entities. The likely error in this counting process is underestimation through failure to include an episode. When respondents are asked to provide a total number of events for a common, consistently performed behavior, they will tend to estimate the number, using a variety of ways to do so. The typical error made in this response formation is overestimation, because of departures from the rule that was used to form the estimate.

How might a statistician view this situation? Let

Y_{ei} = estimated number of times for I-th person,

Y_{ci} = counted number of times for I-th person, and

Y_i = actual number of times for I-th person.

By the social science theory above, $E_t(Y_{ei}) > Y_i$ and $E_t(Y_{ci}) < Y_i$, then for $0<\theta<1$. Then

$$|\text{Bias}(\theta\, Y_{ei} + (1 - \theta)\, Y_{ci})| < |\text{Bias}(Y_{ei})|$$

$$|\text{Bias}(\theta\, Y_{ei} + (1 - \theta)\, Y_{ci})| < |\text{Bias}(Y_{ci})|$$

The practical import of this might be that to improve estimation of a mean or total on Y, a survey should ask two questions, one seeking from the respondent a rate at which the behavior occurs in a week or a month, the other seeking count of total events in reference period. With these two questions the mixed estimator above might be constructed, with more attractive bias properties than either single question estimator.

5. THEORIES ON CAUSES OF BIAS - BIAS REDUCTION THROUGH THEORY-BASED ADJUSTMENT MODELS

Traditional post survey adjustment for unit nonresponse in surveys include weighting class adjustments and propensity model adjustments. In both of these methods the abstract conditions of bias reduction can be deduced from statistical theory. Whether one can act with assurance that the conditions will be met needs some social science theory.

Let's examine the classical weighting class adjustment with ignorable nonresponse. Assume there are J adjustment cells within which participation is independent of (Y,I), where Y is the survey variable and I is the likelihood of inclusion. Here, the Horvitz-Thompson estimator of the mean might be written as

$$\sum\sum y_i \pi_i^{-1} r_j^{-1} \Big/ \sum\sum \pi_i^{-1} r_j^{-1}$$

where r_j is the response rate for the population in adjustment cell j.

In practice, r_j might be estimated by first estimating response propensity from sample data. Let X be observed for both respondents and nonrespondents and

$$R \perp\!\!\!\perp (Y,I) \mid X$$

then $r(x_i) = Pr(R_i = I \mid x_i)$.

The $r(x_i)$ can be estimated indirectly by logit models. Then J adjustment cells can be created by coarsening the estimated $r(x_i)$ into a small number of categories. When propensity models are built, we want to discover desirable X's and model specifications usable over a variety of surveys.

Research in nonresponse behavior has shown that nonresponse is multifaceted. For example, noncontact and refusal are two alternative sources of nonresponse that have very different behavioral bases. Following the logic above, let R_c refer to the probability of contact, and R_p, to the probability of participation, given contact. We want to determine if

$$R_c \perp\!\!\!\perp (Y,I) \mid X$$

but

$$R_p \perp\!\!\!\perp (Y,I) \mid Z.$$

What theories might inform the specification of the X's and Z's? These might include cognitive script theory (Abelson, 1981) to describe the process by which the intentions of the interviewer seeking a survey interview would be interpreted by the householder. They might include psycholinguistic theories of comprehension, which would inform how the meaning of words used by the interviewer are interpreted in the context of the entire conversation about the survey and the interview request. They might include sociological theories of class and race effects (Goyder, 1987).

Such theories motivate a two step adjustment model, with the first step reflecting the process of contacting sample units and the second step reflecting the process of gaining cooperation, among those contacted. The theories identify new observations to collect on respondents and nonrespondents, these involve proxy indicators of the causes of the two nonresponse phenomena. For example, the noncontact model would include as right side variables a set of at-home influences and indicators of interviewer calling patterns. The cooperation model would include influences on participation, as well as correlates of key survey variables.

6. THEORIES ON BOTH RESPONSE BIAS AND VARIANCE - EXAMPLES OF VARIABLE AND FIXED EFFECTS OF INTERVIEWERS

The statistical literature about interviewer effects on survey data contains essentially components of variance models. They model the differences that interviewers create in survey data through different ways of asking questions associated with different response errors. In contrast, the social science literature has focused on a set of fixed effects on respondent behavior from interviewer gender, race, and age. These studies have shown consistent changes in behavior of respondents on surveys dealing with topics relevant to those demographic characteristics of the interviewer.

Understanding and correctly reflecting both these variable and fixed effects of interviewers on survey data require a modification of the specification of the interviewer effect models. For example, let $y_j = \mu_j + \epsilon_j$, for each of the j-th respondents. Now, let's start to acknowledge the error induced by the interviewer:

$$y_{kj} = \mu_j + b_k + e_{jk}$$

when the j-th respondent is assigned to the k-th interviewer. The typical assumptions made in this model, mainly to ease estimation are:

1. $\{b_1, b_2, \ldots, b_k\}$ is a random sample from an infinite population with $b_k \sim$ iid $(0, \sigma^2_b)$

2. $e_{kj} \sim$ iid$(0,\sigma^2_e)$

3. μ_j , b_k , e_{kj} are uncorrelated for all k, j

Then

$$Var(y) = (1/n)(\sigma_\mu^2 + \sigma_b^2 + \sigma_e^2)[1 + (m - 1)\rho_y] = (Var(y_{ij})/n)\,[1 + (m - 1)\rho_y],$$

where ρ_y is an intraclass correlation, within interviewer workloads, of response deviations.

What theories are appropriate for use in the respecification of this model? Most are theories of interpersonal influence. They note that the comprehension of questions is socially constructed. They note that observable attributes of the interviewers are used to aid comprehension and judgement of appropriate answers. Finally, they observe that the race, age, and gender of interviewers have effects on answers when questions are relevant to those attributes.

The alteration of interviewer variance models must acknowledge that some of the variance component associated with interviewers is appropriately associated with the interviewers' demographic attributes. That is, some of the variation across interviewers is due to these fixed demographic attributes; others may be due to attributes that are better modeled as random components. This perspective forces attention to the appropriate inferential population for the survey. For example, do we want to limit inference to essential survey conditions in which there is no variation on the interviewer attribute in question? That is, do we want to measure interviewer variability with the given mix of interviewers used in the project? (Is the demographic attribute a stratifying variable on the selection model for interviewers?) If so, then the appropriate models might be stratified variance models, reflecting "fixed" effects of interviewer attributes.

7. THEORIES OF RESPONSE BIAS AND VARIANCE - DATING OF EVENTS

There is a large literature in survey research on the reports of dates of events within reference periods specified by the questionnaire (e.g., "In the last six months, when did you visit the doctor?"). It is common to observe errors in the dating of reported events. Generally only events occurring before the beginning of the reference period can be misdated into the reference period (forward telescoping). There is some evidence that events evoking vivid memories appear to have occurred more recently (forward telescoping). Response variance on dating appears to

decline for more recent events (lower instability).

There are some models of this behavior, constructed by psychologists studying the topic. For example, Huttenlocher *et al.* (1990) offers a model built on various assumptions. These include a) number of events is uniform in time, b) error rate in reporting independent of number of events, and c) response standard deviation increases linearly in time. From this one can construct a bias adjusted estimator of total number of events, requiring input of response variance, the rate of increase in rate of instability over time, the length of reference period, and the uniform rate of occurrence. Such models, since they require large sets of assumptions, are not practical for use in survey estimation.

Could measurement models of more practical utility be built? The problem is that the theory poses greater complexity than those above. Both response bias and variance are found to change by true date of event. One approach would be to construct an estimator differentially weighting reports by response variance. This would use multiple indicators of date (e.g., free recall of date, measure of how long since event occurred, use of calendrical aid); estimation of response variance (through the covariance analysis of multiple indicators), and the use of response variance in estimation. Such an estimator, however, does not address at all the bias properties of the responses.

When theories specify both bias and variance properties of response then the survey designer must invent multiple indicators that vary in known ways on their variance and bias properties, in order to estimate parameters in a measurement model that reflect both bias and variance.

8. SUMMARY - MOVING NONSAMPLING ERROR THEORIES TO SURVEY ESTIMATION

Theories about error that arise in the social sciences sometimes have implications for the statistical specification of estimators. This requires the statistician involved in survey design, measurement, and estimation to study these theories and invent ways to reflect them in their work. Unfortunately, few social scientists studying causes of behaviors producing survey errors attend to implications for survey estimators. Their focus is on identifying principles of behavior that apply to a wide variety of contexts (including surveys). Nonresponse and response behavior, in that sense, is just one example of these behaviors.

It now seems clear that the "great leap forward" in nonsampling research will come with collaboration between cognitive and social scientists and statisticians. The collaboration will focus on a set of measurement designs and estimators that incorporate auxiliary error models. These estimators will have their error specifications motivated by the theories discovered by the social scientists about behaviors producing the errors affecting the quality of survey data.

9. REFERENCES

Abelson, R.P.. "Psychological status of the script concept," *American Psychologist*, Vol. 36, 1981, pp. 715-729.

Deming, W.E., *Some Theory of Sampling*, New York:Wiley, 1950.

Goyder, J., *The Silent Minority*, Boulder, CO: Westview, 1987.

Groves, R.M., and Kahn, R.L., *Surveys by Telephone*, New York: Academic Press, 1979.

Hansen, M.H. (1951) "Response Errors in Surveys," *Journal of the American Statistical Association*, 46, 147-190.

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

Heneman, G.H. and Paterson, D.G., "Refusal rates and interviewer quality," *International Journal of Opinion and Attitude Research*, 3, 1949, pp. 392-398.

Hilgard, E.R., and Payne, S.L. (1944) "Those Not at Home: a Riddle for Pollsters, " *Public Opinion Quarterly*, 8:2, 254-261.

Huttenlocher, J., Hedges, L.V., and Bradburn, N.M., "Reports of Elapsed Time: Bounding and Rounding Processes in Estimation," *Journal of Experimental Psychology, Learning, Memory and Cognition*, 16, 1990, pp. 196-213.

Hyman, H., Cobb, W.J., Feldman, J., Hart, C.W., and Stember, C., *Interviewing in Social Research*, Chicago: University of Chicago Press, 1954.

Larson, O.N., The Comparative validity of Telephone and Face-to-face Interviewers in the Measurement of Message Diffusion from Leaflets," *American Sociological Review*, 17, 1952, pp. 471-476.

Madow, W.G., Nisselson, H., and Olkin, I., *Incomplete Data in Sample Surveys*, Volumes 1-3, New York:Academic Press, 1983.

Mahalanobis, P.C., "Recent Experiments in Statistical Sampling in the Indian Statistical Institute,"*Journal of the Royal Statistical Society*, 109, 1946, pp. 325-370.

Neyman, J., "On the Two Different Aspects of the Representative Methods: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 109, 1934, pp. 558-606.

Payne, S.L. (1951) *The Art of Asking Questions*, Princeton, N.J.: Princeton U. Press.

Rice, S.A. (1929) "Contagious Bias in the Interview: a Methodological Note, " *American Journal of Sociology*, 35, pp 420-423.

Rugg, D., "Experiments in Wording Questions," *Public Opinion Quarterly*, 5, 1941, pp. 91-92.

Schwarz, N., Hippler, H.J., Deutsch, B., and Strack, F., "Response Categories: Effects on Behavioral Reports and Comparative Judgments," *Public Opinion Quarterly*, 49, pp. 388-395.

Stephan, F.F., (1936) "Practical Problems of Sampling Procedure," *American Sociological Review*, 1, pp. 569-580.

SESSION 1

SURVEY MANAGEMENT AND NONSAMPLING ERRORS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

QUALITY DECLARATIONS AT STATISTICS SWEDEN: PRINCIPLE AND PRACTICE

Claes Andersson[1], Håkan L. Lindström and Lars Lyberg

ABSTRACT

The general principles for quality definition and quality declaration at Statistics Sweden are presented. Their development over the last two decades is discussed in the light of an increasing concern for the users of statistics. Some background is given to explain how variation in ambitions, techniques and resources have changed the possibilities to study and measure quality. For the major quality component accuracy the statements of the quality level of the subcomponents are presented one by one. We mention general approaches to promote measurement and presentation of product quality. Finally we give some examples of products with good quality declarations.

NOTE

This paper was presented at the Council of Professional Association on Federal Statistics in Bethesda, VA, November 12 - 13, 1996. To get a copy please see the reference.

REFERENCE

Andersson, C., Lindström, H.L and Lyberg, L. (1996), Quality Declarations at Statistics Sweden: Principle and Practice. Proceedings of the Council of Professional Association on Federal Statistics, to appear.

[1]Claes Andersson, Statistics Sweden, SCB, U\STM, S-70189 Örebro, Sweden.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

THE MILLENNIUM CENSUS - A TOTAL QUALITY PRODUCT?

Graham C. Jones[1]

ABSTRACT

Conducting the Census is an extremely expensive undertaking and it is therefore vital to plan and manage the entire process from questionnaire design through field enumeration techniques, data coding and capture processes, analysis and presentation as a total quality product.

This paper will outline the approaches being taken in planning the 2001 Census in the United Kingdom to minimise under-enumeration, improve the quality of responses (and response rates); consider the application of Neural network imputation methods and the advances made in image processing and auto coding to enable the relevant analyses to be available to timetable disseminated on the right range of products.

The legislation which underpins the Census requires Parliamentary approval and it is therefore vital that Ministers are convinced that the balance of the burden, the requirement and the sensitivities have been met. Managing the conflict which will naturally occur in such circumstances and ensuring that the decision making process is at one with the overall objective is essential to a successful outcome.

KEY WORDS: Census; Total quality; 2001 planning; Undercoverage.

1. INTRODUCTION

It would be foolish to imagine that in the time available I could address all of the issues connected with conducting the millennium Census in the United Kingdom. Indeed 'I' and the 'United Kingdom' are not synonymous; although I have the title Director Census in the Office for National Statistics, and the Head of the ONS is also Head of the Government Statistical Service - he is only the Registrar General for England and Wales. It is the Registrars General of Scotland and Northern Ireland who have responsibilities for the conduct of the Censuses in their countries.

I do not propose to examine that particular question before this conference either, let it be sufficient to say that the census is directed as one operation (some would prefer the term - partnership) - although some of the specific detail requested in the four countries of the United Kingdom does differ. That is right and proper - for there are different issues to be considered.

2. AIMS AND OBJECTIVES

To carry out the Census of Population and Housing, to process and analyse the resulting data and to disseminate the results. To do this effectively, accurately and with integrity, protecting the privacy of respondents. Which all sounds straightforward! But what that does not say is why we are conducting a Census in the first place. If we consider the Census to be central to official statistics, fundamental to national planning and the only national source of data at the 100% level then we are beginning to address that issue. If we further consider that within those simple statements there is the allocation of billions of pounds worth of government expenditure to regional and local authorities and the Health service, and a substantial amount of information provided in relation to the provision of education and numerous other local services, then we are also coming to the basic business drivers

[1] Graham C. Jones, Director Census, Office for National Statistics, United Kingdom.

behind the Census. None of this, I imagine, will be new, but the context is, I would submit, vital to what follows and the approach we are taking to address the major issues.

3. DEFINING THE REQUIREMENT

In considering the questions to be included in the Census, it is vital to consider what the information is to be used for and by whom. I have already given some indication of that and those involved will know of the seven advisory groups which contribute to the debate. These are from central government, local government, the Health authorities, the academic community (although I do agree this suggests a degree of homogeneity) and the business community (which does not!) and the two groups providing specific input in respect of Scotland and Northern Ireland. These have spawned several content working groups whose proposals and suggestions have been given serious consideration in the run-up to the 1997 Census Test - which is something I will return to shortly. There are clear issues emerging from that consultation and, in addition to the more regular questions which one would expect to find on the Census, there is now a demand for information about income, socio economic group, educational qualification and other issues against a background of considerable social and technological change. But the Census is necessarily a balance; a balance of the demand for information against the burden which can legitimately be placed upon the respondents and all of this within an overall planning framework in which it is quite clear that we cannot devote unlimited resources and therefore there will inevitably be conflict and compromise. That compromise cannot rear its head in terms of the quality of the product - this is vital, for it is important that the users of the Census are absolutely confident with the information with which they are being provided.

4. THE LESSONS FROM 1991

In that respect, there are lessons to be learned from 1991 and the Census Policy Evaluation and Reappraisal conducted in 1992. Having agreed that we need to conduct a census, and note we are right to seek the correct balance in terms of questions and burden to fulfil the defined requirement, then we do need to be entirely clear that we are also producing the right output, at the right level of detail to the right timetable on the right range of products - and the consultation to which I have referred will be vital in this respect too.

In this respect, there were some major successes in 1991 and these cannot and should not be ignored. There was the successful introduction of a question on ethnicity, a high overall coverage and a brand new product - the sample of anonymised records - in many respects the envy of other countries - and is an example of how we work closely with a wide range of users, in this instance Professor Angela Dale from Manchester. There were some downsides too - which must not be ignored either, high under-enumeration in the inner cities, inadequately answered questions and a delay in the production of the results. That some of these results were produced on the basis of only a 10% sample was to many also less than entirely satisfactory.

5. PLANNING 2001

The only way to approach planning such a large task as the census is to break it down into smaller bite sized chunks. Some of these still represent a substantial mouthful and there are issues which arise simply from this disaggregation into projects.

These projects are fundamental to the process of managing the census and are as follows:

- Geography
- Data collection
- Data capture
- Information Systems Strategy
- Output production, policy and marketing

- Data requirements
- Under-enumeration and disclosure
- Edit, imputation and data quality
- Legislation

The reporting mechanisms for each of these are identical in that they all provide input into the Census Programme Board, but the organisation of the census is not the subject of this presentation.

What is important to note is that each of these projects is involved with looking at new methods of conducting the business of the Census.

- The Geography project will be using new GIS systems and digital mapping to plan enumeration districts;

- The Data Collection project is looking at post out/post back methods for distribution and collection of the census forms for the first time in the United Kingdom;

- The Data Capture project is considering automatic scanning in of the data through OMR techniques and is introducing auto coding software to facilitate 100% processing;

- The Information Strategy project is charged with ensuring an integrated system design far removed from our old mainframe environment;

- The Output Production Policy and Marketing project is looking at new outputs on new forms of media; flexibility is vital in this respect - we are very conscious of technological change and the need to keep our options open. We will also wish to identify further partnerships so that we can offer a full range of value added products.

- The Data Requirement project has looked at new questions and new questionnaire design integrated with the requirements for scanning the forms;

- The Legislation project is self explanatory but is central to our relationship with the European Union on a common approach to the censuses being conducted at the turn of the century.

I have deliberately avoided summarising the major statistical projects for it is these projects which are primarily devoted to resolving the problems surrounding the quality of our product and addressing the principal problems arising in the 1991 Census.

6. UNDER-ENUMERATION

Although as I have said, the Census in 1991 achieved a high level of coverage - almost 98% - there was a considerable problem with a differential undercount whereby the under-enumeration was not generally distributed throughout the whole population. Under-enumeration was high for a number of population sub-groups and concentrated in particular areas. This differential under-coverage causes severe problems for the fundamental use of census data at small area level and significant problems in relation to the allocation of resources.

The key groups where under-coverage was a problem were:

- young adults aged 20-29 and in particular young males in the inner cities where the under-enumeration has been estimated to be of the order of 23%;

- infants under one - estimated at 3% nationally;
- armed forces personnel and their dependents;
- elderly women in England and Wales - that is around 6% of those aged 85 and over.

Although the census procedures in 1991 may have missed some addresses, on the whole it is regarded that these were well covered as were persons at those addresses where there was only one household. The principal problem is where there are concentrations of multi occupancy and that is where there is more than one household at one address. It is expected that in 2001, the coverage problems will be at least of the order of magnitude of those experienced in 1991 and in all probability the problem will be greater. We expect more one person households; more people absent during the day; and more households which may be difficult to enumerate because of increased security such as entry phones etc. The mobile population - that is those who reside at more than one address - is also likely to have increased.

So what have we learned from the 1991 situation?

- there was too little paid publicity and it was not effective because it was not directed at the high risk under-enumeration groups;
- publicity did not attempt to tackle the problem of anti-officialdom;
- enumerators were difficult to recruit and there were therefore shortages of field staff;
- multi occupancy procedures were complex and difficult to define;
- the level of non contact was higher than expected - an issue which had not been planned for;
- the census form was cluttered and instructions not clear;
- local contacts were made too late;

and finally, if that was not enough, the whole planning process did not start early enough!

Now it has to be said that these lessons are painful - but they cannot be ignored.

So how are we preparing for the 2001 Census specifically in relation to addressing this particular problem?

We are looking at

- the definition of the population base,
- the date of the census,
- geographical planning,
- form design and content,
- recruitment and training of field staff,
- establishing local contacts,
- improving publicity.

The population base definition will not simply need to be comprehensive - to avoid the holes which might lead to under-enumeration, but must also be understood and communicated both to the public and to field staff.

We will enumerate students at their term time address. As the census date is planned to be in term time, students at universities and colleges and schoolchildren at boarding schools will be enumerated at those places.

The difficulties of defining residents in communal establishments to ensure that they are also correctly enumerated will be reconsidered particularly in respect of the elderly for it is believed that this is the source of under-enumeration of that group. People with more than one address - commonly people who work in one part of the country and live in another will also be enumerated appropriately.

We are improving our geographical planning with the introduction of customised maps, the use of pre-printed address lists, and updated information from local authorities - all aspects which will be included within the 1997 Census Test. These improvements will enable enumerators to be alerted to the likelihood of particular difficulties in their areas in respect of multi occupancy of non residential property.

The design and content of the census form is critical to response rates and the quality of data to be collected. The appearance, layout and wording on the form have all been researched to enable it to be completed more easily. The 1997 Census Test will include a careful assessment of questions which might have an adverse effect on coverage and this will be part of our follow-up survey.

Through the improved recruitment and training of field staff with specific modules addressing the problems identified in 1991 and close working with local authorities, the police and community organisations, we will address directly the problems of coverage in the inner cities.

Finally, in this section our publicity needs to be specifically targeted to address the particular problems which I have alluded to already i.e. young adults, recent mothers and the elderly. It will be simple and have a direct message about what is expected of respondents and we will promote the census further in the education sector with positive messages explaining its purpose.

7. THE DATA QUALITY MANAGEMENT PROGRAMME

One innovation for 2001 is the introduction of the Data Quality Management Programme as a means of improving and measuring data quality. The Programme itself is an aspect of our (loose) implementation of the Total Quality Management (TQM) approach which has been adopted so successfully in Australia.

TQM involves a commitment to delivering a product that meets customer requirements by focusing on the quality of all the tasks carried out in the organisation, this commitment to quality must come from the top down and be understood by all. Examples of the activities we are engaged in that fall under the TQM approach are

- widespread customer consultation,
- a prototyping approach to the development of systems,
- 1997 Census Test.

The Data Quality Management Programme (DQMP) has several aspects

- co-ordination across all areas of the 2001 Census to examine data quality issues and to develop quality standards in consultation with processing specialists

- monitoring and measuring data quality during the census operation to enable problems to be identified and acted upon

- collection of information to inform the user of the quality of the published data e.g. the amount of imputation carried out for each question

- collection of data to feed into the One Number Census estimation procedures.

A key aspect is therefore the development of a system to monitor and measure data quality during the census operation - the Data Quality Monitoring System (DQMS). This will take inputs from the various census operations (collection, data capture, processing and output) as well as from external data sources (previous census, administrative records, etc. to enable pre-determined quality standards to be monitored and provide ad hoc interrogation facilities of the database by staff with specialist subject matter knowledge.

Geographical accuracy was not given sufficient prominence in the 1991 Census; there were problems related to the geographical base and incompatibilities between enumeration districts and postcodes. Much of the information gathered from external sources was incorrect and changes were received quite late in the planning cycle. As a consequence, corrections and amendments were made to separate geography databases one for input and one for output and these were not always synchronised and not discovered until later. The use of Geographical Information Systems and digital mapping together with specific information relevant to the geographical areas - such as a prison or a communal establishment which may be giving a distorted pattern - will also be available at an early stage during processing.

The unclear definitions of residence which lead to confusion in 1991 will as I have already said be revised. This incorrect interpretation of residents lead to incorrect form completion and people being missed. Our revised definitions will be tested during the 1997 Census Test with specific targeting on the rules related to communal establishments and visitors.

The differential undercoverage which I have already described will also be tackled within the Data Quality Management programme. Having available information in a Field Management Information system to provide data about emerging undercoverage as it occurs will allow us to explore how resources can be deployed in response to particular problems.

The scanning technology will naturally improve the quality of the source information - as it will not have been mistranscribed through errors in keying as has occurred previously. But new problems will emerge through multiple ticking of answers which will need to be handled. The Data Quality Management programme will record the incidence of such occurrences - identify the questions which give rise to such ambiguity and ensure that the right edit rules and correction techniques are applied.

Incorrect and inconsistent interpretation of coding rules and errors in reference datasets have led in previous censuses to wrong codes being applied. Auto coding techniques will to a large extent overcome these problems but the DQMP will apply frequency counts to coded fields to allow comparisons of distributions pre and post coding geographically and seek to apply cross field checks for consistency. Planning in advance for higher levels of specific ethnicity in specific areas of inner cities is but one example of this approach.

One of the principal problems identified in 1991 was that errors were found at the tabulation stage when it was too late to react to them and there was considerable incompatibility between the 100% and 10% datasets which was only found after the release of the data. By monitoring data during processing and by using early cross tabulation and frequency counts we will be able to identify problems much earlier in the cycle.

8. IMPUTATION

In 1991 a sequential hot deck imputation method - based on Holt/Fellegi - was used for items not corrected by the edit matrix. But the major problem was that there was no post imputation consistency check for 2001. We are currently running trials in the statistical operational aspects of applying Neural Networks for imputation - and these trials are showing early signs of success. We are however still giving consideration to other methods; modification to the 1991 hot deck system, multi level modelling, and the Canadian New Imputation Methodology (NIM) are examples, but these need to be evaluated so that we can assess the performance of our imputation methods and this together with post imputation consistency checks should address our principal concerns. Professor Ray Chambers is working with us in this evaluation.

9. THE ONE NUMBER CENSUS

The Census provides the benchmark on which we rebase the annual population estimates. As I have already stated, these figures are vital in the allocation of public funds and it is therefore of the utmost importance that all of our users have complete confidence in the information that we are providing. If there is room for doubt or question about our regional estimates and how these compare one region with another then all of the uses to which the Census is put are in jeopardy.

We have therefore decided that one of the principal objectives of the Office for National Statistics in the run up to the 2001 Census will be to undertake research and fully evaluate the prospect of producing information from the Census at all levels on one common basis.

This will not simply require technical expertise, it will also require exceptionally careful management, and handling of the difficulties in producing such a dataset, should not be underestimated. The project will be undertaken under Professor Diamond's direction at Southampton University and report to a Project Board to be chaired by the Director of Statistical Methodology of the Office for National Statistics and quality assured by an eminent panel of experts under the chairmanship of myself. This is another example of where we are building links not just with the academic community, but within the ONS and with other National Statistical Offices.

The estimation procedures will be complex using data from a variety of different sources.

- Census Validation Survey
- Demographic Analyses
- Administrative records

Use these sources to estimate net undercoverage at Regional level and model regional estimates to lower levels of geography (e.g., using regression based approaches). And then, if this is feasible to produce a fully adjusted database at microdata level - consistent outputs.

And finally an agreed methodology will be developed for testing in the 1999 Census Dress Rehearsal.

10. CONCLUSION

I hope today that I have given you some idea about our approach to the 2001 Census, the changes which we are introducing and the innovations which we are researching. The 1997 Census Test which will take place in June of next year will be a vital cog in our decision making process. It will help us decide crucial issues prior to the 1999 Dress Rehearsal and the White Paper to be discussed by Parliament. We are working closely with a wide range of Census users and the academic community to produce what I hope will be a total quality product.

11. REFERENCES

Census Division, The One Number Census: Roma Chappell (Unpublished). Office for National Statistics.

Clark, A., (1995), Problems in large scale data collection: The 1996 Census, *Office for National Statistics.*

Clark, A., (1995), OPCS Census Development Programme, *Census Office for National Statistics.*

Craig, J., (1995), The Background – 1991 Census of England and Wales, *Census Division OPCS (now ONS).*

Dugmore, K, (1995), What do Users want from the 2001 Census?, CACI.

Heady, P., (1995), Census Validation Methods, with special reference to identifying which dwellings are occupied and to capture/recapture estimation, *Office for National Statistics.*

Thomas, J., and Teague, A., (1996), Neural Networks as a possible means for imputing missing Census data in the 2001 UK Census of Population, *Census Division,* Office for National Statistics.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

STANDARDS AND GUIDELINES FOR NONSAMPLING ERROR: PROVIDING BETTER SERVICE FOR LESS

Richard D. Burgess[1]

ABSTRACT

Providing data that are "fit for use" is a primary objective of a statistical enterprise. For the enterprise to be successful users must believe the data meet their information and quality requirements. Generally users must be provided with accurate data, that are consistent with their own conceptual needs, that are current, obtainable at a "competitive" price, and increasingly that yield information that are based on longitudinal data or data integrated from multiple sources. Standards and guidelines can be used to good effect in meeting the broad range of user requirements. Whether explicitly or implicitly applied, and in particular to deal with nonsampling error, standards and guidelines help to facilitate improvement in productivity and quality; thus to provide users with better service at less cost. This paper provides some views on the appropriateness of standards and guidelines in the context of nonsampling error. The strengths and weaknesses of these are briefly discussed and examples of their use at Statistics Canada are given.

KEY WORDS: Standards and guidelines; Nonsampling error; Client service.

1. INTRODUCTION

Clients of a statistical enterprise surely want data that are "fit for use". It is not uncommon to assume that this means that users must be provided with data that are accurate, that are consistent with the user's conceptual or analytical needs, that are current and that are obtainable at a "competitive" price.

Subsumed in their broader requirements, users might not unreasonably expect that a statistical agency or business, as also a scientific or professional enterprise, should and will have explicit standards for concepts, terminology, methods and quality. They might also expect that any differences or deviations from the "standard" practices are for demonstrable and substantive reasons. They will want, although perhaps not expect, that these standards will be based on the most up to date knowledge, will as a result meet the user's current needs and will be consistently reflected in all statistical data products. Given the progress in technology users will eventually expect that such "standardization" will permit an efficient enterprise to create a wide range of cross-sectional, longitudinal and integrated data products, and will meet users needs with microcomputer speed and economy.

To meet user needs and expectations the primary standards of a statistical enterprise must be the knowledge and the experience of its people. Other standards and guidelines do have a place in delivering a complex range of data fit for use and in delivering them efficiently; standards and guidelines for controlling, measuring and describing nonsampling error in statistical data.

2. WHAT AND WHY STANDARDS AND GUIDELINES

2.1 What are standards and guidelines?

Standards might be rules or regulations to ensure compliance or uniformity. For nonsampling error and statistical enterprises, uniformity is not the appropriate context in which to be considering standards. Standards might better be taken as a minimum requirement, the generally accepted model or practices, or a benchmark or model of excellence. There is an element of each of these in most such standards.

---

[1]Richard D. Burgess, Social Survey Methods Division, 15-F, R.H. Coats Bldg., Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Guidelines might be similarly defined but any sense of compliance is reduced and generally so to will be the extent of specifics. Flexibility of implementation may be greater thus reducing any element of uniformity. As a terminology «guidelines» may be more palatable, but perhaps less consistently implemented.

2.2 Why is knowledge not enough?

Given that the appropriate sampling requirements are set, nonsampling error is central to the management of the quality of a survey, or other statistical project, and of its cost. Nonsampling error is controlled through the application of methods, but not necessarily with a clear notion of the cost or overall benefit of these methods. The quality is managed on the ongoing use of judgement, not just on the use of the knowledge of the many people contributing to the project. Those making the day to day judgements may have a restricted perspective and may become more concerned with expediencies of operations and cost management.

Setting of standards or quality targets at the outset focuses knowledge where it is required and permits the expression of this in criteria and acceptable practices. These make clear to all from the design stage onwards what is required and, perhaps, what is good enough. These criteria and practices may then reflect the corporate perspective, as well as those of the particular program and its users. Such criteria, for example, may be embedded in quality control, in edit rules in follow-up requirements, in minimum response rate targets and in a maximum for interview duration. The practices may be requirements for quality control, editing, for testing of questionnaires and of operations and processes, or for reporting of results and methods and data quality.

2.3 What might be appropriate standards and guidelines?

Some of the arguments against standards are:

- they stifle creativity, innovation, and quality and productivity gains;
- they satisfy the lowest common denominator and will eventually bring all down to this level;
- too much effort must be expended to make standards cover all situations and they simply become convoluted, unintelligible or simplistic, and not implemented.

Survey managers and practitioners seem to prefer the practices and criteria of their own choosing. Such choice will too frequently lead to needless diversity, and in some cases to the implicit downgrading of quality as an issue, to the use of dated methodologies, systems and vehicles, and to less satisfied clients.

Standards and guidelines should not and need not be applied in such a way as to restrict the use of knowledge and experience. They should be designed to improve efficiency, effectiveness or productivity. Quality assurance and standards should be used to do more than merely monitor or measure data quality. They should have some positive short and long term impact. Although they require a sound basis in current knowledge and capacities, in some aspects standards and guidelines for nonsampling error should be at the leading edge and supported by research and development activities. Standards should discourage and eliminate bad or inferior practices, encourage good practices, and give direction and goals for technical development and to improvement in client service.

Standards can be a coalescence of corporate quality values and awareness of respondent sensitivities. They should be based on enunciated principles and allow for extending the application of the principles to situations not explicitly covered or not considered. They should reflect the degree to which programs within the enterprise are interdependent and connected, sharing the same clientele, the same fiscal base, the same reputation, and the same knowledge and tools. At the same time not all standards need be generated at the corporate level. Not all programs are of equal importance, or have equivalent resources or quality demands. There should be room and a requirement for individual programs to have their own quality standards and quality assurance practices.

3. USE OF STANDARDS AND GUIDELINES AT STATISTICS CANADA

Over the past of decade there have been significant technological advances - operational, statistical and computing - that have given opportunity to the Agency for improvements in the efficiency and data quality of its programs. This also has been a period of fiscal constraint, increased demands for more and different forms of data and greater emphasis on the quality of service to clients. The Agency is producing a greater diversity of data and products. It is moving to greater use of electronic collection and dissemination, more integration of data, creation of longitudinal data, data analysis and support for data analysis on social and economic issues and policies. The Agency is being asked to do more with less, to do more of what is complex and vulnerable to the impact of nonsampling error.

The assurance and management of quality for all statistical programs of the Agency are to be conducted within a framework of established practices, standards and guidelines, policies and technical knowledge. As in any other statistical organization knowledge and training are the most significant and pervasive standards and guidelines directing the professional activities of its people. Recruitment, training and development of staff, internal committees and external advisory committees, in effect, encourage consistency of good practices across statistical programs. There are as a result implicit standards and guidelines for managing quality and controlling nonsampling error. Follow-up for nonresponse, editing for consistent and complete information, quality control of operations, imputation and weight adjustments for inconsistencies and nonresponse are customary practices. Specific criteria, practices and methods and what is the minimum acceptable level of quality, however, are left to the individual program to workout and justify within its circumstances, constraints, opportunities and objectives.

The specific policies and standards and guidelines affecting the management of nonsampling error that are in place at Statistics Canada include:

- Guidelines for Seasonal Adjustment of Current Observations and Their Revisions
- Policy on Informing Users of Data Quality and Methodology
- Standards and Guidelines on the Documentation of Data Quality and Methodology
- Policy on the Development, Testing and Evaluation of Questionnaires
- Guidelines for Questionnaire Testing and Evaluation
- Standards and Guidelines on the Presentation of data in tables of Statistical Publications
- Standards and Guidelines for Reporting of Nonresponse
- (Draft) Policy on Data Quality Criteria in the Dissemination of Statistical Information
- Standards and Guidelines for the Application of Data Quality Criteria in the Dissemination of Statistical Information
- Policy on Estimates with Future Reference Dates
- Quality Guidelines (a guide to methods to building quality into and assessing quality of a survey)

The implementation of these policies, standards and guidelines, etc., are subject to regular review and revision, usually by the Statistics Canada's Methods and Standards Committee, the Advisory Committee on Statistical Methods and the Agency's Internal Audit Committee.

Generalized statistical systems developed within the Agency for use throughout the Agency are a standard in that they afford consistent and generally current practices and methods for the fundamental statistical functions. These are typically used for new and redesigned surveys. There implementation is at relatively low cost.

There are also standards for the definitions of geographical units, and industrial and occupational classifications. Harmonization or standardization of definitions for subject matter variables and the development of meta-data databases are being pursued. These provide or will provide for more efficient use of the Agency's data, more comprehensive information on concepts, methods and data quality, and presumably reduce misuse and misinterpretation of data.

3.1 What have standards achieved?

A comprehensive view of the impact of policies and supporting standards and guidelines is not intended or forthcoming here. However, there are two important examples for which some assessment may be useful. First,

the policy and standards and guidelines on questionnaire design were developed and practical to implement as a result of the ongoing development of more sophisticated and cheaper methods of questionnaire testing. While there remain examples of entrenched designs and sometimes limited resources being afforded questionnaire testing and development, there is a sense of greater enlightenment on the value of this testing. Much of this has been an outcome of the results of testing and its significant effect on final designs. This is a good example of where considerable improvement in data quality and reductions in respondent burden have been be made inexpensively.

A second important policy is that on informing users of data quality and methodology. This policy and the supporting standards and guidelines, in their third edition, were first introduced to Statistics Canada in 1978. Requirements of the policy started as a voluntary endeavour, a statement of values, but now has mandatory elements as well as extensive guidelines. The requirement to measure quality is a basic principle of the policy, as is the view that the user should be offered all available information on the quality of the data, even if this information is incomplete. There is general compliance with the policy at the survey or statistical program level (but not always at the product level) in regard to providing estimates of sampling error. The results for nonsampling error are less satisfying. Many of the measures or indicators of nonsampling error, when made readily available, are not easily interpretable by users for specific data or even a specific tabulation or set of data. Frequently the only effective assessment of quality given in regard to nonsampling error is the assurances provided through descriptions of methodology. The assurances regarding the quality given by quantified measures, for example a response rate, may only have relative meaning. User comfort with the data may be based more on the reputation of the Agency, on the past performance of the particular program, and on the descriptions of methods which appeal to common sense rather than to quantified or historic results. Only if there is a general or large problem indicated are the measures otherwise useful.

Despite these limitations it is not unreasonable to conclude that those programs which do more for users in describing methods and quality are those for which the measurement of quality and use of the measures are built into the data validation or certification process. Not surprisingly these are the programs for which there is significant focus and a meaningful amount of resources afforded for quality improvement. These are also the programs which are able to deal effectively in public fora with issues of quality, interpretation and relevance raised by users; a particularly important factor for data related to matters of pre-existing and well established opinions.

3.2 Design Standards?

The policy on data quality release criteria is an experimental effort designed to overcome some of the limitations of reporting of data quality measures, to reduce the potential for marketing of data of poor quality and to require pre-specification of design standards or targets for data quality. The policy requires the classification of all data as one of:

- quality appropriate to design
- marginal quality
- unacceptable quality.

Data could be classified to one of the first two categories based on sampling error or to any of the three based on nonsampling error. The classification is to be an overall assessment and is to be reported to users. Data of appropriate quality are to dominate standard Agency products. These products are not to include data of unacceptable quality. Users can request and receive any data but must be advised about quality prior to commitment by either party. The policy is under experimentation through a selected few major statistical programs.

It might be argued that this policy exceeds current knowledge. What it does do is to force the supplier of the data to do what the supplier is currently asking the users to do. While setting levels for categories of quality may be questionable in the context of all potential uses of the data, it seems legitimate in the context of the design targets or expectations. The quality assurance design clearly is stating, knowingly or otherwise, what level of quality is "good enough" given priorities, objectives and constraints.

The policy does exceed current practices in estimating or ascribing nonsampling error to individual data cells. The classification and identification of data cells according to level of sampling error is consistent with current and routine quality reporting practices. Current practices of estimating and reporting of nonsampling error, however, ascribe error to large blocks of data only. The long term viability of this aspect of the policy is specifically linked

to better or more suitable measurement of nonsampling error. Where this seems most likely is for indicators of the potential effects of quality adjustments; the effects of weighting for nonrespondents and the effects of imputation. These effects can be measured for individual cells without serious difficulty. What will still be needed is an interpretation of the potential impact on quality of the adjustment, or lack, on any given cell. Something which might be examined is the effects or distribution of effects across statistical programs to determine what is relatively tolerable or atypical. This is merely an extension of the current use of response and imputation rates.

4. CONCLUDING COMMENTS

Standards and guidelines are merely a means of formalizing certain aspects of what the organization believes should be done by each element within the organization. They do not in themselves cause productivity gains or improvements in quality, or reduction in nonsampling error. They are a means by which good practices are communicated and reinforced and less efficient or satisfactory practices eliminated. Their implementation may impose considerable cost for new and additional applications. However, given that the practice is to be applied the use of standards and guidelines as a vehicle to focus and ensure implementation can realistically deliver efficiency gains through the communication of effective methods, and the experience and developments of others.

Standards and guidelines are a means of giving a presence to issues and principles of the organization in day to day activities. If done properly this will yield quality improvements and consistency, it can lead to research of better and cheaper methods. Unless the standards and guidelines can achieve this they perhaps should not exist.

5. REFERENCES

Bradley, B. and Silins, J. (1995), Data Management for Canada's Health Intelligence Network: Building a Virtual Information Warehouse Through Standards, Cooperation and Partnerships, Proceedings of Statistics Canada Symposium 95: From Data to Information - Methods and Systems, Statistics Canada.

Burgess, R.D. (1990), An Examination of Statistics Canada's Data Quality Release Criteria, Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Statistics Canada.

Early, J.F. (1990), Managing Quality in National Statistics Programs, Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Statistics Canada.

Fellegi, I.P. (1995), Characteristics of an Effective Statistical System, Presentation to Washington Statistical Society, Washington D.C., October 25, 1995.

Kruskal, W. (1990), Introduction to Measurement Errors in Surveys, Edited by Biemer, Paul P., Groves, Robert M., Lyberg, Lars E., Mathiowetz, Nancy A., Sudman, Seymour, John Wiley & Sons, Inc. 1991.

Priest, G. (1995), Data Integration: The view from the Back of the Bus, Proceedings of Statistics Canada Symposium 95: From Data to Information - Methods and Systems, Statistics Canada.

Woltman, H.F., and Thomas, K.F. (1990), Measurement of Content Data Quality in the 1990 Census, Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, Statistics Canada.

SESSION 2

FRAME ERRORS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

MEASURING ERRORS ON THE BUSINESS REGISTER

N. Laniel, L. Mach, H. Finlay, S. Dionne[1]

ABSTRACT

The Business Register is a list frame for business surveys at Statistics Canada. It is mainly based on two independent administrative files. The primary source is the list of Employer Payroll Deduction Accounts which are used by businesses to remit monies from source deductions for all employees. The other source is the Income Tax Filing process through which the annual income of businesses is reported and the income tax collected. The frame data obtained from these two administrative sources is imperfect and thus different types of error (i.e., undercoverage, overcoverage, misclassification, erroneous size and inaccurate contact information) are present on the Business Register. In this paper, all the types of error along with their sources are defined and their impact on survey estimates discussed. Then, the various methods which can be used to measure these errors are reviewed and analyzed. Also, results obtained with some of these methods are presented and discussed. The paper concludes with a brief discussion of the future developments for measuring the Business Register error levels.

KEY WORDS: Administrative lists; Coverage deficiencies; Misclassification; Erroneous contact information; Error measurement methods.

1. INTRODUCTION

The Statistics Canada Business Register is a central repository of information on businesses in Canada based on two administrative lists from Revenue Canada (RC). It is used as the principal source to delineate annual and sub-annual business survey frames for the economic statistics program. As such, the Business Register (BR) plays an important role in contributing to the production of a coherent and integrated set of economic statistics.

Business surveys are used by Statistical Agencies to provide key economic indicators for policy analysis. These indicators should be reliable and reflect the current economic conditions. This need for reliable and current information, expressed in terms of frame requirement, means that a survey list of businesses, and thus the BR, must contain accurate data that is kept as up-to-date as possible. This is not an easy task since the business universe is dynamic in nature.

The BR data accuracy and up-to-dateness depend on two main contributing factors: (i) the quality and timeliness of the two administrative lists and (ii) the amount of resources devoted to its maintenance. Given that administrative systems and statistical systems do not have the same data quality requirements and that the amount of resources that can be devoted to the maintenance of the BR is finite, it is inevitable that errors will be found on the BR. The types of error present on the BR are: (i) coverage errors (i.e. undercoverage, extraneous units and duplicates), (ii) classification errors (i.e. industry and geography misclassification), (iii) errors in stratification variables (i.e. incorrect revenue and employment class) and (iv) errors in contact information (i.e. erroneous name or address).

It is important to measure the level of each type of error in order to efficiently allocate the finite amount of resources. These measures serve to minimize the level of the errors, and to provide statistics which may be used by surveys to account for the BR deficiencies when designing samples or producing estimates.

[1]N. Laniel, L. Mach, H. Finlay and S. Dionne, Statistics Canada, 11th floor, R.H. Coats Building, Ottawa, Ontario, CANADA, K1A 0T6.

As an example of the use of error measures when designing a sample, if one knows the proportion of extraneous units before drawing the sample, one can use this information to increase the sample size to get the desired sampling variances for the estimates. Latouche and Hidiroglou (1987) provide a method of sample allocation taking into account the proportion of unidentified extraneous units on a survey frame.

An example of using error measures for producing estimates is the case of the Survey of Employment, Payrolls and Hours, which uses a measure of the undercoverage of the BR to adjust its survey estimates.

Out of the many types of errors on the BR, some are easy to measure while others are difficult. Also, errors are easier to measure for simple businesses than for businesses with complex organizational structures. Many means can be used to measure errors on the BR. It is the purpose of this paper to review some of them and to present results obtained in the past years when these have been applied to the Canadian register of businesses.

The next section of this paper deals with the design and maintenance of the BR. It provides the basis to define the types of error of the register and their sources. These are discussed in the third section. The fourth one reviews the different methods that can be used to measure the level of errors on the BR and also presents some results obtained for the Canadian BR. Finally, section five concludes with future developments with respect to measuring errors present on the BR.

2. DESIGN AND MAINTENANCE OF THE BUSINESS REGISTER

2.1 Design

The Business Register is a list frame mainly built with two administrative sources from Revenue Canada (RC): the annual Tax returns of both corporations (T2's) and individuals (T1's) and the Payroll Deduction accounts (PD's) (Colledge, 1987). The PD's are accounts used by employers to remit, usually monthly, monies to RC for Pension Plans, Unemployment Insurance contributions and other deductions. There is no common identifier between the Tax and the PD sources. Therefore it is not feasible from a cost and time point of view to link these two sources due to the large number of businesses, over two million in Canada. As a result, it was decided to construct a register composed of two portions: one for the large and/or complex businesses, where the Tax and PD sources are linked, and one for the smaller businesses, using only the PD source. For more details on the system implementation of the BR see Cuthill (1996).

The first portion, named the Integrated Portion (IP) consists of businesses with a complex organizational structure that are active in more than one industry or province (around 10,000 businesses) and of large businesses with a simple structure (about 88,000 businesses) having a revenue above specific thresholds. These are defined by 2-digit Standard Industrial Classification (SIC) code and province. The businesses in the IP cover over 70% of the total revenue in each 2-digit SIC by province class. The information regarding the structure of the complex organizations is stored on the BR. This is done to facilitate the collection of the data from the right part of the organization. However, the two administrative sources lack the structural information and thus profiles must be created via direct contacts in order to satisfy survey taking requirements.

The second set of units, named the Non-Integrated Portion (NIP), is composed of the remaining 800,000 simple and smaller employer businesses. This list of small businesses is based solely on the PD accounts. They provide a current source of information to update the BR. For example, they can be opened or closed by an employer at any point in time. The disadvantage with this source is that non-employer businesses are not covered. The employment and income figures are predicted using a model based on current payroll remittances.

Together, the IP and NIP exclude 1,300,000 PD accounts. Most of are out-of-scope (OOS) to business surveys but some are in-scope. These excluded units are part of what is called the ZIP list. Typically, OOS PD's are accounts without remittances for the last twelve months, household accounts, foreign accounts, accounts owned by businesses which have ceased their economic activities, government special work program accounts, pension plan accounts or accounts used to solve a succession (estate). The in-scope ZIP PD's are essentially accounts,

owned by active Canadian businesses, for which a proper SIC or Standard Geographical Classification (SGC) code is not known. They are referred to as the unclassified records.

2.2 Maintenance

Two types of sources maintain the BR as up-to-date as possible. These are the RC administrative data and the BR surveys collecting frame data. In the following paragraphs, each of the BR sources of updates is described and the types of updates they provide are enumerated.

2.2.1 Revenue Canada Administrative Data

There are two major RC sources of updates based on PD accounts: (i) the PAYDAC file for births, deaths and name and address changes and (ii) the PD20 forms for initial classification data.

The PAYDAC file is an exhaustive list of all the PD accounts opened by employers. It is updated on a daily basis by RC. That is, as soon as an employer requests the opening of an account, a new record is added to that file. Also, when an employer ceases his economic activities and sends all monies due, then the account is closed and the corresponding record deleted from the PAYDAC file in the month of January that follows at least 12 months after closure, providing no reactivation occurred.

Once a month Statistics Canada (STC) receives the latest version of the PAYDAC file. It is then matched with the set of large businesses. Accounts that match the large units are assigned to that set and are used to signal changes in legal, operating or accounting structures of these businesses as well as potential births or deaths of units within these structures. The signals of change trigger contacts with the large businesses in the form of Business Register surveys (see Section 2.2.2).

For the smaller businesses the use of the PAYDAC data is more extensive. It is not only used as a signal but as the major source for maintaining the list of businesses and the frame data. The new accounts that do not match the set of large businesses are assigned to one of three lists. They are assigned to the list of smaller businesses as births if they remitted in the last twelve months and full classification information is available. They are asigned to the set of unclassified accounts if they remitted in the last twelve months but the classification information is insufficient. Finally, they are assigned to the set of OOS units if they did not remit in the last twelve months or the unit meets one of the exclusion criteria.

In principle, every month employers remit monies to RC which then credit their accounts for the dollars received. That information is kept on the PAYDAC file made available to STC on a monthly basis. For the smaller businesses, the remittances are used to update the Number of Employees and Gross Business Income, via models, when the difference between the old and new values is larger than some tolerance limits. Those accounts which did not remit for twelve months get an estimated Gross Business Income of zero. All the accounts in the NIP which get an estimated Gross Business Income of zero are removed (deathed) from that list, that is they are moved to the ZIP OOS list.

Each time an employer remits monies to RC, he is sent a PD7AR receipt. A portion of the PD7AR form may be separated and returned to indicate any changes of name or address. Such changes are incorporated on the PAYDAC file and then used by STC as a signal of change for the large businesses and a direct update for the smaller ones.

When an employer's request to open a PD account is received, RC sends the business a PD20 form to complete. This form essentially asks for legal and operating data (including the expected number of employees and a description of his industrial activity). It can then take a few days or many months before that form is sent back to RC.

Copies of the PD20 forms received by RC are sent weekly to Statistics Canada. These forms are then captured and used to initialize the frame data for the smaller businesses. This constitutes the main source of initial SIC codes and SGC codes for most births. When the copy of the PD20 form of an account is received and captured at STC then that account is birthed in the NIP of the BR if the classification information on it is sufficient and the account remits

monies. Otherwise, it goes to the ZIP OOS if there are no remittances or the ZIP Unclassified if there are remittances.

Another administrative source of updates for the large businesses is the annual Tax returns. Every year STC gets copies of the Tax returns from RC. These reports signal potential changes in business structures that have not been identified through the PD accounts and allow the update of revenue data.

2.2.2 BR Surveys

Statistics Canada conducts surveys which are used to obtain frame data on businesses when some pieces of data are missing, to confirm a change detected, or to refresh the data. For a discussion of some points considered to decide upon the BR surveys needed for updating, refer to Colledge, Estevao and Foy (1987).

For the smaller businesses, one BR survey used is the Business Activity Report. It is a mail survey conducted for remitting employers for which the PD-20 data was not made available to the BR within 90 days from the opening of an account or for which the PD-20 business activity description could not be properly SIC coded. When the BR does not get a response from the business owning the PD account using the Business Activity Report, then it gets followed-up via the Classification Survey which is a telephone survey conducted from the Regional Offices. The Classification Survey collects additional information such as legal, contact and administrative information.

Another survey currently conducted by the BR for the smaller businesses is the Promotion Survey. This survey primarily consists of contacting all the small businesses showing an estimated Gross Business Income larger than the IP/NIP revenue thresholds in order to confirm that their revenue is effectively of large size. When the revenue of a business is confirmed to be of large size, then that business is to be promoted to the large business portion and no frame data is collected. Instead the data is collected via a reaction profiling activity which is discussed below. The promotion of a unit to the set of large units generates the death of a smaller unit and the birth of a large unit on the BR. If the revenue is identified as being small size, then the business stays in the set of smaller businesses and the same frame data as for the Classification Survey is collected on the spot.

The New Entrant Survey (NES) is a survey conducted by the BR for both the large and small businesses on the BR, which enter into economic survey samples. It collects the same frame data as the Classification Survey. The NES is the main source of survey sample information to compensate, via domain estimation, for incorrect SIC codes on the BR.

Another BR survey, which is conducted for both large and small businesses in economic survey samples, is the Survey Frame Feedback. It is conducted for any unit for which an economic survey has signaled a discrepancy between the frame data and the information given by the respondent while economic data was being collected. This BR survey provides the same frame data as the Classification Survey.

In addition, there is an activity called profiling which is a survey of large businesses collecting data on their, often complex, operational, legal and accounting structures. Farrall and Demmons (1987) discuss issues involved in profiling businesses. Two types of profiling activity are in use. First, there is reaction profiling. This consists of, for example, doing a mini-profiling exercise triggered by a change in the data of the set of PD accounts owned by a large business or a change in the information on its Tax returns. Clark and Lussier (1987) discuss in some details the use of administrative data in reaction profiling. The other type, referred to as cyclical profiling, is an extensive and periodic profiling exercise of large businesses with the goal to maintain up to date information on these businesses. Currently the cycle is around two years. The goal is to reduce this to one year as more experience is gained in doing cyclical profiles and the tools used are improved.

3. DEFINITION AND SOURCES OF ERRORS

It is well recognized that perfect business registers do not exist and cannot exist since only a limited amount of resources can be devoted to their maintenance, as well as difficulties in keeping them up-to-date (Tupek, Copeland

and Waite, 1988). Definitions and sources of frame deficiencies are given below along with their impact on economic survey estimates. The deficiencies being described in this section are not peculiar to Statistics Canada BR but also are common with those of other countries such as the U.S. frames (Konschnik, 1988).

3.1 Missing Units

Missing units are businesses that should be present on the IP or NIP of the BR but are not. There are six categories of missing units for the Business Register. The first category includes the unclassified PD remitters in the ZIP for which contacts are being made via the Classification Survey. The second one consists of units wrongly categorized as out-of-scope in the ZIP list, for example, active employer businesses that failed to remit monies to RC for over 12 months. The third category is composed of those new active employer businesses that recently opened a PD account but have not yet been added to the BR. The process of adding new active accounts introduces a lag of a half to one and half month. The fourth category includes all of the smaller active non-employer businesses, which are excluded from the BR by design since the NIP is based solely on the PD administrative source. The fifth category consists of large active non-employer businesses not yet introduced on the BR due to time lags introduced by the annual Tax administrative source and the processing of the information. Finally, the sixth category comprises the sub-units within large organizations, already represented on the IP, but not yet added to the BR due to the time lags for profiling complex units and processing the resulting profiles.

Since the missing units cannot be included in the survey sampling frames, they introduce a downward bias in the survey estimates.

3.2 Extraneous Units

Extraneous units are units that should not be present on the IP or NIP of the BR. There are three categories of such units. The first one consists of NIP remitters whose accounts are owned by organizations that are not involved in an economic activity (e.g., household accounts). The second one includes NIP inactive employer businesses that remitted in the last 12 months. Note that this is a deficiency by design as an attempt is made to minimize the undercoverage due to the active employer businesses which are slow in remitting monies to RC. The third category consists of large inactive non-employer businesses, or part thereof, not yet identified as inactive via reaction profiling (from annual tax sources) nor cyclical profiling.

There is no source that can identify all of the extraneous units. As a result, extraneous units may introduce a bias and an increase in variance of the survey estimates. An upward bias in the estimates will occur when extraneous units are not identified as such by the sample survey and are then imputed. When the extraneous units are identified by the sample survey, the bias is reduced but there is an increase in variance because the effective sample size is reduced.

3.3 Duplication

Duplicates are units which are represented more than once on the BR. There are two categories of duplicates. The first category is the duplication between IP and NIP units due to the inherently imperfect matching process of the PAYDAC file with the IP. When matching the new PDs with the IP, some of the non-matches are in fact already represented in the IP. The second category involves units in the IP that are represented more than once due to the imperfect matching process between the PD source and the Tax source.

Note that, in principle, duplication is not a problem within the NIP for two reasons. First, the NIP is made up of a list of PD accounts which have unique identifiers so that there is no duplication of accounts. Secondly, surveys select PD accounts in the NIP and the businesses owning the selected PD accounts are contacted to determine their list of PD accounts. That information is then used to adjust the weights at the estimation stage of the surveys. This works well as long as businesses do provide their complete list of PD accounts when contacted.

The impact of IP/NIP duplicates on survey estimates is to introduce an upward bias since the observed population is larger than the target population.

3.4 Misclassification

A unit is misclassified when its SIC code or SGC code is incorrect. This can be the result of the use of inaccurate information for coding or a coding error by clerical staff. Another reason for misclassification can also be that it is unknown that the unit has changed its class since the last time it was asked to provide one.

Misclassified units introduce coverage biases and/or a variance increase in survey estimates. When misclassified units are not included in the population of an industry specific survey due to their misclassification then a downward bias is introduced. For an industry specific survey, when misclassified units are incorrectly included in the survey population due to their misclassification, then an upward bias is introduced if these units cannot be detected via the data collection process. However, if the sample survey can detect misclassified units, then these units are treated as inactive businesses and thus the upward bias is not introduced but the variance is increased with the necessary use of domain estimation.

3.5 Erroneous Size Measure

An erroneous size measure for a unit is a measure which is far from its true value for that unit. For the large units in the IP, this type of error may happen as the main source for updating size measures is the Tax data, which is out of date when it becomes available. For the smaller units, the size measures are based on models which, as with any model, introduces a certain bias and variability in the predicted values.

For a measure of size, the impact of erroneous values on the survey estimates can be important. At the design stage, this may make the size stratification and the sample allocation less optimal. Also, at the estimation stage this may result in the presence of many outlier values in the sample, for units with a size measure much smaller than its actual size.

3.6 Inaccurate Contact Information

The contact information for a business is inaccurate when it disables an economic survey from effectively collecting its data. This may happen because the initial contact information available on the BR comes from the RC administrative data which are not always appropriate for survey collection purposes.

The problem with inaccurate contact information is that it can lead to non-response when the appropriate contact person cannot be reached in time. The presence of non-response means that a bias may be introduced in the survey estimates and also that their variance is increased. This is true whether imputation is used or not.

Inaccurate contact information can also lead to systematic response errors for complex businesses. Not contacting the right person in a complex organization may result in not getting all of the data required from that organization but collecting only for a part, that is the part known by the person contacted.

4. MEASUREMENT OF ERRORS

Essentially, the BR deficiencies to measure are coverage errors (i.e., duplication, extraneous units, missing units and units classified to the wrong survey population) and the accuracy of the different frame data items (i.e., SIC code, SGC code, measure of size, name and address). Ideally, these should be measured in terms of counts of units, to evaluate the cost of reducing the level of a given type of error, and in terms of measure of size to quantify the impact of these deficiencies on survey estimates. These measures should also be periodic in order to be able to observe improvement or deterioration in the data quality. This would allow action in a timely and appropriate manner as well as adjustment of survey estimates. These are the goals that the Business Register Division of Statistics Canada targets for quality measurement.

In this section, types of error that are difficult to measure are discussed, then methods that have been used, so far, to measure Business Register errors are reviewed and, finally, some results are presented and discussed.

4.1 Errors Difficult to Measure

There are two groups of units for which it is presently difficult to get measures of errors. These are the complex units and those incorrectly assigned to the OOS list.

a) Complex Units

There are no measures of error currently available for the complex IP units. All that is known, is that the Survey of Employment, Payrolls and Hours evidenced an increase in reported employment for those businesses for which recent BR profiles had been made. This can be seen as a measure of the impact of the BR missing sub-units within complex organizations.

The reasons for which there have not been any quality measurements yet are several. First, that the response burden is already large for these businesses as they are often selected with certainty in many surveys on top of the profiling activities. Secondly, the costs currently incurred for profiling these complex businesses are already important. Consequently it is not appropriate to devise a measurement tool that would increase burden and costs. The only viable approach would be to make use of the current reaction and cyclical profiling results. Work is required in that area.

b) Incorrectly OOS Units

To get an estimate of the number of incorrectly OOS units, the unbiased approach is to draw a sample from the set of OOS units, not linked to an IP unit, and to directly survey the sampled units (to verify if they are active employers and assess their SIC and SGC codes). One disadvantage with this approach is that it would require a large sample size to provide reliable estimates since the proportion of erroneously OOS units is expected to be very small. Also, it may be difficult and costly to try to contact a sample composed of mostly inactive units and then to assess whether they should be in-scope or not. There is no obvious approach to measure this undercoverage. The use of external sources should be considered.

4.2 Methods Used to Measure Errors

There are many methods that can be used to measure levels of error on the BR. A few have been used so far and are reviewed in the following paragraphs.

a) Quality Measurement Survey

Statistics Canada conducted a Quality Measurement Survey (QMS) to assess the quality of frame data on many occasions (see Lorenz and Laniel, 1992). The last one was in 1995 and it collected most of the data items on the BR for both the large units with a simple structure and for smaller units. The survey used a stratified random sample of 5,000 units. It was a telephone survey from the ROs and the collection was performed over 3 months. The response rate was about 90%.

b) Internal Source

As mentioned in section 2, there is a list of unclassified units which can be accessed. These units are all in the process of being surveyed either via the PD-20 form, the Business Activity Report or the Classification Survey. In other words, at some future point in time, all of these units are classified. One approach here to measure undercoverage is to select a reference period, to obtain a list of all the unclassified units in the category and to wait a number of months in order to get the vast majority classified. A previous study showed that the median time for classifying the unclassified units was eight months. This study suggests that after a year most of the unclassified units are finally classified. This approach leads to biased estimates if the units are not all classified.

c) External Sources

There are three external sources that can be used to estimate the undercoverage of the Business Register: the Survey of Employment Payroll and Hours, the Labour Force Survey and the Revenue Canada list of tax filers. The Survey of Employment Payroll and Hours surveys the unclassified PD accounts and thus provides an estimate of the number of employees working for those businesses owning the unclassified accounts. However, that estimate does not have an industry breakdown.

The Labour Force Survey provides an estimate of the number of self-employed workers which includes those with employees in their business. That estimate is an upper bound for the number of non-employer businesses which is also the number of employees in these businesses.

The list of tax filers provides a list of the self-employed filers which includes those with employees in their businesses. That list can be used to calculate an upper bound for the number of non-employer businesses, the number of employees in these businesses, as well as the total revenue of the non-employers.

d) Record Linkage Study

It is possible to estimate the duplication between the IP and NIP by linking a sample of NIP units to the IP units and then manually assessing the correctness of the links. Such a study was performed in 1992 (see Chun *et al*,1993). In that study a stratified random sample of 13,400 NIP units was selected.

e) Study of Model-Based Revenue

For the NIP businesses, the errors in the model-predicted Gross Business Income are evaluated annually by comparing the predicted value with the values of these variables from the Monthly Retail Trade Survey data. This study essentially consists of calculating correlations (see Dionne and Hawley, 1996).

4.3 Results

Over the past years, estimates of errors have been produced using the methods listed in the previous section. These numbers are presented and discussed in this section.

a) Missing Units

The number of unclassified businesses on the BR represents around 1.8% of the entire population on the Register. According to the Survey of Employment Payrolls and Hours, these units contribute 0.6% of the estimated number of employees within employer businesses. This small contribution can be explained by the fact that the unclassified units are usually new businesses of small size. Although there is no hard number about the contribution of the unclassified to the total revenue, it can be speculated that they contribute less than 0.6% of the total revenue of the employer businesses. This is due to the distribution of revenue being more skewed than the one for employment.

A study of the unclassified units was performed in 1995 and showed that the distribution of these units amongst the industries was similar to the distribution of businesses on the BR.

When pooling 1994 estimates of self-employed workers from the Labour Force Survey and the list of tax filers, the upper bound for the proportion of non-employers among all businesses can be approximated at less than 60%. In terms of number of employees, this represents less than 10% of the total for the entire business population. As far as revenue is concerned, it is estimated that the non-employers contribute to less than 1% of the total.

b) Extraneous Units

From the 1995 Quality Measurement Survey, the percentage of extraneous units found amongst the large and simple IP units is 14% and amongst the smaller NIP units 9.1%. The difference in rate can be attributed to the fact that the IP updating process (i.e., via profiling) has been slower than the NIP process (i.e., PD accounts not

remitting for 12 months) to identify dead units. Note that the two rates may be downward biased since it is likely that a larger proportion of inactive units than active units exists among the non-respondents.

c) Duplication

In 1991, the Quality Measurement Survey was conducted for a sample of NIP units (see Lorenz and Laniel, 1992), and it was then found that only 0.2% of the smaller units duplicated the large units. Note that the procedure to identify duplicates relied on the willingness of the businesses to provide the entire list of PD accounts owned. As this can be viewed as confidential information by some businesses, one could suspect that this duplication rate is downward biased.

The record linkage study of 1992 estimated that about 1.6% of the NIP units duplicated IP ones. This estimate of the duplication rate is larger than the one found in the 1991 survey and seems to confirm that respondents did not provide the required linkage information to the interviewers. However, it is also suspected that a number of false linkages have not been identified in that study.

d) Misclassification

The 1995 QMS also provided industry misclassification rates among the active units on the BR. These rates are presented in the tables below.

Table 1 presents estimates of percentages of units out of industry for 4 industry levels by size of business (i.e., IP and NIP). Note that these error rates at the 3 and 4 digit levels should be interpreted with caution. In many industries there is no requirement to code industries at the third or fourth digits as they require only a 2-digit or a 3-digit industry code.

The table indicates that the rate of units out of industry is very similar for large and small units. This suggests that, if we were calculating these rates as employment or revenue shares, the percentages would be in the same range.

Table 1. Estimates of percentages of units in incorrect industry

Category	IP single	NIP
incorrect industry division (2-digit group)	7.4%	7.5%
incorrect major group (2-digit)	10.4%	9.6%
incorrect industry group (3-digit)	15.9%	16.5%
incorrect industry class (4-digit)	18.7%	22.7%

The table shows that the average overcoverage rate of an industry division is 7.5%. This means that a survey which target population is solely in an industry division may significantly overestimate the number of businesses in its industry if the correctness of the industrial activity is not assessed for the sampled units. Note that these rates may be underestimated as the interviewer asked the respondent to confirm the description he got from the BR (i.e., a dependent approach). However, asking the respondent to provide a description without feeding him with the BR data may lead to overestimation as a new description may be incorrectly interpreted as a different industrial activity.

The 1995 QMS also estimated the proportion of missing units in each industry division. It was observed that the proportion is ranging between 1% to 15%. It should be noted that for an industry specific survey this source of undercoverage is likely the most important one.

The 1995 QMS provided geography misclassification rates among the active units on the BR. These rates are presented in Table 2 for two geographical levels (i.e., province and census metropolitan area/census agglomeration) by size of business (i.e., IP and NIP).

Table 2. Estimates of geography misclassification rates

Category	IP single	NIP
incorrect province	0.05%	0.16%
incorrect cma/ca	3.2%	2.2%

It can be observed that errors at the province level are very small and those at the subprovincial level are small.

e) Erroneous Size Measure

With the 1995 study comparing the predicted GBI to Retail Trade survey data, the following results were obtained. A straight correlation was first calculated and the value obtained was 0.29, which is fairly low. However, when removing 1% of the extreme values for the predicted GBI, the correlation went up to 0.61, which is a good correlation. The increase in correlation after removal of the extremes indicates that the model error has a relatively large variance.

f) Inaccurate Contact Information

The 1995 QMS provided error rates for the contact information among the active units. These are presented in Table 3 for the name and postal code by size of business (i.e., IP and NIP).

Table 3. Estimates of error rates for the contact information

Category	IP single	NIP
incorrect legal name	10.1%	25.6%
incorrect business postal code	21.6%	11.4%

For the legal name, the error rate is larger for the NIP than for the IP units. This certainly reflects the fact that the IP uses tax data (which is legal entity based) as a source of updates. For NIP units, the 1991 QMS gave a 12% error rate. The difference is due to the inclusion of spelling errors in the calculation of the error rate for 1995. The error rate on the postal code also reflects the difference in the source of updates. The address of the business is often different from the address of the legal entity but closer to the address of the operating entity. Note that it is not possible to evaluate the impact of contact information errors from the rates above. The reason is that, for a given survey, the real impact depends on its data collection procedures.

5. CONCLUSION

As we have seen in Section 2, the design and maintenance of the Business Register are complex as they involve the use of administrative sources and surveys to model the business world. This complexity leads to the types of error (coverage errors, classification errors and contact information errors) and sources of error as discussed in Section 3.

In Section 4, it was shown that it is not a trivial task to measure the level of these errors especially for complex units and missing units. It also showed that some of the errors are important in size. For example, the errors in the industry classification probably constitute an important component of biases and/or sampling error of survey estimates. When reviewing the allocation of the resources for the regular maintenance of the BR, the possibility of investing more resources for identifying the sources of errors and improving procedures for SIC coding should be considered. It should be noted that in the near future, the implementation of the 1997 Standard Industrial Classification (also called the North American Industrial Classification) will provide an occasion to improve the quality of SIC codes. From July 1997 to December 1998, about half of the businesses on the BR will be contacted to confirm their industrial activity.

There are some improvements that can be made to the quality measurement of the BR. For example, error measures should be developed for the complex units in order to obtain a complete picture of the BR quality. Also, procedures to link duplicates and assess linkages identified should be reviewed with the goal to improve them. Economic data should also be obtained for the duplicates found to weigh their importance in terms of revenue and employment. As well, the design of the QMS should be modified in order to be capable of estimating efficiently the errors in terms of business size (i.e., revenue and employment). This would be useful as estimates based on size could be used to adjust estimates of surveys that are industry specific.

6. REFERENCES

Chun, B., Hutchinson, D. and Patak, Z. (1993). Methodology Report of IP/NIP Duplication on the Business Register. Internal document, Business Survey Methods Division, Statistics Canada, Ottawa.

Clark, C. and Lussier, R. (1987). The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Colledge, M., Estevao, V. and Foy, P. (1987). Experiences in the Coding and Sampling of Administrative Data. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Colledge, M. (1987). Use of Administrative Data in the Business Survey Redesign Project. *Proceedings of the Statistics Canada Symposium on Statistical Uses of Administrative Data*, Ottawa.

Cuthill, I. (1996). The Canadian Business Register. Internal document, Systems Development Division, Statistics Canada, Ottawa.

Dionne, S. and Hawley, M. (1996). Testing of the Final 1993 Gross Business Income Tables. Business Survey Methods Division Internal Report, Statistics Canada, Ottawa.

Farrall, K. and Demmons, P. (1987). Profiling for Statistics Canada's Central Frame Data Base. Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

Konschnik, C.A. (1988). Coverage Error in Establishment Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Latouche, M. and Hidiroglou, M.A. (1987). Sample Size Determination and Allocation for the Monthly WRTS. Business Survey Methods Internal Report, Statistics Canada, Ottawa.

Lorenz, P. and Laniel, N. (1992). Measuring the Quality of the Business Register: Methodology and Results. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Tupek, A.R., Copeland, K.R. and Waite, P.J. (1988). Sample Design and Estimation Practices in Federal Establishment Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 298-303

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

ESTIMATION OF COVERAGE ERROR IN THE 1996 CENSUS OF POPULATION

Claude Julien[1]

ABSTRACT

One of the main uses of the Census is to serve as a basis for estimating the size of each province for the purpose of allocating tax transfers. Since 1991, estimates from coverage studies have been used to correct the basic census figures. Hence it is even more important to improve these studies so as to reduce the sources of sampling and nonsampling error. The Reverse Record Check is the most important coverage study in terms of its size and its impact on tax transfers. This article shows how we have improved this study so as to produce, with the same resources, more accurate estimates while reducing sources of potential bias.

KEY WORDS: Coverage error; Census; Nonsampling error; Matching.

1. INTRODUCTION

Since 1961, Statistics Canada has been publishing estimates of coverage error for the censuses of population held every five years. Those estimates are produced by means of the technique known as the reverse record check (RRC). Before 1991, those estimates served to inform census managers and the main data users about the quality of the coverage. In 1991, Statistics Canada decided to include these estimates in the population estimation program. To do so, it had to reduce the sources of errors that had affected the quality of the 1981 and 1986 RRCs, as described in Burgess (1988). In 1991, Statistics Canada almost doubled the size of the RRC sample, introduced major improvements in the processing of the data and, for the first time, produced official estimates of overcoverage.

Estimates of coverage error now have a direct impact on the division of federal tax transfers to the provinces. This new use of the estimates has had two consequences. First, the 1991 estimates have been analysed much more closely, and all sources of sampling and nonsampling error have been identified as being major. Second, in light of the sensitivity of the transfer formulas, sources of error must be further reduced. Thus, for the 1996 RRC, the survey design was altered in order to produce more accurate estimates at the provincial level with the same sample size as in 1991. On the other hand, owing to the demographic makeup of the 1996 sample, there is likely to be an increase in nonresponse, and hence in nonsampling error. To reduce this risk, we have further improved the processing of the data.

The four sections that follow describe the known or potential sources of error that characterize each operation of the RRC. The four main operations of the RRC are (1) survey design, (2) tracing of selected persons and data collection, (3) searching of census documents and classification of selected persons, and (4) weighting and analysis. The first operation consists in combining several data sources to create the population that should have been enumerated in the Census and selecting a sample from that population. The second operation consists in contacting the selected persons (SPs) after the Census and asking them to provide all the addresses at which they could have been enumerated. The third operation consists in finding the census document completed at each address in order to determine the number of times an SP was actually enumerated. This operation may provide new leads for contacting an SP, or it may call for collecting additional information. The last operation consists in weighting the sample, compensating for

[1] Claude Julien, Statistics Canada, 15-C R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

nonresponse, producing estimates and comparing them to independent sources. This operation enables us to identify problems that might make it necessary to review the classification of a group of SPs.

2. SURVEY DESIGN

2.1 Sampling frame

The sampling frame is constructed from the following six lists or frames:

Census	All persons enumerated in a province in the last census; for the RRC96, all persons enumerated in a province in 1991.
Births	All births that occurred during the intercensal period (frame constructed from vital statistics files).
Immigrants	All immigrants who were landed during the intercensal period (frame constructed from files of Citizenship and Immigration Canada).
Permit holders	All persons residing in Canada under a temporary student or employment authorization or Minister's permit, as well as persons claiming refugee status (frame constructed from files of Citizenship and Immigration Canada).
Missed	Sample of persons missed in the last census; for the RRC96, the group of persons who, according to the RRC91, were missed (there is no complete list of this group, but only a sample of these persons accompanied by their final sampling weight in 1991).
Health care file	All persons registered in the health insurance files of Yukon and the Northwest Territories (frame used only for these territories).

These frames are supposed to include all persons who should be enumerated in the 1996 Census without duplicates or overlap. Royce (1993) reports that these frames cover more than 99% of the target population. In other words, the coverage of the frames is not a major source of error. Nevertheless, certain deficiencies should be noted.

Some persons are not included on these lists. This is the case with Canadians who were abroad at the time of the last census and who have since returned to Canada. This group represents approximately 0.5% of the Canadian population, and there is no register on which it appears. This has been the case since the first RRC was conducted in 1961. For the same reasons, the frames used for the RRC96 do not cover those persons who were residing in one of the territories during the last census and who subsequently moved to a province. There are fewer than 4,500 persons in this group, and their contribution to the undercoverage estimate is less than 0.1%. Bureau, Julien and Provost (1995) contend that the decision to exclude this population is justified by the need to use a more recent source to measure the coverage error in the territories, as well as by the cost of processing a sample from the census frame drawn from the territories.

Most of the duplicates come from the census frame. In 1991 for the first time, the number of duplicates present in the census could be estimated, as a result of quality studies. We will be able to use these estimates to correct the weights of the units selected from the census frame and eliminate the contribution of duplicates.

Two sources of overlap between the frames deserve special attention. First, some landed immigrants and permit holders were residing in Canada at the time of the last census under a temporary permit. These persons are part of both the census frame and the missed frame. This potential source of error was introduced into the RRC by the inclusion, for the first time in 1991, of the population of permit holders in the population targeted by the census. Using administrative data of Citizenship and Immigration Canada (CIC), we can identify most of these selected persons in the sample. The other source of overlap is classification errors from previous RRCs. A person who was classified as missed in the preceding RRC

but who was in fact enumerated constitutes an overlap between the missed frame and the census frame. This source of error is inherent in the RRC methodology. To counter it, we make operational improvements to each RRC. Nevertheless, it will always be present, and we assume that its effects are negligible and are even largely cancelled out by other sources of errors.

Lastly, although complete nonresponse to the census is a fairly marginal phenomenon, for the RRC it is a potentially major source of error. In the case of a complete nonresponse to the census, we use special procedures in the collection and processing of the data in order to add persons through imputation. In the 1991 Census, the number of persons added increased by 92% over the previous census. A bias in the number of persons imputed or their characteristics directly affects the quality of the census frame.

2.2 Sample selection

The survey design is the main methodological improvement introduced for the RRC96. Except for the missed frame, the sample for which is obtained from the RRC91, we draw a sample of persons from a single-stage survey design with demographic stratification and an optimum distribution based on the historical under-coverage rate and strata size. The previous RRCs used a two-stage design with geographic stratification in the census frame. As a result, the 1996 design will produce more accurate estimates, for as shown by Boudreau and Germain (1990) and Royce, Germain, Julien, Dick, Switzer and Allard (1994), coverage error is a phenomenon that is more demographic than geographic in nature. Using this survey design, we selected more persons who are difficult to enumerate. By observing more cases of missed persons in the sample, which is what we want to measure, we will produce more accurate estimates of undercoverage at the provincial level.

Table 1 - Sample sizes comparison

STRATUM	proportion in sample RRC96	RRC91	ratio 96/91	response rate
MEN 30-39 NON MAR	3.1%	1.4%	220%	91%
MEN 40-59 NON MAR	2.9%	1.3%	220%	93%
MEN 20-29 NON MAR	7.8%	4.2%	185%	93%
WOMEN 30-39 NON MAR	2.4%	1.3%	184%	93%
PERMIT HOLDERS	2.3%	1.3%	180%	79%
IMMIGRANTS	5.1%	2.9%	176%	86%
WOMEN 20-29 NON MAR	5.2%	3.0%	173%	95%
MEN 40-59 MAR	5.0%	6.8%	74%	98%
WOMEN 15-19	4.4%	6.0%	72%	96%
WOMEN 30-39 MAR	3.4%	5.0%	68%	98%
MEN 60+ MAR	2.7%	4.1%	66%	99%
MEN 20-29 MAR	1.8%	2.8%	65%	96%
WOMEN 60+ MAR	2.0%	3.2%	63%	99%
WOMEN 40-59 MAR	3.9%	6.4%	62%	98%
WOMEN 20-29 MAR	2.0%	3.7%	53%	97%

Table 1 shows the groups of persons for which the sample size in 1996 either increased or decreased the most in relation to the RRC91. This same table also shows the non-response rate observed in the RRC in 1991 for these groups. The persons who are difficult to enumerate are young and mobile and therefore hard to trace and classify accurately. The dilemma that faces us here is that by reducing sampling error, we risk increasing nonsampling error. For this reason, we are introducing a few major operational improvements to control nonresponse and the attendant potential for error.

3. TRACING

In terms of potential error and its effects, tracing is the most important operation and the most difficult one to improve. On the basis of the data from the various selection sources, we must contact each SP as soon as possible after the census. Yet some data sources, such as the census frame, are five years old. The

challenge, then, is a major one. By using administrative data and drawing on the talent of the interviewers in our regional offices, we manage, from one RRC to the next, to contact more than 96% of SPs.

3.1 The problem of persons not traced

Despite this excellent result, even the low percentage of persons not traced can cause significant errors in estimating an undercoverage rate of 3%. The problem is to establish the undercoverage rate that is to be "imputed" to the persons not traced. We expect their rate to be higher than that of traced persons, since a missed person is also a person who is difficult to trace. The RRC91 "imputed" a rate of 12.7% to persons not traced, as compared to an estimated rate of 9.3%[2] for traced persons. Is this difference accurate? As Burgess notes, "To the extent that it is not correct, there may be some distortion in provincial estimates of undercoverage as well as a bias in overall estimates of undercoverage." Burgess hypothesizes that persons not traced tend to have moved from one province to another since they were selected (i.e., they are interprovincial migrants). Interprovincial migrants have a higher undercoverage rate (9.9%) than the general population (3.4%). If Burgess's hypothesis is correct, the RRC would tend to underestimate undercoverage.

Table 2 shows the results of a study in which we simulated the impact of various undercoverage rates imputed to persons not traced. The first part of the table shows the impact on the estimate of the number of missed persons. An imputed rate between 15% and 20% generates a bias greater than one standard deviation only for the estimate for all ten provinces. It takes an imputed rate greater than 21.4% (instead of the 12.7% actually imputed) for the estimate in the largest provinces (QC, ON, AB et BC) to differ by more than one standard deviation from the final estimate from the RRC91. The second part of the table shows the impact on the estimate of each province's relative share. This estimate corresponds to the census count adjusted by the estimate of missed persons in the province divided by the sum of these adjusted counts for all ten provinces. As may be seen, even with an imputed rate of more than 33%, the impact on the relative share is less than one standard deviation. These results confirm what Burgess reported, namely that a slightly higher imputed rate can cause major bias at the national level. By contrast, at the provincial level, the bias is low in relation to sampling error, and it therefore has little impact on the tax transfers that are determined by each province's relative share.

3.2 Control of the not traced rate

Efforts to reduce the number of persons not traced begin with the selection of the sample. In this operation, we record information relevant to the tracing of the SP and any other person residing with the SP or associated with the SP (e.g., the sponsor of a landed immigrant). This information is quite useful, since it is easier to trace a household than a lone individual. The census, births and missed frames naturally provide us with households. On the other hand, it is harder to retrieve households from the CIC administrative data and the territory files.

The crucial stage of the tracing process consists of matching SPs and the members of their households to various administrative data sources in order to find their most up-to-date address prior to the census. The data sources consulted are taxation data, CIC data and telephone directories available in electronic format. Table 3 shows the proportion of SPs for whom at least one member of the household was matched to the taxation data. The RRC96 exhibits the best results for SPs in the census, births and missed frames. For the first time, we also matched immigrants. In addition, in 1996 the CIC administrative data give us quite recent addresses for most permit holders.

---

[2] In the RRC91, nearly half of the SPs enumerated were classified without tracing, in a special operation incorporated into the census processing. This explains the high rates in comparison to the overall rate of 3.4%. This operation will not be carried out in the RRC96.

Table 2 - Impact of imputation rate of non traced SP's

Impact on estimation of the number of non censused people

Province	RRC91		imputation rate of non traced SP's					
	estimate	standard error	13.7%	14.7%	15.6%	21.4%	28.0%	33.2%
Newfoundland	17747	1771	33	64	95	293	530	730
Prince Edward Island	2928	306	57	61	65	89	117	140
Nova Scotia	26072	3379	11	73	132	500	913	1240
New Brunswick	32360	3260	123	243	360	1102	1986	2725
Québec	281444	14487	1443	2840	4196	12681	22507	30471
Ontario	521880	31328	3423	6727	9921	29661	51963	69595
Manitoba	32119	4125	149	293	433	1309	2321	3139
Saskatchewan	32591	3329	129	254	374	1121	1971	2653
Alberta	87826	7202	705	1393	2063	6341	11469	15759
British Columbia	144016	8441	915	1803	2667	8109	14479	19686
Total	1178983	36536	6988	13751	20306	61206	108256	146138

Impact on relative share of each province (in %)

Province	RRC91		imputation rate of non traced SP's					
	estimate	standard error	13.7%	14.7%	15.6%	21.4%	28.0%	33.2%
Newfoundland	2.1	0.007	0.000	-0.001	-0.001	-0.003	-0.006	-0.008
Prince Edward Island	0.5	0.001	0.000	0.000	0.000	-0.001	-0.001	-0.002
Nova Scotia	3.3	0.012	-0.001	-0.001	-0.002	-0.005	-0.009	-0.012
New Brunswick	2.7	0.012	0.000	0.000	-0.001	-0.002	-0.003	-0.004
Québec	25.3	0.047	-0.001	-0.002	-0.003	-0.010	-0.017	-0.023
Ontario	37.4	0.091	0.003	0.006	0.008	0.024	0.041	0.053
Manitoba	4.0	0.015	0.000	-0.001	-0.001	-0.004	-0.007	-0.009
Saskatchewan	3.6	0.012	0.000	-0.001	-0.001	-0.004	-0.007	-0.009
Alberta	9.3	0.026	0.000	0.000	0.001	0.002	0.005	0.008
British Columbia	12.1	0.030	0.000	0.001	0.001	0.003	0.005	0.007
Total								

Note: the imputation rate for RRC91 is 12.7%

After obtaining the most recent addresses possible, we send the information to the regional offices, where interviewers try to contact the SPs by telephone at one of the addresses supplied or by tracing them with public sources acquired by each office. Owing to the improvements that we made to the survey design and the selection operations, we are able to begin this operation earlier than in past RRCs. This is especially true for the samples of immigrants, permit holders and the territories. We hope that this will have a positive impact on the tracing rate. In addition, concurrently with the tracing being carried out in the field, we are continuing to consult more recent administrative data (for example, the taxation data for 1995 and 1996) in order to obtain other addresses and send them, as required, to the regional offices.

3.3 Collection

The interviewers try to directly contact each SP or a relative or spouse to have him or her complete a survey questionnaire. During the interview, the respondent reports all addresses at which the SP could have been enumerated and the characteristics of the household in which the SP was living. These addresses include the current residence, the address on Census Day, the previous address (if the SP has move recently), any addresses at which the SP temporarily resided during the census collection period, and the address of any relatives or friends of the SP who might have included the SP on their census questionnaire. By combining the selection source, administrative sources and collection, we can obtain up to 10 addresses for a given SP. On average, we expect to have at least three. The characteristics of the household include the names, sex and dates of birth of those persons who were living with the SP at the latter's usual address on Census Day. Lastly we obtain information such as marital status, language spoken and dwelling type, so as to establish a better profile of the missed person. The next operation consists in entering this information and starting the search for the census document completed at each address obtained.

Table 3 - Tax data linkage

	matching rate	
	RRC91	RRC96
FRAMES		
CENSUS	85%	92%
BIRTHS	73%	80%
NOT CENSUSED	70%	80%
IMMIGRANTS	no attempt	72%
PERMIT HOLDERS	no attempt	no attempt
CENSES FRAME STRATA		
MEN 30-39 NON MAR	76%	87%
MEN 40-59 NON MAR	72%	86%
MEN 20-29 NON MAR	85%	92%
WOMEN 30-39 NON MAR	75%	89%
WOMEN 20-29 NON MAR	87%	93%
MEN 40-59 MAR	86%	94%
WOMEN 15-19	89%	94%
WOMEN 30-39 MAR	91%	95%
MEN 60+ MAR	82%	94%
MEN 20-29 MAR	89%	94%
WOMEN 60+ MAR	78%	93%

4. SEARCHING AND CLASSIFICATION

The searching and classification operation consists of processing the information obtained from tracing and collection, in order to determine whether the SP should have been enumerated and how many times he or she actually was enumerated. Among the persons who should not have been enumerated are deceased persons and persons who left Canada before Census Day. For the other SPs, it is necessary to check the census document corresponding to each available address. By census document, we are referring to visitation records, fully or partially completed questionnaires and forms indicating the reason for a complete non-response (non-contact, refusal, etc.).

4.1 The challenge posed by searching and classification

For an SP to be classified as enumerated, he or she must be found to be "clearly" enumerated on one of the documents checked. For an SP to be classified as missed, it is necessary to be certain of having found documents corresponding to all the addresses at which the SP could have been enumerated but at which he or she was not "clearly" enumerated. An SP is "clearly" enumerated at an address when his or her name or characteristics (sex and date of birth) appear on a questionnaire.

By contrast, an SP cannot be "clearly" enumerated at an address at which a census procedure has imputed a number of persons because of a complete nonresponse. In this case, as well as in others in which the information and results obtained are not specific enough to make a definite classification, the SP is "not

classified" and serves to increase the nonresponse in the RRC, just as persons not traced do. The percentage of SPs not classified is low (1.1%), but these persons pose the same problem as SPs not traced, namely, what undercoverage rate should the "imputed" to them? The RRC91 "imputed" a rate of 12% to them, as compared to an estimated 9.5% for classified persons.

Classification errors cause bias. On the one hand, it is difficult to incorrectly report an SP as enumerated. On the other hand, we can never be certain that an SP is missed without consulting all the census documents. It may be that the respondent has not given us all the addresses or that we have not correctly processed all the addresses supplied. Therefore classification is characterized by an unavoidable bias that we must reduce as much as possible.

Reducing nonresponse and classification error poses a major operational challenge. We expect to have to search the documents corresponding to more than 150,000 addresses obtained in the tracing and collection stages. This operation is made especially difficult by the fact that names and addresses are not captured in the census. We meet this challenge by using computer processing to reduce the number of documents to be checked in order to classify an SP. This will enable us to devote more time and resources to correctly processing the SPs who are eventually found to be either missed or not classified.

4.2 Geographic coding and matching

Computer processing is composed of two stages: geographic coding and matching. Geographic coding consists in assigning to each address a search area in which it should be located. This area is made up of from 1 to 10 enumeration areas, and it is basically obtained by converting the postal code into a set of enumeration areas. This stage was put in place for the RRC91 and was improved for the RRC96.

The matching stage consists in matching the household in which the SP was living to all households present in the census database in the search area identified by means of geographic coding. Using only the characteristics (sex and date of birth) of the SP and the other members of the household, matching finds all households in the database that resemble the SP's household. If one of the households in the database "strongly" resembles that of the SP and if the SP "participates" in this resemblance, we can classify the SP as enumerated without having to consult the census questionnaire of the household in question to confirm the SP's presence.

In one study, the primary goal of which was to evaluate a method of detecting overcoverage in the census, Bernier (1995) showed that two households in the database that were found to have several persons with the same sex and date of birth were necessarily composed of the same persons. For example, two households that have at least three persons in common are necessarily the same household enumerated twice. Another way to interpret this result is to state that in as limited an area as the one produced by means of geographic coding, households of at least two persons are unique in terms of the combination of sex and date of birth characteristics of the members. Thus, in the RRC96, we will automatically classify an SP as enumerated when we match the characteristics of that person as well as those of the members of his or her household to the same household in the census database.

We applied this method to the data of the RRC91. It enabled us to classify more than half of the enumerated SPs automatically without checking census documents. However, the method is not error-free. It identified two persons as enumerated when in fact they were not. Since there were 2,341 missed persons, the error rate is less than 0.1% and is negligible in comparison to the benefits that the method offers: consistent, systematic and orderly classification and better use of resources.

In addition, matching identifies households in the database that are largely similar to the SP's household, but are not similar enough to be classified automatically. In these cases, the verification task can be reduced to checking the census questionnaire of one or two households. The strategy that consists in matching the SP's household, and not merely the SP, to the census frame is the most significant improvement made in the processing of the RRC96 data. Using this strategy, we will be able to process a number of SPs very quickly, leaving us more time and resources to conduct, verify and correct the processing of those SPs who are eventually found to be either missed or not classified.

4.3 Monster matching and additional searches

This operation, new in the RRC96, supplements geographic coding and matching. It consists in matching the SP's household with the census database so as to identify all households in the database that have at least two persons in common with it. This operation uses an algorithm described in Julien and Mayda (1995). It serves to classify SPs as enumerated at addresses that are not provided in the collection process or at addresses that are incorrectly processed in the searching process. Monster matching is also a simple and inexpensive method of solving address problems and finding new leads for correctly classifying the SP. It is also used to reduce nonresponse and classification error without having to re-contact the SP.

Prior to final classification, persons not found to be enumerated after the searches conducted are the object of additional searches. We try to obtain other addresses from administrative sources. The results of monster matching can provide us with new leads for contacting the SPs. Finally, all missed SPs and some SPs not classified are contacted again, and at that time we confirm and clarify their address on Census Day and try to obtain other addresses. All this new information is processed by going through most of the operations already described.

In summary, the searching and classification operation consists in processing a sizable amount of information under difficult conditions. It is important to carry out this operation as correctly and thoroughly as possible. On the one hand, classifying persons as missed when in fact they are enumerated leads to an overestimate of undercoverage. On the other hand, if some persons cannot be definitely classified, this can result in an underestimate of undercoverage. To reduce the incidence of these errors, we have developed automatic classification procedures that quickly process a number of SPs with negligible risk of error. With the same resources as in 1991, this enables us to devote more effort to correctly classifying the other SPs. However, complete nonresponse to the census makes a few SPs impossible to classify and increases nonresponse in the RRC without there being anything that can be done about it. If complete nonresponse to the census increases further, it will be necessary to consider new methods of dealing with this situation.

5. WEIGHTING AND ANALYSIS

Weighting is composed of three stages. During the selection process, we assign each SP an initial weight equivalent to the inverse of that person's sampling ratio. For the missed frame, the initial weight is equivalent to the final weight from the last RRC. At this stage, we select the sample in five replicates, to allow for variance estimation. Then this weight undergoes two corrections: one to take account of nonresponse and the other to improve the representativeness of the sample in relation to the frames from which it is drawn (poststratification).

In past RRCs, these corrections were made within groups defined on the basis of the survey design (frames and strata) and also, if possible, on the basis of demographic and mobility characteristics, while also taking account of the division into replicates. The fact that the census frame was geographically stratified greatly limited the demographic groupings that could be made within it. The entirely demographic stratification of the survey design of the RRC96 allows for finer groupings to correct for nonresponse, and it practically eliminates the need for the second correction. We are currently looking into ways to make greater use of the tracing and searching results with a view to making better corrections.

The final weights are used to estimate undercoverage, of course, but also to estimate the number of enumerated persons, deceased persons, persons who have left Canada, etc. These estimates are compared to outside sources in order to identify potential problems. The undercoverage estimates are compared to errors of closure derived from demographic projections and estimates from past RRCs. The profile of missed persons is analysed jointly with the Demography Division. In 1991, we conducted a thorough review of missed persons in New Brunswick and in certain age groups. In the case of New Brunswick, the RRC estimates proved to be accurate, whereas the estimate for girls aged 0 to 4 was adjusted to ensure demographic consistency (male/female ratio).

The estimates of enumerated persons by province and by demographic group are compared to the census counts. This enables us to analyse the weighting and assess whether corrections for nonresponse properly distribute the weight of persons not traced and persons not classified among enumerated persons and missed persons. Historically, the RRC has achieved a good estimate of the enumerated population by province. At the level of demographic groups, the comparison becomes difficult because of greater sampling error and the increase in item nonresponse and complete nonresponse to the census, which creates distortions in the demographic distribution of the enumerated population.

The estimates of deceased persons by province are compared to the counts obtained from Vital Statistics. Suspect estimates lead to a review of the information and a re-evaluation of the results obtained. At this stage, we are proceeding to match the deceased person with the register of deaths in order to confirm the classification results. Similarly, estimates of the number of persons who have left Canada are compared to estimates based on demographic models.

In summary, weighting is an important operation, since it serves to correct the weights to compensate for nonresponse. It is essential for the adjustment to be made within groups made up of persons with similar characteristics, such as sample frame, age, marital status and sex, which influence their enumeration. All these characteristics are part of the survey design of the RRC96, which will thus allow for more appropriate groupings than in 1991. The procedures used in tracing, searching and classification serve to detect and deal with problems at the microdata level, while comparison of the RRC estimates with outside sources serves to detect problems at the macrodata level. This analysis can lead to a complete review of all documentation accumulated for a group of SPs, for purposes of correcting or confirming the RRC estimates.

6. CONCLUSION

Since 1961, coverage studies of the Census of Population have been providing census managers and users with information on the quality of coverage. Since 1991, these studies have also had a direct impact on the finances of the provinces, since they yield estimates that are used to correct the census counts and produce estimates of the population of each province. It is therefore essential to review the methodology and processing procedures in order to further reduce sampling and nonsampling errors. For the 1996 RRC, we have carried on with improvements that began with the 1991 RRC. First, we have changed the survey design to obtain more accurate estimates with the same sample size. Second, we have improved the tracing, searching and classification procedures in order to reduce nonresponse and the other processing errors that can cause bias. We are currently focussing our research efforts on weighting and adjustment for nonresponse. We plan to use the new survey design and the processing results to make better groupings for distributing the weight of SPs who are not traced or not classified.

7. REFERENCES

Bernier, J. (1995) « Étude pilote: estimation du surdénombrement détecté par appariement automatique », Internal Report, Statistics Canada.

Boudreau, J.-R. and Germain M.-F. (1990) *User's Guide to the Quality of the 1986 Census data: coverage*, catalogue 99 - 135E, Statistics Canada.

Bureau, M., Julien, C. and Provost, M. (1995) « Les études de mesure de l'erreur de couverture dans les territoires en 1996 », Internal Report, Statistics Canada.

Burgess, R. D. (1988) « Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population », *Survey Methodology*, Vol. 14, No. 2, Statistics Canada, pp. 137-156.

Julien, C. and Mayda, M. (1995) « Improving Census Coverage Error Measurement Through Automated Matching », *Proceedings of the Survey Methods Section, American Statistical Association*, pp. 849-854.

Royce, D. (1993) « Comments on Documents from Bureau de la Statistique du Québec Concerning Adjustment of the Population Estimates for Net Undercoverage », Internal Report, Statistics Canada.

Royce, D., Germain, M.-F., Julien, C., Dick, P., Switzer, K. and Allard, B. (1994) *Coverage - 1991 Census Technical Reports,* Catalogue 92 - 314E, Statistics Canada.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

WHAT IS THE ROLE OF DEMOGRAPHIC ANALYSIS IN THE 2000 UNITED STATES CENSUS?

J. Gregory Robinson[1]

ABSTRACT

Demographic analysis (DA) is a well-developed coverage measurement and evaluation program in the United States. DA has served as the standard for measuring coverage trends in recent censuses and differences in coverage by age, sex, and race at the national level. In this paper, we explore the role that demographic analysis can play in the 2000 census:

* Should DA be only a coverage evaluation tool in support of the survey-based coverage estimates (CensusPlus or Dual System Estimation) used in the Integrated Coverage Measurement (ICM) operations, or

* Should DA be formally integrated with the survey estimates into the ICM coverage measurement process, drawing from the particular strengths of the demographic approach?

The role of DA should be based on the balancing of the strengths and limitations of the demographic method and of the survey-based coverage estimates. We believe that demographic analysis can play an important and expanded evaluation role in the 2000 census. DA also has the potential to enhance the ICM coverage measurement in the areas where DA is strong and the survey estimates have been weak--(1) the measurement of undercoverage of adult Black men and (2) the production of detailed age, sex, and race estimates that are both longitudinally and internally consistent. By integrating the DA results into the 2000 ICM, the age-sex-race differences between DA and survey estimates will be reconciled before producing the one-number census estimates.

KEY WORDS: Demographic analysis; Coverage evaluation; Undercount.

1. INTRODUCTION

One of the goals of the 2000 census is to reduce the differential undercount and cost of the census with the use of sampling and estimation. We will use statistical techniques, and administrative records where possible, to estimate the number and types of people missed in the census. We will do this with the Integrated Coverage Measurement (ICM) program. The missed persons will be added to produce a "one-number" census total by the December 31, 2000 release deadline.

The ICM program was tested for the first time in the 1995 census test. We tested two coverage measurement techniques--CensusPlus (CP) and Dual System Estimation (DSE). These are sample survey-based estimates, involving case-by-case matching of persons in an independent survey with persons in the census (see Mulry and Singh, 1995, for a description of the CensusPlus and DSE methodology). In the 1990 coverage measurement program, the survey estimates were based on the Post Enumeration Survey (PES) (see Hogan, 1993). The 1990 PES was a dual-system estimation technique.

The Census Bureau has another coverage measurement and evaluation program--Demographic Analysis. Demographic analysis (DA) represents a macro-level approach to measuring coverage, where analytic estimates of net undercount are derived by comparing aggregate sets of data. The demographic approach differs fundamentally from the survey estimates, which represent a micro-level approach.

---

[1]J. Gregory Robinson, Chief, Population Analysis and Evaluation Staff, Population Division, United States Bureau of the Census, Washington, DC 20233-8800.

The method of demographic analysis relies heavily on aggregate administrative records, which are essentially independent of the census. The estimates for the population below age 65 are derived by the basic demographic accounting equation:

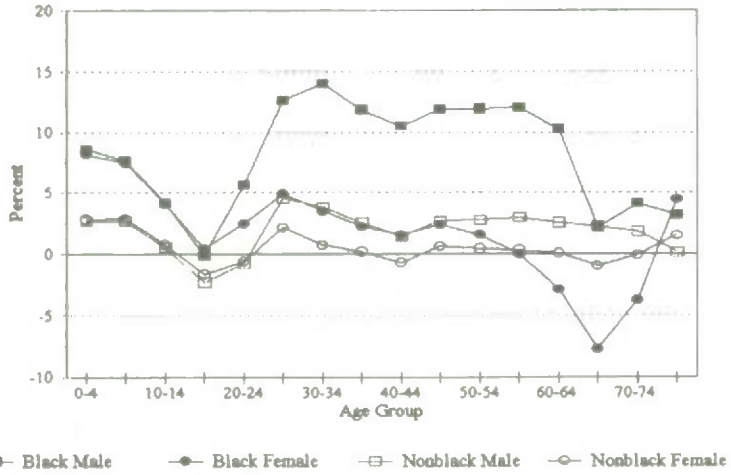
Population = Births - Deaths + Immigrants - Emigrants

Aggregate Medicare data are used to estimate the population aged 65 and over.

The estimation process involves a number of assumptions about the completeness of the administrative data used to develop the demographic estimates. Also, since there are no records for some population groups such as undocumented immigrants, the size of some groups must be estimated (see Robinson et al, 1993, for a discussion of the 1990 demographic results and Himes and Clogg, 1993, for an excellent overview of the use of demographic methods).

Demographic analysis as a tool for coverage evaluation has been well developed over time. The national demographic estimates have become the benchmark for assessing differences in coverage by age, sex, and race. Figure 1 displays demographic undercoverage rates for 1990--the figure shows the relative high undercount of Black children and adult Black men. The most notable pattern is the high levels of undercount of Black men between ages of 25 and 64, where the estimated undercount ranges from 10 to 15 percent. A principal goal of Census 2000 is to reduce these differentials. In keeping with the theme of this conference, this is a nonsampling error we are trying to reduce in the census.

Figure 1. Percent Net Undercount: 1990 by Race, Sex, and Age



2. THE POSSIBLE ROLES FOR DA IN 2000

2.1 Coverage Evaluation versus Coverage Measurement

How can demographic analysis be used in Census 2000, where we plan to release a "one-number" census? DA can serve one or two roles in 2000:

(1) As a **coverage evaluation** tool only, in support of the CensusPlus or DSE-driven ICM operations.

In 1990, DA served as a coverage evaluation program. It was used to evaluate the quality of the PES results, and provided important historical benchmarks (1940-1980) to assess completeness of coverage in 1990. Research conducted over the past four years demonstrates the DA can play an expanded evaluation role in the 2000 census, including the use of subnational DA benchmarks (see Robinson, 1994 and Robinson and Kobilarcik, 1995).

(2) As a coverage evaluation and an "active" **coverage measurement** program, where the demographic coverage estimates would be **integrated** with the CensusPlus or DSE survey estimates.

This "best set" would serve as the ICM standard for producing the one-number 2000 census products. In 1990, the PES estimates were used exclusively as the coverage measurement vehicle for any adjustment of the 1990 census counts. DA estimates were not used, because we believed the limitations of DA at that time (e.g., no geographic detail, uncertainty of the estimates) offset its strengths (e.g., independence, internal consistency).

2.2 Strengths and Limitations of DA

Should DA be only an evaluation tool in 2000--or should it also play an active integrated role in the ICM coverage measurement operations? These decisions will depend on how we can minimize its limitations and more clearly maximize its strengths. In the following review, we will identify where the strengths or limitations have changed since 1990 to build a stronger case for the integration of DA.

2.2.1 The Limitations of DA

1. **Lack of geographic detail**--Independent DA estimates in full age-sex-race detail are not available below the national level. For coverage measurement purposes in 2000, the survey-based estimates would remain the principal vehicle for the subnational ICM estimates.

Since 1990, extensive research has been conducted to develop "subnational" DA benchmarks of coverage, mainly for ages under 18 and 65+ for States and large county areas. For the younger ages, birth and death data are readily available and school enrollment data can provide an independent source for measuring the school- aged population and estimating migration. Administrative Medicare data are an excellent independent source for the population 65 and over. Further, sex ratio analysis provides clues about coverage differentials for ages 18-64 (See Robinson, 1994). Finally, we are developing a housing unit estimates program (for States and counties), which may ultimately be integrated with and strengthen the population estimates.

So there is a new geographic dimension to the demographic program, which can serve as an important evaluation tool in 2000 to compliment the survey-based ICM activities. We successfully used DA to evaluate the CensusPlus and DSE results in the 1995 test sites of Oakland, CA, Paterson, NJ, and six rural parishes in Louisiana (Robinson, 1996).

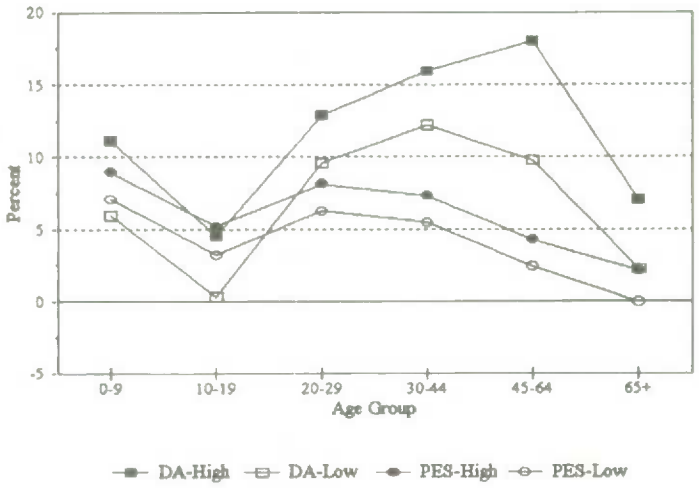
2. **Limited race/ethnic detail**--The principal DA race categories are Black and Nonblack. Although research is being conducted to produce DA estimates for Hispanics and Asians, these measures would not be as reliable as those for Blacks and Nonblacks. The CensusPlus/DSE would provide the coverage measurement standard for Hispanics, Asians, and American Indians (as well as important classifications by tenure).

3. **Inconsistencies in race classification**--The DA estimates of net undercount will be biased if persons who are classified as Black in DA are reported as another race in the census. We need to conduct more research to assess the degree of inconsistency and identify ways this "classification error" can be minimized. Also, the effect of a multiracial designation in the census race question for 2000 needs to be considered.

4. **Uncertainty in the DA estimates**--The principal concern of the DA estimates in 1990 was the uncertainty of the measured undercounts. For the first time, the 1990 DA estimates were accompanied by statistically-based measures of uncertainty (Das Gupta, 1991). The results demonstrated the DA estimates were subject to considerable uncertainty in measured undercount levels (see Figure 2 for 95 percent confidence intervals for the 1990 DA and PES undercount estimates of Black men). Nonetheless, it is clear that the demographic estimates

of percent undercount for Black adult males remain relatively high under any reasonable "uncertainty" assumption. The "lowest" alternative estimate for Black males is above 8 percent for each broad age group between 20 and 64. And these lowest DA estimates were consistently higher than the comparable PES estimates that included uncertainty bounds (see Adlakha et al, 1991).

Figure 2. Undercount Confidence Intervals Black Males: 1990 DA and PES



It is important to note that the DA estimates are subject to less uncertainty in terms of measuring differences in coverage according to age, sex, and race. This property--that demographic analysis provides better measures of coverage differences rather than absolute coverage levels--is attributable to the fact that many of the errors in the estimates are consistent and hence tend to "cancel" in comparisons across sex, race, and time. This particular strength could be exploited in 2000. For example, the DA sex ratios (ratio of males to females) are less error-prone than the DA undercount estimates themselves.

2.2.2 The Strengths of DA

Demographic analysis possesses certain advantages over the survey-based approach that can be utilized in a comprehensive ICM system for 2000. Some of these strengths, while existing in the 1990 census setting, gain "standing" in the cost-conscious, one-number census environment of 2000:

1. **Low cost**--With the reduction of cost an important goal of the 2000 census, the relative low cost of the DA program becomes very attractive. DA is very cheap because it draws extensively from the Census Bureau's ongoing population estimates program. Even with a stepped-up research program, the DA method is much less expensive than the survey-based approach.

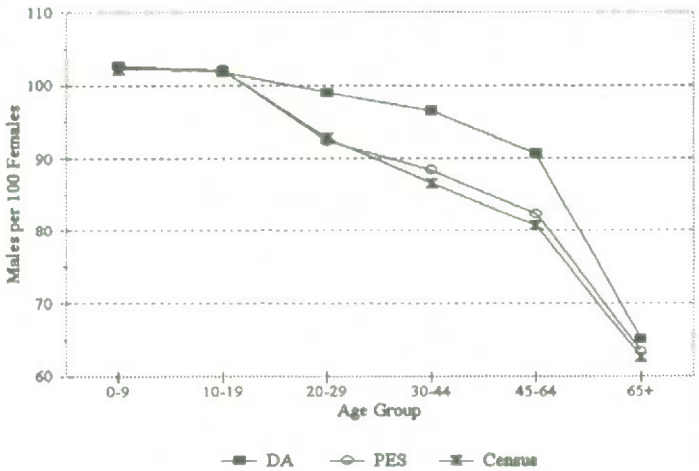
2. **Operational feasibility**--The DA method is battle-tested in previous censuses, with continued improvements in data and techniques and results available for review. The CensusPlus technique is still in the testing phase, in fact, it encountered unforeseen problems in the 1995 test. The DSE approach faces a very tight critical path to produce results by the December 31, 2000 deadline. The independent, administrative record-based DA estimates would provide a back up if the CensusPlus or DSE encounters problems.

3. **Timeliness**--Since field operations or census matching aren't involved, the DA estimates will be available in 2000 before the CensusPlus or DSE. First, independent housing unit benchmarks could evaluate completeness

of the Master Address File even before the 2000 census begins. Second, DA population estimates can give important readings on the differential undercount in the "pre-ICM" counts (e.g., July-August of 2000). For example, the indication from low sex ratios of large relative undercounts of adult Black men (like in previous censuses) would stress the importance of the ongoing ICM operations. Of course, the DA estimates will also be available to immediately evaluate the survey estimates when those are ready (October-November 2000).

4. **Independence**--Since DA is based largely on aggregate administrative records, it provides an independent basis to validate the ICM survey estimates. In 1990, the independent DA undercount estimates (1.85 percent) were used to validate the overall PES estimate (1.58 percent). The detailed DA estimates indicated, however, that the PES significantly understated the net undercount of adult Black men--the well-known "correlation bias" problem. For example, Figure 3 shows how the 1990 PES sex ratios for Blacks are much closer to the implausible census sex ratios than to the DA ratios. Even after taking into account the measured uncertainty of the DA and PES estimates, the DA sex ratios are significantly higher than the PES or census ratios.

Figure 3. 1990 Expected Sex Ratios: Comparison of DA and PES to Census



For 2000, we are looking at ways to integrate the DA results (such as sex ratios) in the ICM to minimize this problem. Here, DA would clearly be serving a dual coverage evaluation and coverage measurement role (see Wolter, 1990, and Bell, 1993, for research on the use of DA sex ratios in coverage estimation).

5. **Internal consistency**--The foundation of the demographic method is the logical and longitudinal consistency of the underlying demographic data. DA follows the demographic process of population change as it occurs, starting with births, then incrementing or decrementing cohort size with subsequent information on mortality and net migration. The estimates created for 2000 from this process will be longitudinally and internally consistent. The time series linkage of the DA estimates (for multiple censuses) provide a consistent basis to assess the plausibility of the demographic estimates themselves. On the other hand, the survey estimates have no longitudinal dimension and cannot check for both longitudinal and cross-sectional consistency.

One distinct advantage of the DA method in this regard is that it provides detailed single-year of age estimates. The administrative data for DA is virtually complete (no samples involved) and available annually (e.g., births, deaths, immigration data). The demographic process automatically produces detailed single-year of age estimates. The survey estimates are necessarily based on sample data, which will compromise the quality of the single-year age estimates. Among other uses, accurate single-year data are an important ingredient for the Census Bureau's annual population estimates program. The quality of the ICM age data for 2000 could be enhanced if the DA estimates were integrated in the coverage measurement process.

6. **Historical benchmarks**--A major goal of the 2000 census is to reduce the differential undercount. The DA estimates provide the only consistent historical series of detailed age-sex-race undercount factors to document the possible reduction of undercount in 2000 compared to earlier censuses. The survey estimates simply don't have this historical breadth. Further, the detailed 1990 PES estimates for Blacks are flawed for the purposes of making valid 1990-2000 comparisons (e.g., the PES sex ratios for Blacks are implausible compared to the DA estimates).

3. DISCUSSION

In designing a comprehensive Integrated Coverage Measurement system for the 2000 census, we need to balance the strengths and weaknesses of DA and survey-based techniques. Clearly, demographic analysis should play an important role in the evaluation of the census and the ICM operations. The independent demographic estimates will be available on a timely basis to take multiple readings on coverage patterns, before and after the ICM. And it can do this at a relative low cost.

The question is: Should we take the next step forward and formally integrate demographic analysis into the ICM coverage measurement process? In particular, can we enhance the ICM estimates in the areas where DA is strong and the survey estimates have been weak?--(1) the measurement of undercoverage of adult Black men and (2) the production of detailed estimates by age, sex, and race that possess the demographic properties of longitudinal and internal consistency. By integrating the DA results into the 2000 ICM, the age-sex-race differences between the DA and survey estimates will be reconciled before producing the one-number census estimates, not after.

We are developing a research agenda that will spell out how DA can be integrated in the ICM. This agenda also documents the research tasks needed to improve the basis DA estimates themselves. Our goal is clear: To selectively and creatively draw from the unique strengths of demographic analysis to enhance the survey-based ICM estimates used in the final one number census counts for 2000.

4. REFERENCES

Adlakha, Arjun, Hogan, Howard, and Robinson, J. Gregory. "A Report on the Internal Consistency of the Post-Enumeration Survey Estimates," *1990 Coverage Studies and Evaluation Memorandum Series S-1*, U.S. Bureau of the Census, 1991.

Bell, William R. 1993. "Using Information from Demographic Analysis in Post-Enumeration Survey Estimation," *Journal of the American Statistical Association*, Vol. 88, No. 423, P. 1106-1118

Das Gupta, Prithwis. 1991. DA Evaluation Project D10: "Models for Assessing Errors in Undercount Rates Based on Demographic Analysis." *Preliminary Research and Evaluation Memorandum No. 84*, U.S. Bureau of the Census.

Fernandez, Edward W. 1995. "Using Analytic Techniques to Evaluate the 1990 Census Coverage of Young Hispanics," *Technical Working Paper Series No. 11*, Population Division, U.S. Bureau of the Census.

Himes, Christine L. and Clogg, Clifford C. 1993. "An Overview of Demographic Analysis as a method for Evaluating Census Coverage in the United States," *Population Index*, 58(4): 587-607, Winter 1992.

Hogan, Howard. 1993. "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association*, Vol. 88, No. 423, P. 1047-1060.

Mulry, Mary H. and Rajendra P. Singh. 1995. "Development and Evaluation of Census Methodology for 2000 Census," *Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol, United Kingdom, April 1-4.

Robinson, J. Gregory and Edward L. Kobilarcik. 1995. "Identifying Differential Undercounts at Local Geographic Levels: A Targeting Database Approach." Paper presented at the Annual Meeting of the Population Association of America. San Francisco.

Robinson, J. Gregory. 1994. "Use of Analytic Methods for Coverage Evaluation in the 2000 Census." Paper presented at the Population Association of America, Miami, May 5-7.

Robinson, J.G., Ahmed, B., Das Gupta, P., and Woodrow, K.A. 1993 . "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," *Journal of the American Statistical Association*, Vol. 88, No. 423, P. 1061-1071.

Wolter, Kirk M. 1990 "Capture-Recapture Estimation in the Presence of a Known Sex Ratio, *Biometrics*, Vol. 46, P. 157-162.

SESSION 3

MAXIMIZING RESPONSE RATES

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

RESPONSE RATE AND THE CANADIAN LABOUR FORCE SURVEY: LUCK OR GOOD PLANNING?

Mike Sheridan, Doug Drew and Benoit Allard[1]

ABSTRACT

The unemployment rate and employment estimates from the monthly Canadian Labour Force Survey (LFS) are among the key current economic indicators produced by Statistics Canada. Since its inception the LFS has, over the long term, enjoyed excellent response rates. This paper attempts to quantify the reasons for that success. It examines the trends in LFS response and nonresponse patterns including seasonality, and reviews some of the primary factors that on occasion have contributed to increases in the nonresponse rates. Further the paper provides an assessment of the factors and processes that help account for the continued level of high response. The discussion of these factors focuses on items such as training, personal and telephone interview methods, publicity, respondent burden, rotation pattern and a number of factors that affect response rates in both a positive and negative fashion.

KEY WORDS: Labor force survey nonresponse; Components of nonresponse; Nonresponse rates; Techniques for reduction of nonresponse; Interviewers and nonresponse.

1. INTRODUCTION

The title of this paper really begs, in a rather tongue and cheek fashion the question... Are good response rates about luck or are they about being smart? To answer that, we try to do two things. First, we provide a bit of a management perspective to the issue of nonresponse rates in major statistical programs. This is accomplished without any statistical formulae or notation. The second is to say few words about the things Statistics Canada has done to reduce nonresponse and the inherent bias that usually accompanies it. It is fair to suggest that there are a number of uncontrollable factors that define the final nonresponse rates for any survey. Unlike the scientific methodological approach suggested by Bob Groves in the opening keynote address we play the two wild cards of nonresponse - namely the subject of the inquiry and the mood of the respondents. It must be stressed that, as in all endeavors in the world of survey research, the king pin in control of nonresponse is to a high degree predicated, unfortunately, by MONEY. As a manager, money dictates to high degree many of our collective decisions around how much or little nonresponse any particular survey or survey program can or will tolerate.

We are also not given to dismissing outright to the notion that luck plays in the equation but come grudgingly to the conclusion that is not fundamental. Rather it seems to work like a statement made by Martin B. Wilk, former Chief Statistician of Canada, when he occasionally pronounced that sometimes it is much better to be lucky than to be smart.

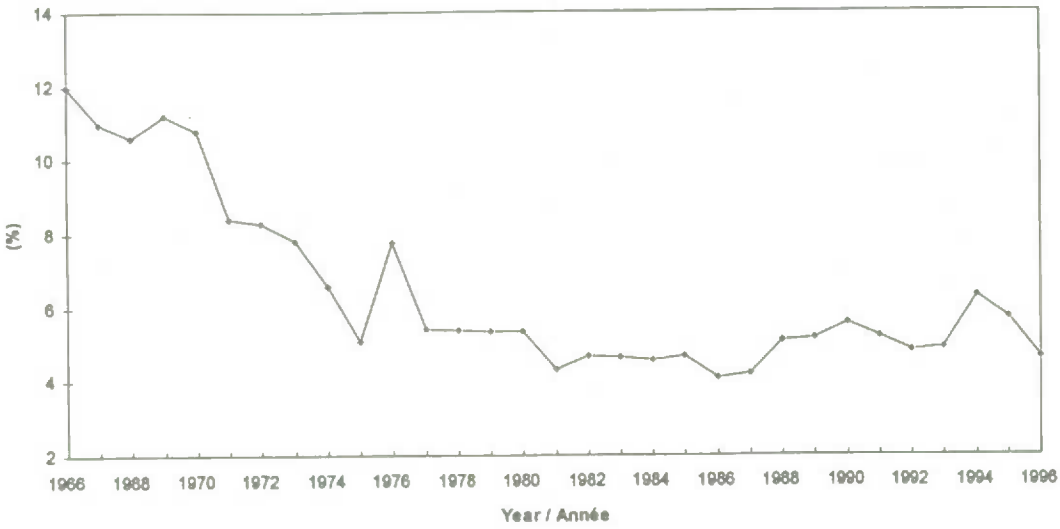
2. NONRESPONSE IN THE LFS: FROM 1966 TO 1996

The first point made is done through the examination trends in overall nonresponse for the LFS over the past 30 years, and discussion of some of the operational factors influencing those trends. As shown in Figure 1, the annual nonresponse rate for the LFS decreased substantially in the late 1960s and early

[1] Mike Sheridan, Director, Household Surveys Division; Doug Drew, Assistant Director, Labour Force Survey Sub-Division; Benoit Allard, Methodologist, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

1970s. This decrease appears to have been associated with increased emphasis on interviewer training and controlling the quality of the interviewers' work. In the early 1970s, the interviewer's manual was revised to make it more detailed and to include specific procedure for various situations arising on the field. For example, it included specific instructions on how to make contacts and call-backs. Also, a special Monday follow-up procedure was introduced in the July and August surveys, in an effort to contact households who were temporarily absent (mostly on vacation) during the survey week.

Figure 1
Overall LFS Nonresponse Rate (Yearly Averages)
Taux de non-réponse global de l'EPA (moyennes annuelles)



The introduction of telephone interviewing was another factor in reducing nonresponse. In the early 1970s, telephone interviewing was introduced for subsequent interviews in urban areas (initial interviews were still done via person visit). By 1975, almost all Canadian cities were converted. Since telephone calls are easier and less costly to place than personal visits, the number of contact attempts was significantly increased and the no-contact component of nonresponse was reduced.

In 1976, the LFS sample size was increased from 36,400 to 55,700 households. This addition of nearly 20,000 households required the hiring and training of a large number of new interviewers. As a result (or so it seems), nonresponse rates were high in 1976 (7.6%), but they reverted back around 5-6% once the sample increase was implemented, and new interviewers gained experience.

The overall nonresponse rate was at its lowest between 1981 and 1987; the annual rate was below 5% throughout this period. Between 1988 and 1993, the annual nonresponse rate was again in the 5-6% range. Another sample increase of about 16,000 households was implemented at the end of 1989. This increase may explain why the nonresponse rate has remained somewhat higher in the early 1990s than in the mid-1980s.

In late 1993 and early 1994, the survey was converted from paper-and-pencil interviewing (PAPI) to computer-assisted interviewing (CAI). This change triggered an increase in the nonresponse rate. Initial versions of the CAI application and case management software slowed down interviewing so there was less time to for multiple contact attempts - leading to an increase in no-contact nonresponse. There were also a variety of technical problems resulting in data being lost on interviewers machines or in transmission. Both of these problems were gradually reduced as improvements to CAI applications and processes were improved and as a consequence the response rate improved until mid-1994.

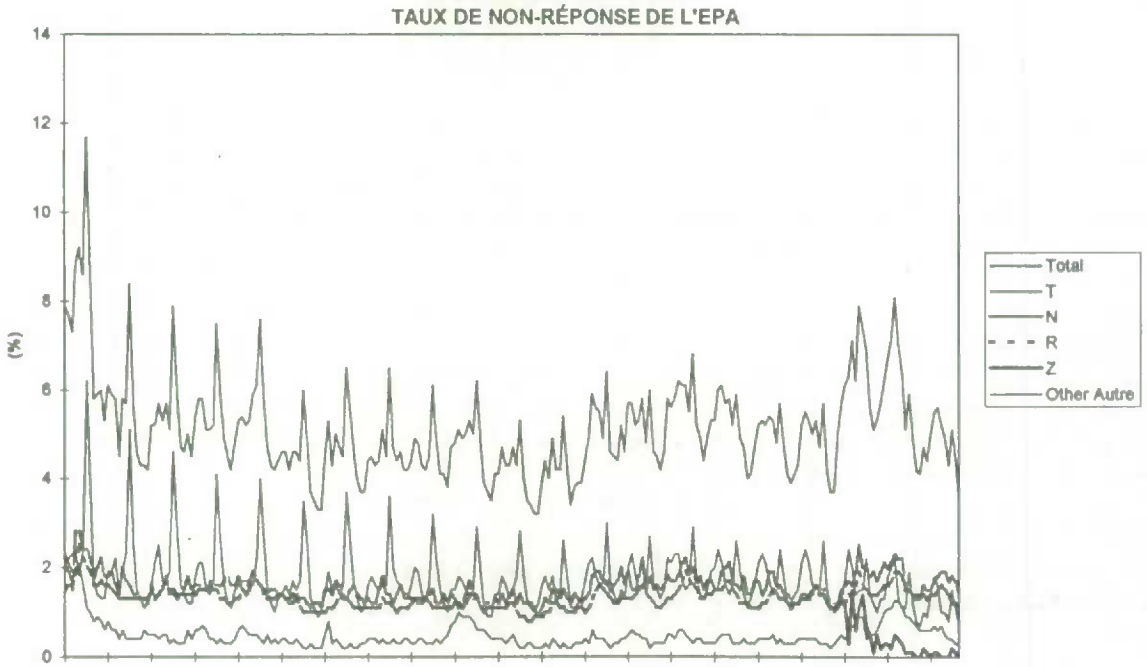
In late 1994 and early 1995, a redesigned sampling plan was phased in, again pushing nonresponse upwards. Many new interviewers were hired at that time, and workloads increased because of the listing activities involved in introducing a new sample. Also, to avoid excessive respondent burden, a household sampled under the new design was not required to answer the survey if it had been part of the sample in the past two years. These "sample overlap" nonresponse cases represented 0.6% of all the sampled households when phase-in of the new sample design was complete in March 1995.

In the months that followed, the nonresponse rate declined rapidly, returning to the low level of magnitude which was experienced before the introduction of CAI. Response rates were affected by the two major changes implemented since late 1993 (the conversion to CAI and the redesign of the sampling plan) but only temporarily; the 1996 average nonresponse rate was 4.6%, the lowest since 1987.

2.1 The Components of LFS Nonresponse

Figure 2 presents the monthly LFS nonresponse rate since 1976. The lower curves in the graph show the breakdown of the overall nonresponse rate in five categories:

FIGURE 2
LFS NONRESPONSE RATES
TAUX DE NON-RÉPONSE DE L'EPA



1. *Temporarily Absent* (T): The household is away from home during the interview period (on vacation, for example). The interviewer was informed of this in a previous interview, or by a neighbour.

2. *No One at Home*(N): The interviewer was unable to make contact.

3. *Refusals* (R): The selected household refused to answer the survey.

4. *Technical Problem* (Z): These include the transmission problems related to CAI which are mentioned above, and cases lost due to laptop computer breakdowns.

5. *Other:* This includes all other reasons for nonresponse: bad weather conditions, unusual circumstances within the household such as death or sickness, language barrier, lack of an interviewer, etc..

These detailed and scientific descriptions really boil down to three sources of nonresponse:

1. we cannot find them - either they are not home or we miss them,

2. we find them and they refuse to answer, or

3. with new technology, we find them, they answer us, and then we cannot get it out of Cyberspace.

As is evident from Figure 2, the biggest contribution to monthly nonresponse for the survey for an extended period of time has been the temporary absents, followed by no one at home, followed by refusals and then technical problems. Seasonal factors aside, and we will come back to them, Figure 2 leaves one wondering why temporary absents would fall over a period of twenty years. We expect that to find the reason would lead on a long interesting voyage. Our hypothesis includes both economic and social factors and we leave that for another paper.

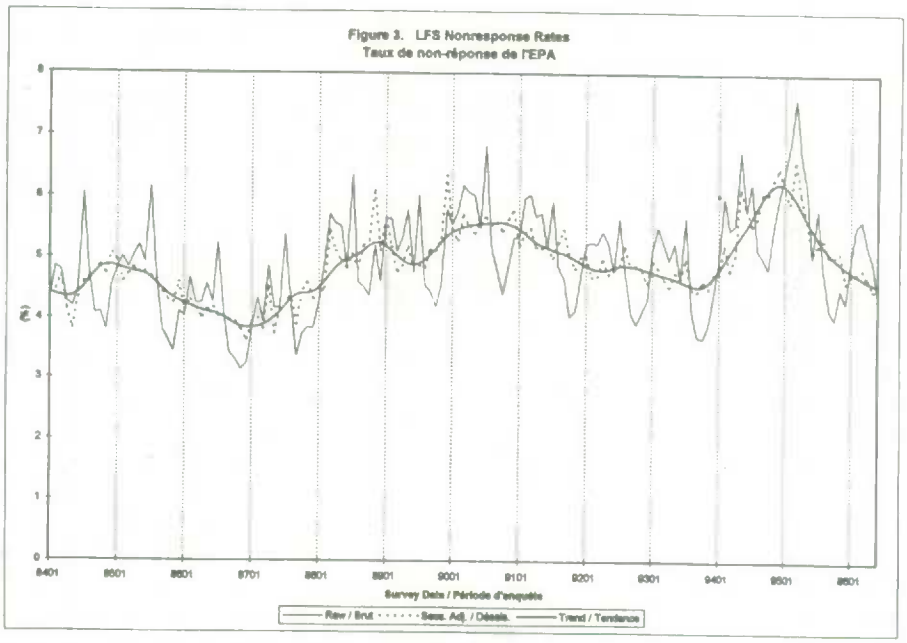
Over the last 20 years, refusals have, if anything decreased slightly and really show a declining long term trend with the exception of some rather small seasonal and monthly fluctuations. The business decision here is of course how much more money does one spend to move the nonresponse down further?...

2.2 Seasonality

As we mentioned earlier there is also a seasonal component to the nonresponse. Figure 3 shows nonresponse, both actual and seasonally adjusted, with a trend line integrated into the graphic. The overall nonresponse rate displays an obvious seasonal pattern with a peak in July. This peak was very high in the late 1970s, but gradually decreased during the 1980s. Meanwhile, nonresponse increased for the winter months. As a result, the seasonal pattern observed in the 1990s is characterised by strong lows in the early fall months (September and October) rather than by the high July peaks of the late 1970s and early 1980s (the July peak still appears in the 1990s, but is far less influential than before). Looking back to figure 2, it is clear that the seasonality in the overall nonresponse rate is driven mostly by the "Temporarily Absent" (T) component. The shift in seasonality is also very apparent in the temporarily absent component. The refusal rate (R) also contributes somewhat to the seasonality of the overall nonresponse rate; peaking in May and bottoming in the fall. With response burden high in April and May as a result of two extensive supplementary surveys, (the Survey of Consumer Finances and the Household Facilities and Equipment Survey), which are administered during these months to a large subsample of the LFS. Also, Canadians must complete their income tax return by the end of April every year; this exercise may affect the public mood towards additional perceived government burden and make some households less inclined to respond.

The trend shows an increase in nonresponse rates over the period from 1987 through 1991 and again between 1994 and 1995. Since 1995, the rate has dropped, returning to about the same level as 12 years ago in 1984. This is perhaps where one could say few words about "luck" and the "wild card factor" - our respondents and the public at large. The increase in response rates during the period from mid 1987 to the first quarter of 1991 are the results of a number of factors some of which might include the possible destabilizing effects of the introduction of a sample increase, the impact of the economic hard times as reflected by the recession or perhaps just the public's mood. One wonders about the impact of a governments popularity and the potential relationship with response rates. With some data polling we reviewed one could make a case to say the public mood does impact respondent relations and especially when the relationship is with a Federal Government Department like Statistics Canada performing as an Agent of the Crown. During this period of increased nonresponse approval ratings as measured by various private sector polling firms declined creating a reciprocal relationship between LFS nonresponse rates (up) and the government approval rating (down). The decline in the approval ratings and perhaps the link to our increased nonresponse may have been reflecting some respondent discontent with government policy. On the other hand, it may have simply been the economy. Or it may have been

operational factors, such as the effect of bringing on large numbers of new interviewers when a major sample increase was introduced. We choose to leave the regression analysis and modeling to someone else, and just leave the speculation there.



2.3 Rotation

The LFS has a rotation scheme that sees one sixth of the sampled households rotated each month. One of the much studied phenomena for the survey is the nonresponse rates over six months households remain in the survey. Figure 4 gives nonresponse rates by the number of months households have been in the survey. The table shows a 12 month period in 1995 and 1996 which is representative of the typical pattern observed. The pattern is one of higher nonresponse in the first month households are in the sample, mostly the result of higher no contact rates. Nonresponse falls in subsequent months. There are a number of reasons for this. Once interviewers have contacted a household, they find best times to call in later months. Another factor is that the first interview is generally a personal visit, so there is a practical limit to how many contact attempts can be made in the short collection period.

One of the big factors in the decline in aggregate nonresponse was the introduction of the telephone interviewing for months 2 through 6. This procedure - telephoning after an initial presonal visit interview - is referred to in the literature as *warm telephone* interviewing. This approach has changed little since its introduction in the early part of the 1970s. The impact of the telephone saw the nonresponse rates drop from the 10% to 12% range in the early seventies to the current 5% to 6% range.

The LFS warm interview approach was predicated on the long established sense of comfort and well being from having the first interview of the six conducted as a personal interview. In fact that comfortable feeling extends not only to responses rates, but also to some degree to the belief that somehow that personal first contact improves overall data quality. This belief is held, despite the fact that there is not a lot of empirical evidence to support the data quality contention. We took an in-depth look at this phenomena to see if we were dealing with a reality or a myth as far as the issue of the first month person interviews and their relationship and generally assumed positive impact on response rates for subsequent months are concerned.

FIGURE 4
NONRESPONSE RATES BY NUMBER OF MONTH IN THE SURVEY FOR CANADA - TOTAL AND BY COMPONENT /TAUX DE NON-RÉPONSE PAR NOMBRE DE MOIS DANS L'ENQUÊTE POUR LE CANADA AU TOTAL ET PAR COMPOSANTE.

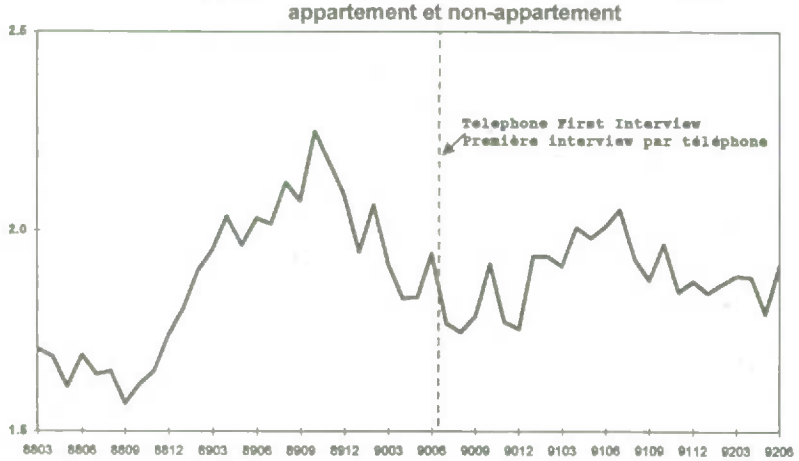
MEAN FOR SURVEYS 9508 TO 9607/MOYENNE DES ENQUÊTES 9508 À 9607

Number of month/ Nombre de mois	1	2	3	4	5	6	Total
Total	7.53	4.76	4.33	4.23	4.19	3.93	4.82
Temporarily absent/ Temporairement absent	1.84	1.21	1.06	0.99	0.96	0.84	1.15
No one at home/ Personne à la maison	2.96	1.63	1.45	1.31	1.21	1.03	1.60
Refusal/Refus	1.72	1.28	1.26	1.34	1.38	1.39	1.39
Other/Autre	1.01	0.64	0.56	0.59	0.64	0.67	0.68

First, it is acknowledged that the telephone has become a "key" element in solving the riddle of how to contact very mobile, seldom at home and tending to be young single persons living in high rise apartments. The LFS decided in 1990 to implement telephone birth interviews in the high rise apartment portion of the LFS sample, partly to combat the problem of the security arrangements associated with these buildings, and as well simply to help overcome the difficulties in ever finding anyone at home during a personal visit to this class of dwellings. Figure 5 illustrates the impact of the decision to do first month interviews for the high rise portion of the sample over the phone. It shows the differential nonresponse rates between the apartment sample before and after the implementation of the first interview by telephone, and there is not much difference in the before and after picture.

Figure 5
Ratio of first month nonresponse rates: apartment vs non-apartment

Rapport de taux de non-réponse pour le premier mois: appartement et non-appartement

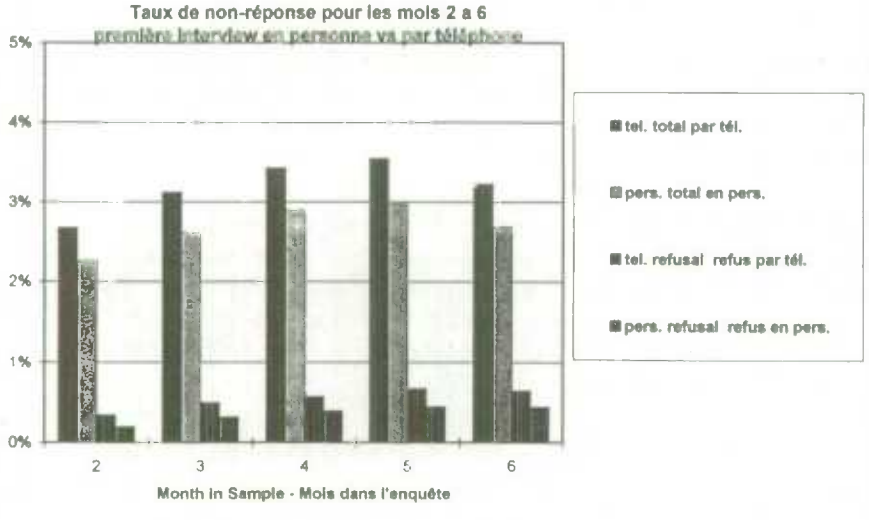


Nonresponse in the high rise portion of the sample has remained about twice as high as that for remaining dwelling types and the introduction of the first telephone interview has made the situation neither a whole lot better or a whole lot worse. This biggest impact in the use of this approach in the high rise portion of the sample is the reduction in the interview costs for those dwellings sampled in high rises.

Right now, for the overall LFS sample, about 75% of first month interviews are conducted by telephone, leaving about 25% conducted as births by telephone. The number of births by telephone has been creeping up over the last few years. Telephoning in the first month is done primarily as a means of follow-up for no contact cases. Interviewers are instructed to make their best effort to conduct a personal visit interview, but if they cannot make contact they are permitted to use the telephone in subsequent attempts.

Figure 6 examines the relationship between nonresponse rates in months 2 through 6 depending on whether the interview was personal or by telephone in month number one. These results are, of course, conditional on being a respondent in month number one. And since telephoning in month one is primarily used to follow-up households that were no-contacts after a personal visit, the telephone group is one more or less prone to nonresponse - at least of the no-contact variety - to begin with. This caveat aside, there are a couple of trends that are immediately obvious. The first is the total nonresponse for households where the first interview is conducted by phone is always higher in remaining months in the sample. The other element which is evident is that the refusal rates for telephone interviews in month one are only slightly higher than for those who were personal interviews in the first month.

Figure 6
Nonresponse rates months 2 to 6:
personal vs telephone first interview



So, even ignoring the differences in the population, and accepting the notion that the first month interview contributes to better response rates in ensuing months, this still begs the question from a management stand point and from an efficiency stand point of whether these quite small differences in performance are worth the really significant investment and cost in first month personal interviews. Having posed this question, we will want to take a more detailed and rigorous review of the first month personal interview. The biggest of the questions that would have to be answered, of course, would include the impact on data quality. Further complicating the picture is the requirement for an assessment of the impact of personal interview on the response rates for other surveys that use the supplementary survey

capacity of the LFS. Additionally samples of LFS households, after they have completed their six month tenure in the LFS are sometimes used as the initial sample for longitudinal surveys. So while there does not seem to be much of an impact on the LFS response rates, other users of the LFS capacity may have less success in retaining respondents cooperation without the benefit of a "theoretical" rapport building first interview.

The other factor in the LFS response rates is something that ad hoc and private research ventures do not benefit from and that is the monthly publicity that the survey garners each month with the media and policy setters. This exposure coupled with the provisions of Chapter 19 of the Statutes of Canada has some benefits on the response side of the equation. These obvious advantages aside, those associated with the LFS program would argue that there are a number of key elements that are critical to the definition of the success of this survey and those associated with its supplementary capacity. In fact, one could argue that these are generic elements applicable to all survey research.

1. Recruiting and hiring. The importance of getting the right people for the job cannot be over emphasized. In the case of the LFS, these people are also committed to the program.

2. Training. There is absolutely no point in even trying to conduct decent survey research work without a strong investment in training. In the LFS program the focus for training is two fold. The first is learning what is necessary to do the job, the second focus is with maintaining and improving these skills.

3. Culture. Nonresponse is not acceptable either by the regions or in head office. The importance of both high response rates and good data quality sometimes seem at odds but over the long run they have balanced each other out.

4. Follow-up with respondents who do refuse. There is a Senior Interviewer follow-up, a refusal letter from the regional director and a return visit the following month.

5. Interviewer tools and back-up. These include a notice of visit letter, introductory letter to respondents, phone number for hard to contact respondents. In addition editing reports are produced for each interviewer assignment that provides information on edited item discrepancies which is reviewed each month.

6. Observation and validation programs. Over and above annual training, each interviewer is observed by the senior interviewer and a review and evaluation session is held with the interviewer. In each month except April and December, ten percent of LFS interviewers have part of their assignment checked. A random sample of seven households in each selected assignment are recontacted. The respondents are asked a number of questions by regional office staff to confirm that the household was contacted and interviewed.

7. The Census effect. I have some serious reservations about putting this forward as one of the elements that improves response rates. However, there are those who are strong proponents of this school. There is a claim to a bonus effect from both Census advertising urging people to respond and also from the publicity surrounding the census prosecutions that is a windfall gain to the LFS.

8. Building rapport. Despite the fact that we showed that this rapport building did not make a really big difference in response rates it still plays as a strong intangible role. Perhaps the key factor in the overall process is the actual interviewer. Some people are really good at it, like it and stay on as interviewers for many years, others are not and tend to leave the job

3. CONCLUSION

1. First and foremost, is that LFS nonresponse rates as such do not appear to be a problem.

2. The public factor and their reaction, or for that matter the media have been and will continue to exert influences on nonresponse rates that do not lend themselves to precise scientific formulation and that will be difficult if not impossible to predict much less control.

3. The impact of personal interviewing in the first month, the long term effects on response rates deserves another hard look from a cost benefit analysis approach

4. The LFS has a very simplistic model for eight things that we do right in collection. We assume they are all predicted upon commitment to the household surveys program and Statistics Canada. The model says that as long as we enjoy the positive approval of our interviewers and hold up to their expectations they will continue to deliver the data quality and the response rates.

4. REFERENCES

Kumar, S. and Bérard, H. *Nonresponse Rates and Trends in Household Surveys.* Methodology Working Paper SSMD-94-002. Statistics Canada.

Statistics Canada. *LFS-100, Interviewer's Manual.*

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

EXAMINING ALTERNATIVE METHODOLOGIES DEVELOPING COMMUNICATIONS STRATEGIES: INCREASING RESPONSE RATES VS. INCREASING NON-RESPONSE

Scott D. Nowlan[1]

ABSTRACT

The premise of this discussion paper is to explore whether the development of a communications strategy with potential respondents leads to the augmentation of response rates. Will such a communication endeavour introduce respondent bias and/or non-sampling error into the fold once contact is made with the potential respondent?

The paper has several objectives. First, is an explanation of what a communications strategy is for survey research. The paper will surmise that during the planning stage of any data collection methodology including the development of a sampling strategy and frame, the researcher should consider conducting a "pre-contact" of respondents to encourage participation in the survey. Although not a universal tool, targeted, more specific research where the sampling frame is well defined and accessible for example, is the type of research which can be a candidate for such a communications approach. Examples of these types of surveys include respondents drawn from administrative data. The surveys that I will discussing include the 1995 Canada Pension Plan Disability Survey (Statistics Canada on behalf of Human Resources Development Canada) and the 1995 National Inmate Survey (Price Waterhouse on behalf of Correctional Service of Canada), both of which benefited from the development of a thorough communications strategy. This paper also examines the components of a workable and effective strategy as well as discussing how such a strategy assists in the augmentation of response rates.

KEY WORDS: Communication; Non-response; Response rates.

1. INTRODUCTION

1.1 Opening Remarks

Under the most ideal circumstances, the conduct of survey research can be challenging. As survey researchers, we understand and appreciate the methodological and analytical consequences of these challenges. Indeed, if these challenges were not sufficient enough, the entrance of numerous participants in the survey research business has placed substantial pressure on the researchers ability to gain the confidence, trust and most importantly, the participation of potential survey respondents. This expansion of the marketplace has had a direct impact on the burden placed on the respondent. These are significant challenges and are the main reasons for this discussion.

To this end, we as researchers need to explore beyond what I would call "pure" methodological issues that can be found in the design, editing and data processing stages of survey research; we must engage in a discourse of "non-traditional" methodologies. In this competitive research environment we find ourselves in - the arena where academics as well as public and private sector researchers compete for research respondents - this discussion and debate is not only important, I would argue that it is essential if we are to continue to rely on the collection of primary data from the same pool of respondents.

One of the other perpetual challenges that survey researchers face is the issue of non-response. The level of non-response has a direct impact on the ability of any methodologist or analyst to accurately measure

[1] Scott D. Nowlan, Price Waterhouse, 1100-180 Elgin Street, Ottawa, Ontario, Canada, K2P 2K3.

outcomes and discuss facts and conclusions. This challenge is also a reason why this investigation of methodological alternatives should be explored.

1.2 Outline of this Paper

The objective of this paper is to enter into a discussion, a debate perhaps, over the use of non-traditional survey research methodology. More specifically, I want to examine how a ""re-contact" of respondents impacts on both response rates and issues surrounding non-response. This paper does not present a compendium of data to prove my hypothesis. Instead this paper examines an approach used to develop a detailed communications strategy for respondents. I will discuss the framework of such a strategy and the detailed components of this approach. Two case studies will be presented to illustrate this framework. The relative success of this framework being presented is still open for debate. With that said, I have also examined some of the perceived advantages and disadvantages and started to look at ways that such a strategy can be improved. This strategy then has not been proven quantitatively. Rather, it builds on similar past studies (Dillman, 1978; deVaus, 1991) in examining alternative methodologies which strive to increase response rates, decrease non-response and augment the validity of the data that we collect.

2. WHAT IS A COMMUNICATIONS STRATEGY?[2]

2.1 Framework

A communications strategy is designed with specific objectives in mind. One of the most important objectives is to increase the response rate of a survey. The debate however is whether such a strategy increases potential response bias. This point, which I will discuss later however, for the purpose of this paper, discussion of response bias will be limited.

Time is an important variable to discuss as part of the strategy. A thorough communication plan takes time to develop and implement. In a world where many clients require information quickly, the process of using this strategy is ignored. Although easier said than done, the expected results realized when such a communication strategy is employed are significant enough to warrant this investment of resources. If the concept is discussed with clients and fellow researchers during the initial design stage of a product, it can become part of the overall project. It should become part of your overall methodology and not added as a last minute thought. Time is always a challenge but implementing such a strategy will assist the research team throughout the remainder of the research project.

The development of a communications strategy rests on a number of fundamental principles. These principles are important to discuss separately in order to tackle the challenges of non-response. In many cases, respondents may refuse simply because "he or she found no convincing explanation about why it should be completed." (Dillman, 1978.) These "sub-components" may themselves seem self evident to researchers or methodologists however, each part must be examined and designed separately as each component has varying objectives. When placed together however, the components form to meet the main objectives of increasing survey response, informing respondents and obtaining buy-in. I propose that there are eight components of a successful communication strategy: event, objective, expectation, confidentiality, encouragement, contact, information reporting and appreciation.

---

[2] Another important source of information for the development of a "pre-contact" strategy is Dillman, 1978. His discussion of a Total Design Method (TDM) also provides valuable insight into reasons why the development of a communications strategy is useful. While this additional resource was only recently uncovered by this author, it has proven to be valuable supporting documentation.

2.2 Event

The opening statement should clearly outline the study that is taking place. Potential survey respondents will be informed of the questionnaire or survey contact itself. In essence, it may be deemed a "warning" that they will be contacted in some manner either via mail or telephone. It should be noted that this could also be the first potential "red flag" which could lead to survey non-response. One of the purposes then of the other components of this contact is to attempt to address these immediate concerns or questions that may be raised in the beginning. If the communication is structured appropriately, one hopes that the other components to be discussed will work towards the elimination of this possible "threat" to completion.

2.3 Objective

It is vital to articulate the objective of the survey to potential respondents. This is important because it commences the "building of a trust" between researcher and respondent. Scepticism or lack of understanding can lead to refusal or non-response - communicating the stated goals of the research will assist in alleviating such mistrust. Another important aspect of this agenda is that it must be designed and agreed upon by the research stakeholders. By stakeholders, I mean the "client" (or owner of the research data) and the "researcher" (or producer of the research data). The objective should state what the data will be used for - what are the research goals?

2.4 Expectation

The communication should address how the research will be conducted. An understanding of the logistics of contact and completion will be important for the respondent to understand. Without addressing this, respondents may not feel comfortable completing the survey. Addressing expectations in a few circumstances has assisted in augmenting response by allowing the researcher to be made aware of any challenges that must be overcome. Examples of this include, becoming aware that respondents need access to T.D.D. (Telecommunications Device for the Deaf) or that the respondent does not have access to a telephone. Alternative data collection logistics can be arranged in advance to ensure participation and therefore lessen the introduction of (non) response bias. Again, this disclosure of information works toward the building of a relationship with potential respondents. By demonstrating that we are forthright, not "hiding" anything, respondents will be more willing to discuss and respond to our questionnaires.

2.5 Confidentiality

This component of an effective communication strategy cannot be emphasized enough. Researchers are aware of mounting cynicism regarding privacy issues. The general population is concerned that personal and corporate information may be both shared and/or used against them.[3] The ethical standards of our industry prevent us from doing so. However, ignoring industry standards will not address the public reluctance to participate in research practices and projects. I believe that concepts of confidentiality must not only be adhered to, they should be repeatedly articulated to potential respondents. Respondents who can be certain that their personal data is "safe from harm" will be more likely to participate. Further, I believe that addressing issues of confidentiality and anonymity should be done on more than one occasion. It may be repeated in a communications strategy but as well, once contact to complete a questionnaire is made, it needs to be repeated. We should not underestimate the value of this practice. Ensuring that the potential participant not only understands the concept of confidentiality but also believes our sincerity and trust, is more likely to participate in our research work. Any evidence shown to the contrary will work to break any trust that is built.

---

[3] Price Waterhouse has conducted numerous focus group sessions and related qualitative research on behalf of Statistics Canada. This research in testing questionnaires with potential respondents suggests there is consistent confusion between confidentiality and anonymity. This was expressed not only as a concern by respondents but one of the key elements as to whether respondents would consider completing a questionnaire.

2.6 Encouragement

It is important that we demonstrate that as researchers we rely on the respondents participation to ensure not only our success, but more importantly, the success of the project. To this end, potential respondents need to be encouraged to participate. Examples of this encouragement includes expressing the fact that their opinions and attitudes are important to the success of the project; (Dillman refers to this as explaining "you are important to the success of the study")replacing how their responses will impact on the data and research outcome such as a program evaluation or employee survey; and, describing that they have been randomly selected and their views therefore represent others (e.g., as in post-weighted data). The goal of this is to instil a sense of importance. Appealing to their sense of altruism will not always be successful. Explaining their role and its importance will assist in this endeavour.

2.7 Contact

The addition of a contact person and/or address/telephone number is another example of articulating the seriousness of the research. One must be aware however that this component of the communications strategy may "turn the table" on the researcher and can augment the burden of collection logistics. For example, researchers must be prepared to field calls and queries. Anticipating this means that one can be prepared for such response. It should be noted as well that this contact may not necessarily be used to build and augment a sense of trust. Many respondents will use a contact to test the legitimacy of the research project.

It provides a few advantages however. The contact will be used by some respondents to make collection arrangements. This may include confirming or changing contact logistics (i.e., collection date or mode), requesting clarification on the stated objectives and asking specific questions on contact. The response to such requests should be discussed during the development of the communications strategy. How the requests are addressed will be an important factor in a potential participants decision whether to respond to a survey. This contact however can become a source for up-front refusal leading to non-response. This is a potential risk, but it is a risk worth taking as you attempt to build a relationship of trust with all potential survey participants.

2.8 Information Reporting

A brief explanation to the respondent of "how" their information will be reported will assist in the respondent's understanding of where their information will go. An example of this would be that reporting will include only summary data or data will be shared only with a specific organization etc. This can be addressed at the same time as the explanation surrounding confidentiality as there is a logical connection between these two components.

2.9 Appreciation

To build on the other components of the strategy, researchers should not overlook the importance of demonstrating to potential survey respondents that their participation is important and appreciated. (Dillman, 1978.) Although perhaps a "simple" concept, demonstrating appreciation for an individual's time and effort will encourage people to participate.

3. CASE STUDY ONE

3.1 1995 National Inmate Survey

The National Inmate Survey was conducted by Price Waterhouse on behalf of Correctional Services Canada (CSC) in the Autumn, 1995. Price Waterhouse (PW) and its National Survey Centre completed over 4,000 surveys of inmates at forty-four federal penal institutions across Canada. The survey itself covered a broad range of topics intended to meet both legislative requirements as well as the recommendations of a federal task force looking at health risks such as AIDS faced by inmates housed at federal institutions. A total of nineteen specially trained field staff were recruited to enter these institutions and facilitate the collection of the self-completed, scannable questionnaires.

During the preparatory and design stages of the project PW designed a comprehensive communications strategy. This was particularly challenging given the number of "stakeholders" involved. In order to ensure the success of the research project PW needed to communicate the objectives and arrangements (all the components as described in the previous sections of this paper) to not only the potential respondents - the inmates - but also to inmate committees, staff, wardens and regional Corrections officials, all of whom could be defined as "gatekeepers". Bringing all of these individuals on side would prove to be a challenge given that they could be a potential source of non-response.

The process of communication strategy development was an iterative one with PW providing initial drafts and requesting successive input from the CSC Project Team. One point of contact was established at each institution to ensure that the collection logistics and the communication vehicles (i.e., letters) be directed and monitored to assist in this complex procedure.

CSC sent formal communications to all wardens and regional staff informing them of the logistics. Letters were sent by PW to all points of contact as well as all inmate committees throughout the penal system. These letters included the seven strategic components that have been discussed in this paper. Optimally, these letters were intended to bring these people "on-side". Site contact participation would be necessary to ensure access to potential respondents. The support of inmate committees was sought to obtain buy in which, if accomplished, would lead to an endorsement and thus a higher response rate among the inmate population. The risk that was taken throughout the communications process was that without an inmate committee endorsement, an unusual (or "non-traditional") type of non-response would be experienced, that of proxy refusal. More specifically, non-response would take place outside of the control of the respondent - (an interesting point of debate at this point, perhaps is whether these respondents would be considered out of scope). An additional point of contract was initiated after the receipt of the letter by means of face to face meetings with both site contacts and inmate committees to address questions or concerns and to finalize collection logistics. Given the circumstances of this research project, a detailed and thoroughly implemented communications strategy was necessary for the eventual success.

Overall non-response rate for this survey was calculated to be 34.6%. As expected, non-response varied widely across institutions and regions. These varied from 0% to 69.0% for institutions and 29.9% to 39.9% for regions.[4] Specific reasons for refusal varies however the majority were found to be due to inmate attrition such as transfers and release (particularly at minimum security institutions) and outright refusals.

I contend that with the adoption of a communications strategy, non-response rates would have been much higher. Further, the inclusion of the eight components as described were instrumental in addressing those issues respondents felt were important during the pilot test of the research.[5]

---

[4] Correctional Service Canada. (1995) *Research Report.* Government of Canada.
[5] Correctional Service Canada. (1994) *Pilot Study Report.* Government of Canada.

4. CASE STUDY TWO

4.1 1995 Canada Pension Plan Disability Survey

The 1995 Canada Pension Plan Disability Survey was conducted by Statistics Canada (STC) on behalf of Human Resources Development Canada (HRDC). The main objective of the survey was to identify and profile Canadians currently receiving disability benefits under Canada Pension Plan (CPP) legislation and programs. A random stratified sample was drawn at the regional level to allow for the analysis of regional differences across Canada. Subjects explored included reasons for benefits, individual and family demographics and most importantly, detailed financial and income information. The results of the survey were also linked to HRDC CPP administrative files to conduct further, more detailed analyses and modelling.

A detailed communications strategy was developed in this instance for similar reasons as the above case study, however, there were also very different issues which needed to be addressed by this pre-contact. It was hypothesized that the respondents for this study would be fearful of discussing their pension benefits for fear of losing them; we therefore wanted to address this concern. As well, we were collecting detailed income and financial data therefore we wanted to allow respondents time to prepare in order to share this information. A contact (in this case, a 1-800 number) was also established and was used quite extensively by respondents. The reason for this seemed to be to check on the validity of the study and to express concern over losing benefits. Once the study was further explained and the use of the data was articulated, respondents stated their willingness to participate.

The overall non-response rate for this survey was 45.9%.[6] Response rates were notably lower among younger respondents than older individuals. There were three levels of non-response, a portion of which were refusals to complete the entire questionnaire. Almost half of all non-response cannot, I would argue, be attributed to the "pre-contact" through the implementation of the communications strategy. Of the 45.9% who are part of the non-response statistic, 60.7% did not complete the questionnaire or were only considered "partial" completes. 23.9% did not have a matching phone number to the address on the administrative file, even after tracing and the remaining 15.4% did not complete the final question on informed consent which was needed to link the survey data with the CPP administrative records for detailed analysis.

Despite what some may consider to be a higher non-response rate than is traditionally the case for Statistics Canada, the use of a communications strategy for this project was deemed to be successful. The subject matter was both controversial and sensitive and if one examines the breakdown of non-response, almost half of these individuals were not considered complete because of reasons that this strategy could address.

5. THE IMPACTS

5.1 The Results of a Communications Strategy

At this point in the debate, I would argue that an effective means of determining the true statistical impact of a communications strategy as defined in this paper would be it's adoption as part of an experimental design. In the absence of such "hard" or empirical evidence, we need to briefly explore both the perceived advantages and disadvantages of employing such a strategy within data collection methodology.

---

[6] Statistics Canada (1995). *1995 Canada Pension Plan Disability Survey: Final Report.* Government of Canada.

5.2 Advantages

5.2.1 Building a relationship with respondents early in the collection process. Establishing trust with respondents is a constant challenge for survey researchers. Without trust and understanding respondents will be less likely to participate. Ensuring that all important concepts are conveyed and allowing for early detection of problems, concerns and necessary alternatives are imperative to the success of any data collection project.

5.2.2 Increasing response rates. One of the main objectives of any survey is to produce valid, rigorous data. This statement is self evident. Higher response rates to any survey greatly assist this process. While response rates in and of themselves do not guarantee this, higher response rates reduce bias found in non-response, reduce the need for complex interpretation, decrease the dependence on weights to explain a variable and as well, significantly increase the confidence that researchers and analysts have in the results. The communication strategy is a tool -- not an end in itself -- to assist the researcher to deliver important messages to respondents. These messages, if delivered properly, will foster the trust and confidence that is necessary to build respondent participation.

5.2.3 Decreasing collection time. The adage that "time is money" holds true for our business. One of the objectives of employing a communications strategy is to address some of the concerns and questions that respondents may have up front rather than during the collection process, thus saving interviewers' time. This is not to say that communications strategies are inexpensive, nor do they replace interviewer/respondent communication. I would hypothesize however that such pre-contact will minimize the time that interviewers need to spend on handling discussion pertaining to research objectives, goals and uses of data. I would argue then, that the employment of a communication strategy may impact revenue as well.

5.3 Disadvantages

As with all approaches, both theoretical and practice, there are limitations. There are reservations regarding the universal applicability of this model. The proceeding points briefly outline some reservations about such a strategy.

5.3.1 Potential for early refusal. A pre-contact could make participant refusal easier. In contacting potential respondents early and providing a contact number and/or person, we could make it simple to opt out of participating. As is the case with all potential refusals, we need to attempt to dissuade potential non-participants. This is perhaps analogous to not asking potential respondents if they have time to complete a questionnaire.

5.3.2 Increase non-response leading to potential bias. There are arguments which can be made, as stated above, that it can lead to immediate non response however, any non response must not be ignored. Non responses must be classified and discussed both prior to data analysis in order to measure impact and during the interpretation and report writing stage. To this end, any project employing a communication strategy should discuss its perceived or potential impact on response.

6. MAKING IMPROVEMENTS

6.1 How the Communication Strategy can be Improved

We have the luxury of being able to adjust and experiment with various alternative methodologies. A communications strategy is not different in this regard. If a component is not working or if it is not applicable for a specific audience, we can make changes. Through such changes we can explore alternatives in the hopes of improving this strategy and in the end, increase response rates thus decreasing

non-response. I believe there are four main areas that must be explored if we are to further enhance this type of strategy. First, researchers must continue to examine alternatives for obtaining buy-in and increasing response rates. What else could we be doing outside of the traditional realm which may lead to increase response rates and lower non-response? Secondly, we must examine whether the information that we are conveying in our communication to respondents is accurate and correct. Each project will require that different needs be met and therefore different points be addressed in the pre-contact. A continuous review of objectives and specific goals is therefore important. Thirdly, it is important to stress the issue of confidentiality. There are numerous ways this concept can be described - it must be clear and accurate, if the respondent is to understand the concept and therefore be willing to participate. Lastly, we must explore alternative media for such a pre-contact. We have traditionally utilized a mail methodology. However, with increasing communications choices, we may have other opportunities to contact potential respondents.

7. CONCLUDING REMARKS

The debate over contacting respondents prior to the completion of a questionnaire will continue. One way to convince researchers is of course providing empirical evidence that supports a communications strategy, (see Dillman, 1978). As we search for a critical mass of such empirical evidence, however, we need to continue to look for options. The competitive world of survey research and the search for the illusive respondent is here to stay. We can no longer rest on corporate reputations or the expectation that people will take time to answer our calls. Researchers must begin a discourse of alternative methodologies - one which challenges traditional methods. With new technologies on the horizon, and the Internet here to stay and flourish, competition will only increase. Sharing our knowledge will help to ensure that respondent's knowledge is shared with us.

8. REFERENCES

Blalock, Hubert M. (1982) *Conceptualization and Measurement in the Social Sciences.* Beverly Hills: Sage Publications.

Buckland, William R., Fox, Ronald A. (1963) *Bibliography of Basic Texts and Monographs on Statistical Methods, 1945-1960.* Edinburgh and London : Oliver and Boyd.

Church, Allan H. (1993) "Estimating the Effect of Incentives on Mail Survey Response Rates." *Public Opinion Quarterly.* Volume 57:62-79. American Association for Public Opinion Research. Correctional Service Canada. (1994) *Pilot Study Report.* Government of Canada.

Correctional Service Canada. (1995) *Research Report.* Government of Canada

De Vaus, D.A. (1991) *Surveys in Social Research.* Third Edition. Sydney: Allen and Urwin.

Dillman, Don A. (1978) *Mail and Telephone Surveys: The Total Design Method.* New York: John Wiley and Sons.

Goyder, John. (1987) *The Silent Minority: Nonrespondents on Sample Surveys.* Cambridge: Polity Press. Cambridge.

Mills, Frederick C. (1955) *Statistical Methods.* New York: Holt, Rinehart and Winston.

Rubin, Donald B. (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

Statistics Canada (1995). *1995 Canada Pension Plan Disability Survey: Final Report*. Government of Canada.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

ENCOURAGING RESPONSE TO AGRICULTURAL SURVEYS

Patricia Whitridge[1]

ABSTRACT

Over the past few years, the issue of encouraging farmers to respond to Statistics Canada's surveys has become increasingly important. The population of interest is finite and, as is common for business surveys, highly skewed. Collecting relevant information from the largest farmers is virtually impossible. The burden placed upon the smaller farmers was also considerable. Between the amount of government money spent managing the industry and the concentrated interest in agriculture, the large number of surveys was inevitable. As a result of these factors, a concerted effort has been made to reduce the respondent burden and to encourage farmers to respond to our surveys.

The measures taken to date include the heavy use of administrative taxation data, coordination between surveys with similar reference periods, minimization of overlap between major surveys, and deliberate attempts to improve the quality of the sampling frames to eliminate unnecessary contacts. These have been complemented by the use of focus groups to understand farmers' concerns with Statistics Canada's surveys, mixtures of collection modes according to respondents' preferences, and establishing personal relationships with the largest farms. For certain surveys, respondents are provided with summary publications about agriculture in Canada.

This paper will describe the various initiatives in more detail, highlighting the individual surveys affected, and presenting some results.

KEY WORDS: Agricultural Surveys; Survey Response.

1. INTRODUCTION

Statistics Canada has a mandate to collect agricultural information from Canadian farms on an ongoing basis to produce statistics on crop acreages and yields for field crops and fruits and vegetables, livestock inventories, and financial information. All farms in Canada, regardless of size, location, or enterprise are of interest. There are approximately 17 agriculture surveys that are conducted on an annual or sub-annual basis. In addition, every year there are some smaller ad hoc surveys that take place to provide information of specific interest to the sponsors. In 1995, approximately 220,000 contacts in total took place, from a population of 280,000. However, it should be noted that many farmers were contacted more than once. The total collection cost for 1995 was $2,900,000, a significant amount.

Most of the surveys are conducted over the telephone, using computer assisted telephone interviewing (CATI) technology. Sample sizes range from several hundred to thirty thousand respondents, with collection periods between one week and several weeks long. The response rates for CATI surveys range from 89% to 99%.

The sampling frame for most agricultural surveys is derived from the Census of Agriculture, which is conducted every five years. The Census list frame is considered frozen for the five year period; in intervening years, the major surveys use an area frame to compensate for births and changes to existing farms. Practically, the Census information is loaded onto the Farm Register, which serves as the repository of administrative information about farms.

[1] Patricia Whitridge, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario Canada, K1A 0T6.

The agricultural survey program, when considered as a whole, poses a significant response burden to Canadian farmers. The majority of farms are run as family farms, with few employees, if any. For these farms, it is the principal operator or a member of his family who is asked to take the time to respond to surveys. Since the level of education of farmers is quite variable, there is an impact on the quality of data provided, especially to financial surveys.

There have been no astounding advances to improve response rates to Statistics Canada's agricultural surveys. Rather, there have been a number of smaller initiatives, each with a specific target group of farms, over the past few years. It is the set of these initiatives that will be discussed in this paper.

2. CENSUS OF AGRICULTURE 1996

Every five years Statistics Canada conducts a Census of Agriculture, most recently in May of 1996. Information on all aspects of farming is collected, covering crop acreages, livestock inventories, basic finances, farm management practices, and land use. All farms with the intent to sell agricultural products are in scope for the Census, regardless of actual sales. The data from the Census of Agriculture are then used as the sampling frame for the ongoing surveys.

Over the past few years, the number of telephone answering machines on farms has increased. Some farmers now have several phones: a house phone, a separate phone (and possibly number) in the barn or office, and sometimes a cellular phone in the tractor. This has made it possible to talk to farmers as they work on the fields, which has increased the rate of contact. All of the information about phone numbers has been collected for all farmers involved in a given farm; in past Censuses, only the principal operator information was captured and available to interviewers for survey data collection. The collection of the additional information should make it easier to contact the farmer, or one of his partners.

Once the information from the Census of Agriculture is loaded onto the Farm Register, we expect to see an improvement in the rate of contacting farmers, due in part to the additional information collected. As well, if we are able to reach farmers where they are not too busy to talk, such as on the tractor, they may be more likely to provide the data we need. Previously, farmers were often contacted at meal times when they were more likely to be in the house, but also more likely to resent being disturbed.

3. SURVEY REDESIGN

Every five years, after the Census of Agriculture, Statistics Canada undertakes a redesign of the ongoing agricultural surveys. This redesign program revolves mainly around the crop and livestock surveys, both of which are conducted on a sub-annual basis. Other smaller surveys, such as fruit and vegetables, greenhouse, sod and nursery, and potatoes are also redesigned after each Census. Different aspects of the sample design are considered, such as sample frame, stratification methods, sample allocation and selection, then sample maintenance over the five year period (Denis and Whitridge, 1995). Estimation methods to take advantage of the sample design are tested and robustness over time is of particular concern. Moreover, the overlap between the crop and livestock samples, as well as other survey samples is important, since many surveys are conducted within short time spans of each other.

This time, budgetary constraints are of special interest, since certain savings need to be realized over the five year life of the sample design. As a result, much thinking took place about the possibility of combining the sample design for the crop and livestock surveys, or at least combining the field collection. Historically, both approaches have been tried, with different advantages and disadvantages. It was feared that a combined survey would require a long questionnaire covering all aspects of farming, which would discourage response. On the other hand, it would alleviate the problem of contacting crop farmers for the livestock survey, then not collecting the information about their farm beyond the livestock. Efforts to

control the overlap between the crop and livestock samples were considered, either to maximize the overlap and do a combined collection, or to minimize the overlap and have separate collection. Either of these options is potentially problematic, since both samples undergo annual rotation for the five year inter-censal period.

It was decided to keep collection separate for the two surveys, and to minimize the overlap between the samples. Where possible, the sample size for specific occasions and the questionnaire content will be reduced, to encourage response. Field interviewers will be informed if respondents were contacted for another survey within a short time span, to take special care with the respondent.

4. LARGE FARMS

A special effort has been made over the past two years to address the problems associated with collecting information from large farms, defined based on a sales threshold. Some of the largest farms in Canada are part of multi-holding corporations, such as McCain and Nabisco. It is very difficult to obtain information about their farming activities when they are approached during the regular interview process. Large multi-holding corporations are not always able or willing to report the commodity information in the format required by the surveys. In addition, different people in the organization may be responsible for answering a survey, depending upon the questionnaire content.

A group was put in place to contact each of the multi-holding corporations and large farms to undertake a profiling exercise. As a result of this, profiles were established, showing the most appropriate person to contact for each of the surveys, and the best way to approach them (Blais, 1995). This information has been used as part of the data collection for agricultural surveys for about one year now, including the Census of Agriculture.

For the ongoing surveys, the large farms are contacted by telephone by the same person each time, who quickly becomes familiar with the farm and how the data will be supplied. A personal relationship is then established between the interviewer and the respondent, which increases the feeling of trust, and encourages the respondent to provide the information. Experience over the last year leads us to believe that we are receiving better quality data and enjoying higher response rates than before, especially considering that these units had been unlikely to respond to our regular surveys in the past.

Different scenarios were considered during the development of a collection model for these large and complex farms. A choice had to be made between contacting the farms once a year with an omnibus questionnaire and contacting them as necessary for the individual surveys. The first option would have required each survey to transform the data they receive into a format they could use. After consulting the large farms, it was decided to contact them as necessary with individual survey questionnaires.

Response rates for the large enterprises to the 1996 Census of Agriculture are approximately 99%, a dramatic increase over past experiences.

5. ATLANTIC PROVINCES

Traditionally, the agricultural statistics program at Statistics Canada covers all 10 Canadian provinces. Newfoundland has been surveyed separately from the rest of Canada, due primarily to the very small number of farms in the province. The quality of estimates produced for the remaining Atlantic provinces (Prince Edward Island, New Brunswick and Nova Scotia) is uneven. For some surveys and some commodities, the coefficients of variations (CVs) produced are very good; however for others they are quite weak. Given the small number of farms in these provinces, it was felt that the response burden was quite high, and so increasing the sample sizes to improve the reliability was not a viable option.

As a result, the survey redesign that we are undertaking using the new Census of Agriculture data for a sampling frame, treats the Atlantic provinces as a separate survey program. Consultations have been ongoing with experts in the Atlantic provinces to develop a program of surveys that will not impose an unreasonable response burden, while at the same time improving the quality and usefulness of the data.

The proposed program will have a reduced number of survey occasions and use a combined sample design for crops and livestock surveys, along with an integrated collection for poultry and potato surveys (Denis, 1996). A large sample will be taken in November to provide final data on seeded acreages, production and livestock inventories, producing data at small area levels for important commodities, with a small follow-up survey occurring in July. This should significantly reduce the number of contacts while making small area data available for the first time, as has been requested through the consultation process.

The new survey program for the Atlantic provinces is scheduled to be implemented in a phased approach, with the first stage taking place in the fall of 1997.

6. ADMINISTRATIVE DATA

The agricultural statistics produced by Statistics Canada have included tabulations based on administrative data since the early 1980's, as part of the Tax Data Program. Information is obtained from Revenue Canada for a sample of farms, for which detailed financial statements are transcribed. Based on this data, analyses of income and expenses are possible. Some farms also supply a balance sheet, which permits estimation of assets and liabilities, although with limited quality, since the balance sheet is not a required part of the financial statement. Changes taking place in the forms used to report business income and expenses will improve the timeliness and quality of the tax data being disseminated, since income and expense information will be available for all tax filers reporting self employed income (for individuals) and corporations. Statistics Canada has been involved in negotiations with Revenue Canada over the exact layout of the forms. These changes will take effect over the next two years. Due to the richness of the taxation information, Statistics Canada will not need to conduct a survey to collect detailed income and expense data.

In the early 1990's, the Whole Farm Data Project was initiated to assemble and integrate information about all aspects of farming - crop and livestock variables, as well as financial items (Whitridge and Ménard, 1994). Attempts were made to reconcile survey commodity data and administrative financial data to produce a view of the "whole farm". As part of this effort, different statistical techniques were examined, always with the goal of avoiding a separate survey to collect information on both physical commodity variables and finances. Data from the Whole Farm Data Project have been marketed to different stakeholders, who take advantage of this information, collected without any increase in respondent burden.

As part of their mandate, Agriculture and Agri-Food Canada requires information to administer effectively different agricultural support programs. Wherever possible, data already collected from government agencies, such as insurance programs and inspections, is used. In some cases where administrative data is not available, small surveys are run.

7. FARM FINANCIAL SURVEY

The Farm Financial Survey is a biennial survey conducted jointly by Statistics Canada and Agriculture and Agri-food Canada. Its main purpose is to collect information to analyze the financial situation of Canadian farmers. The results are used to evaluate farm subsidy programs and identify potential changes that could be made. Detailed balance sheet information along with some income and expense data are

collected. The collection mode for the survey is rather complex, being initiated by a letter sent to all sampled farmers introducing the survey and explaining its purpose. The farmers are then contacted by phone by interviewers, to arrange a time for a longer personal interview. At that time, the questionnaire is completed, based largely on the farm accounts.

Past survey occasions have experienced non-response rates as high as 20% for some important variables. It was felt that the refusal rate was particularly high due to the sensitive nature of the financial questions and the length of the interviews. Some efforts to use incentives to encourage response have been made, with limited success. The incentives in question ranged from calendars and pens to current practices of supplying some final survey data to interested respondents. In an attempt to improve the situation, Statistics Canada conducted a study using qualitative techniques to investigate methods to encourage response to the 1993 Farm Financial Survey questionnaire.

The study comprised a critical review of the questionnaire, personal interviews, in-depth interviews and focus group discussions with farmers in three regions of Canada (Lawrence and Laffey, 1993). The methodology of the study started with the questionnaire review by questionnaire design experts. This helped to identify potential problem areas, which could be highlighted during the interviews and focus groups. Then, a sample was drawn from the survey frame of farms that would not be included in the actual 1993 survey sample. Participants were contacted and invited to participate in a one-on-one personal interview, to be followed by a group discussion a few days later. Some participants were unable to attend the focus group, so in-depth interviews were conducted with them.

The recommendations from the study were numerous: change the initial approach to respondents, revise the questionnaire wording, content and order, think more about the timing of the collection period, and do not use local interviewers, to name just a few. The results of the study were useful in the development of the collection methodology for the 1993 Farm Financial Survey. The most important findings for Statistics Canada involved how to influence participation in the survey.

8. USER CONSULTATIONS

Statistics Canada has many different forums for user consultations, particularly with respect to agriculture. It is felt that these consultations lead to increased response, as more respondents are part of the survey process and are aware of the value of the data.

A starting point for these consultations is an annual meeting between federal and provincial representatives of agricultural statistical agencies. Information is exchanged about recent progress in different programs and an opportunity for questions and input into the agricultural statistics program at Statistics Canada is provided.

In addition, as part of the Whole Farm Data Project, meetings are held with many groups of stakeholders across the country. These meetings take place approximately each year, sometimes more often, depending upon the progress of the project. Input is solicited about what data would be of interest to users, as well as in what format, in terms of content, timing and dissemination media.

Finally, there is the Advisory Council on Agricultural Statistics. This body meets annually to discuss the work of Statistics Canada with respect to its agricultural statistics program. Experts from across the country, both academics and practitioners, provide critical input and examination of our programs. Questions are sometimes raised that must be answered at subsequent meetings.

9. CONCLUDING REMARKS

Over the past few years, as the burden being placed upon respondents has increased, much effort has been expended to develop surveys that are more respondent-friendly, in terms of collection modes, questionnaire design, and survey procedures. In agriculture, many small steps have been taken that affect almost all aspects of the agricultural statistical program: surveys, the Census, and programs that use administrative data.

As part of the survey redesign, consideration has been given to the best manner of collecting the data from the respondents, be it via a unified sample design, integrated collection, or two separate designs with minimum sample overlap. The statistical program has been made more flexible, which has allowed special procedures to be established for large farms, as well as a separate, more appropriate, survey program in Atlantic Canada.

When requests for information are received, the possibility of using administrative data is now seriously considered before a new survey is proposed. Procedures are in place whereby new surveys routinely use pilot tests and focus groups to establish questionnaire design and collection methods. These help ensure that new surveys, while placing an increasing burden on our limited population, do not jeopardize the response rates we have worked so hard to maintain. Increasing avenues for input from stakeholders, both respondents and data users, are being made available.

As this paper has illustrated, there have been no astounding advances to encourage response to agricultural surveys. Rather, there have been many small initiatives across different survey programs to improve and maintain the response rates. Field and head office coordinating staff must be commended for their work in maintaining these high response rates.

10. REFERENCES

Blais, K. (1995). Multi-Agriculture Operations, Large Farms, Special Farms and Specified Farms (MULES) Project, Presented at the Federal-Provincial Meetings on Agriculture Statistics, Ottawa, Ont.

Denis, J. (1996). Atlantic Canada - Survey Redesign Methodology Proposal, Statistical Internal Technical Report, Ottawa, Ont.

Denis, J. and Whitridge, P. (1995) 1996 Post-Censal Redesign of Agricultural Surveys, Presented to the October 1995 Advisory Committee on Statistical Methods, Ottawa, Ont.

Lawrence, D. and Laffey, F. (1993). Qualitative Testing of the Farm Financial Survey Questionnaire, Proceedings of the International Conference on Establishment Surveys, Buffalo, New York.

Whitridge, P. and Ménard, M. (1994). Methodological Issues Involved in the Creation of a Farm Level Database, Proceedings of the Annual Research Conference, Washington, D.C.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

MINIMISING NON-RESPONSE IN A PANEL SURVEY

Heather Laurie, Rachel Smith and Lynne Scott[1]

ABSTRACT

This paper provides an evaluation of some of the fieldwork procedures and survey systems used on the British Household Panel Study (BHPS). The BHPS procedures for dealing with non-response through panel maintenance systems, tracking procedures, and refusal conversion during fieldwork are described. The analysis uses data from the first four waves of BHPS from 1991 to 1994, to examine longitudinal patterns of response and reasons for refusal. The reasons for refusal or for becoming a non-contact over the life of the panel are discussed. The process of refusal conversion is described together with conversion outcomes. Finally the effect of interviewer continuity on maintaining the co-operation of sample members is examined. The paper argues that in the context of a longitudinal panel survey having a relatively complex set of procedures in place is critically important to minimise non-response and maintain high response rates over time.

KEY WORDS: Response rates; Longitudinal survey methodology; Fieldwork procedures.

1. INTRODUCTION

Conducting a longitudinal panel survey presents a number of specific problems which have a direct bearing upon data quality. For longitudinal studies such as cohort or panel surveys, minimising non-response to counter the potentially damaging effects of attrition and to maintain a viable sample is essential (Kazprzyk et al,1989). Survey non-response has long been recognised as a complex and multi-faceted phenomenon (see for example Sudman and Bradburn, 1977). While longitudinal panels share many of the difficulties faced by cross-sectional surveys in gaining a high response rate, the very nature of the panel design imposes additional complexities in terms of response rate requirements. Panels face two main problems specific to their design which can result in attrition over time. The first major source of loss from a panel survey is due to the geographical mobility of sample members. If respondents move and, despite all efforts cannot be traced, they are effectively lost from the survey. Moreover, the respondents who are most likely to be geographically mobile tend to differ from those who maintain a stable home address. So the problem of differential attrition arises where a particular category of respondent can become under-represented within the sample. The second, and more extensive source of loss, is due to refusals, very often the result of what we will call panel fatigue. At every interview point, respondents have the option of refusing to take part in the survey. After co-operating for what can be some years of a panel, respondents may become bored or disinterested in taking part any further or simply feel they have 'done enough'. While the majority of respondents become rather committed to taking part and actively enjoy the interview process, inevitably there are some respondents who decide they do not wish to carry on. As with the geographically mobile, those who refuse to be interviewed tend to have specific characteristics, potentially producing differential patterns of attrition and, at worst, bias within the data collected.[2] The aim of this paper therefore is to assess the effectiveness of the procedures used on the BHPS for minimising non-response over time.

---

[1]Heather Laurie, Rachel Smith, Lynne Scott, ESRC Research Centre on Micro-social Change, University of Essex, Colchester, CO4 3SQ, England.

[2]The issue of differential attrition from the sample and the weighting techniques used by the BHPS to compensate for non-response bias are not directly addressed by this paper. Please see Taylor,A.(1993) for a description of respondent characteristics in relation to differential attrition and weighting procedures.

2. THE BRITISH HOUSEHOLD PANEL STUDY

The British Household Panel Study is a national household panel survey of over 10,000 individuals in some 5,500 households in Britain which is carried out by the ESRC Research Centre on Micro-Social Change based at the University of Essex. The sample covers non-institutional residences in England, Wales and Scotland. The BHPS began in September 1991 and returns to re-interview panel members on an annual basis. At Wave 1 of the survey 13,840 individuals, including children under 16 years of age, were enumerated in 5,511 households. Of these, 9,912 eligible adults i.e. aged 16 years or over were interviewed and 352 proxy interviews taken giving an upper response rate (full interviews with at least one member of the household) of 74 per cent. The fieldwork for the sixth wave of the survey began in September 1996 and we will be returning to our respondents for the seventh time in September 1997. The BHPS collects information at both the household and individual level. At the household level the questionnaire covers household composition; housing tenure and costs, consumption items and household expenditure on fuel and food. The individual questionnaire collects a wide range of information on migration, health status and usage of health services, detailed employment and income information, values and opinions, household organisation and a self completion questionnaire containing attitudinal items and some GHQ items (see Rose et al, 1991 for a full description of the content and design of the BHPS). The household interview takes around ten minutes to administer and each individual interview 40 minutes, on average, keeping the total interview package for any one person to no more than one hour maximum. In addition, since 1994, children between eleven and sixteen years of age living in our sample households have completed a short self-completion questionnaire (Scott et al,1994). While the aim is to gain a full interview with every eligible adult, we also collect proxy information or conduct a short telephone interview as a means to gain basic information about as many sample members as possible. As other panels have found, the use of flexibly constructed data collection instruments and a mix of methods helps to maintain contact with sample members who might otherwise be lost altogether (Schupp and Wagner, 1996). For the first five years of the BHPS, respondents completing a full interview have received a £5 gift voucher as a token of our thanks for taking part. Young people completing a youth interview receive a £3 gift voucher. Both of these are mailed to the respondent with a thank-you letter and change of address card after the interview[3]. In addition, we use small gifts given by the interviewer at the point of interview, such as a pen with the survey logo or a small diary for example. While there is some evidence that the incentive increases response at the margins, particularly for those on low incomes such as the single elderly, it is used primarily as a means to register our thanks for the respondent's co-operation rather than being a payment for their time.

All panel studies adopt following rules which designate which sample members are to be followed and under what circumstances they should or should not be followed (Burgess, 1989; Kalton and Lepokowski, 1985). Whatever following and eligibility rules are adopted in the overall survey design impose certain constraints on how sample members are followed year on year, requiring a relatively sophisticated sample management system to be in place. In the case of the BHPS sample members are followed as they move out of a household, create a new household or rejoin a household of which they were a member at a previous wave. All members originally sampled at Wave 1 of the survey are designated as permanent sample members (PSMs) and are followed when they move, including children under the age of sixteen. As children reach the age of sixteen they become eligible for interview. New household members are included in the sample and are eligible for interview as long as they continue to share a household with a permanent sample member. The BHPS is effectively an individual level sample, as it is the individual who is followed as they move in and out of different household circumstances. As Duncan and Hill (1985) have argued there is no such thing as a longitudinal household, only longitudinal individuals.

---

[3]From Wave 6 of the survey, the voucher incentive is being mailed in advance of the interviewer calling to respondents who co-operated the previous year and to rising 16 year olds becoming eligible for interview at the current wave. Interviewers will have vouchers to hand directly to all other respondents at the point of interview. The value of the voucher has also been increased from £5 to £7 per individual interview and from £3 to £4 for a youth interview.

Households change in composition, new households are formed and households dissolve through the combined movements of individuals, making the individual the only sensible unit for tracing in a longitudinal context.

3. LONGITUDINAL RESPONSE RATES

In a panel survey the issue of how to describe response rates longitudinally becomes somewhat problematic. For each cross-sectional wave of the survey we can calculate the household response rate for all issued households or for all contacted households if we include new households created during fieldwork. Similarly we can calculate the individual response rates for each cross-sectional wave. While cross-sectional response rates give some purchase on the success or otherwise of each fieldwork period, they tell us little about longitudinal patterns of response over all the years of the survey. Nor can we assess the impact of attrition from the survey over time. From a longitudinal perspective therefore, we need to calculate the wave on wave response rates at the individual level. This is because we are dealing with a sample of individuals who move between households, making wave on wave household response rates problematic to derive. For carrying out substantive panel analyses, it is those respondents with continuous interview records, that is respondents who have done a full interview at every wave of the survey, who provide the core longitudinal information. One means of assessing wave on wave response rates is to look at the wave on wave re-interview rate at the individual level (see Table 1). Of the 9,912 respondents who did a full interview at Wave 1 of the survey, 87.7 per cent of those still eligible for interview were re-interviewed at Wave 2. At Wave 3, 90.4 per cent of eligible Wave 1 respondents who were interviewed at Wave 2 were re-interviewed. And at Wave 4, 94.8 per cent were re-interviewed. While these re-interview response rates may seem high in comparison with many cross-sectional response rates, it has been necessary to achieve these levels in order to maintain a viable longitudinal sample. Of the total 9,912 respondents at the first wave of the survey, 7,131 have continuous interview records over the four year period. This means that, after excluding those who have become ineligible at any point, we have retained 74 per cent of our original interviewed sample with continuous information for the first four years of the survey.

Table 1: BHPS Individual Re-interview Response Rates - 1991 - 1994*

	1991	1992	1993	1994
Original respondents	9912			
Continuing W1 respondents		8568 (87.7%)	7629 (90.4%)	7131 (94.8%)
All continuing from last wave		(na)	8216 (90.2%)	8278 (94.0%)
Total number interviews		9459	9032	9062

* Full individual interviews for eligible respondents at each wave.

In addition to the core longitudinal sample, we have in many cases information for respondents who have been interviewed at one or more points in the survey, but not at every wave, and therefore have discontinuous data. For example, there are 207 of our Wave 1 respondents who have completed a full interview at Waves 1, 2, and 4 but not at Wave 3. Similarly, there are 173 Wave 1 respondents who have a full interview for all waves except Wave 2. Depending on the analysis being carried out, these respondents clearly have longitudinal information which can be used. Some of the respondents with discontinuous information are members of originally sampled households at Wave 1 who were either not interviewed at Wave 1 or were under 16 years of age and have since become eligible for a full interview. At each wave of the survey approaching 200 youngsters turning sixteen become eligible for a full interview, all of whom we attempt to interview. In the three years since Wave 2 of the survey, 432 of our younger original sample members have been interviewed. Others with discontinuous information are temporary sample members who have joined the survey since Wave 1. Despite the fact that these respondents are not part of the original sample, many have been with the survey for two or three years,

providing important contextual information for any longitudinal analyses, while also adding to the overall sample size for cross-sectional analysis. When we look at the original interviewed sample according to the number of waves at which they have responded and the type of interview data collected, the percentage of sample members who have any form of longitudinal information over the four years of the survey is higher than when calculated on the basis of a full interview at all four waves. Of the 9,391 originally interviewed sample who were eligible for interview at Wave 4, 80 per cent (7,573) were re-interviewed at Wave 4. This proportion rises to 83 per cent (7,805) when information collected by proxy or via the telephone interview is included. While the proxy and telephone interview provide a somewhat limited amount of information about the respondent, they do enable us to keep these respondents within the longitudinal interviewed sample population.

4. PROCEDURES TO MINIMISE NON-RESPONSE

A number of procedures have been adopted on the BHPS to minimise non-response as far as possible. These procedures are built into the survey process as a whole with some, such as tracing respondents, being ongoing throughout the year between interview points. Running a panel requires the implementation of a range of quality control measures throughout fieldwork, all of which are aimed to maximise response and collect high quality data. On the BHPS only experienced interviewers who have previously worked on random sample surveys are employed and, where possible, the same interviewer is assigned the same households at each wave of the survey. All interviewers new the survey attend a two day briefing prior to going into the field while those who have worked on the survey in previous waves attend a one day briefing. Fieldwork is closely monitored throughout with a weekly progress chase of all interviewers to establish the current status of each household and individual sample member. Interviewers are required to make a minimum of six calls on each address at different times of day before returning the household as a non-contact. In addition, where six or more calls were made at the previous wave, the call records are fed forward to interviewers. The content, design and length of the questionnaire documents are also critical elements in gaining the on-going co-operation of sample members. However, beyond these elements which apply to any survey data collection operation, there are three areas in which a panel must commit additional resources. These are i) running a panel maintenance operation; ii) having tracking procedures in place for movers, and iii) implementing a refusal conversion programme.

4.1 Panel maintenance

An early decision taken after consultation with other panel surveys, was to develop a custom designed Panel Maintenance Database (PMDB) to keep track of our panel members. The priority in an ongoing panel survey is to maintain up to date and accurate records of the whereabouts of each sample member. The PMDB is maintained as a separate database of names and addresses of sample members for two reasons. First, the issue of confidentiality has to be considered, not only to comply with the UK Data Protection Act but also to maintain our own ethical standards as researchers in protecting our respondents. Respondents in the BHPS are given a promise of confidentiality which guarantees that their name and address will never be linked to any of the information they provide. Holding names and addresses separately from the survey database ensures we maintain this promise as direct links between the two can only be made by a limited number of authorised staff. Secondly, we update the PMDB in the year between interview points, so need some facility to do this separately from the survey data collected at each wave. In designing the panel maintenance procedures our main aim has been to keep contact with respondents through means other than the interview itself. We use a variety of techniques to do so including:

- providing a named contact person, freephone number and answerphone for respondents
- recording details of contacts with respondents between interview points

- passing any relevant information about respondents to the interviewer before each round of interviewing e.g. news of a family bereavement/illness
- an annual pre-fieldwork mailing of a short Respondent Report of research findings and activities with a confirmation of address card for freepost return
- the inclusion of a change of address card with gift vouchers and thank-you letter post-interview
- sending a £5 gift voucher incentive to any person returning a change of address card between interview points
- updating address details between interview points
- maintenance of an historical record of all addresses ever occupied for each sample member
- ongoing tracing of respondents both during and between fieldwork periods

We have taken the view that our respondents make quite a commitment in agreeing to continue with the survey and deserve some feed back about how the data they provide are being used. Anecdotal information from respondents indicates that receiving feedback about the survey in the Respondent Report is much appreciated by them, makes them feel they are contributing to a worthwhile project, and are considered to be individually important to the survey as a whole. Indeed we have many respondents who request more information than we provide in the respondent report, requests which are handled on an individual basis. Maintaining a rapport with respondents through mailings between waves encourages a feeling of belonging to the survey while providing us with an additional opportunity to update our address records. This means we can not only update our addresses at the point of interview but also in the months between interviews, a process which feeds into the survey's tracking procedures. Approximately 500 change of address cards are returned to us every year and the confirmation of address card is returned by around one third of respondents before we issue the sample into the field each year.

4.2 Tracking

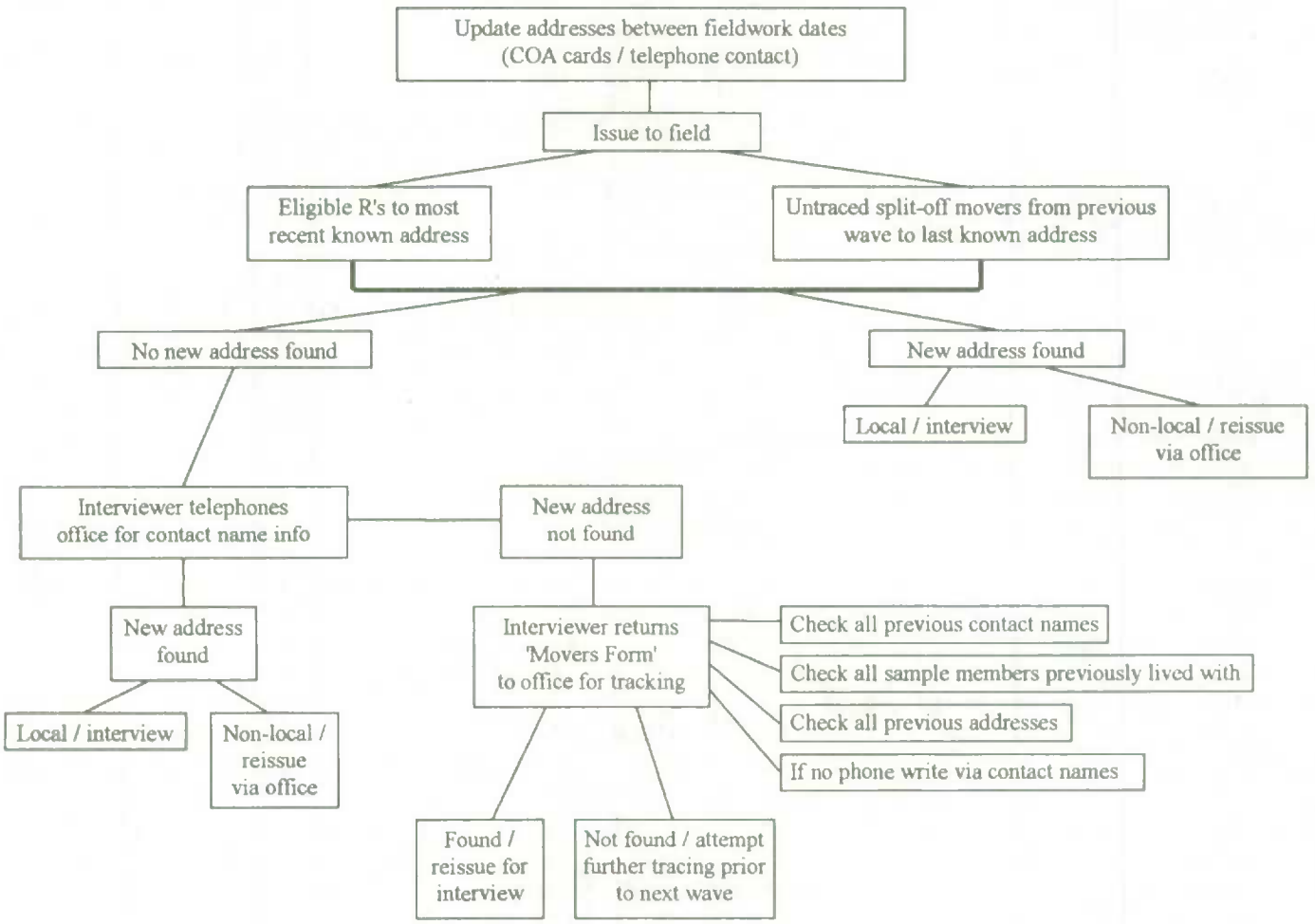
Updating addresses in between interview points so that we issue as many households as possible to the most recent address is what Burgess (1989) calls forward tracing methods. Retrospective methods are used at the point of interview when the interviewer calls, discovers someone has moved and tries to find a new address for them. Approximately 10 per cent of the sample (1,000 individuals) move in a given year. In up to one half of these cases we will have received some notification of the change of address through contacts between interview points via the change of address card, the confirmation mailing or by telephone. For the remainder, the tracking process begins at the point when the interviewer makes their first call and is unable to find a new address for the respondent. One of the advantages of a panel is that the interviewer's knowledge of the respondent's circumstances, their tracking skills and local knowledge build up over the years of the survey, increasing the chances that they will be able to trace someone without needing any help from the office. However, it is inevitable that interviewers will not be able to find everyone who moves and they then complete a Movers Form with details of the respondent(s) they are unable to find. At each wave of the survey interviewers return between 200 and 250 tracking forms to the office for further tracking. One of the main means we use for tracking is via contact names supplied by the respondent in previous years. Every year all respondents are asked for details of a contact name who would know where they are if they happened to move and in our experience this is the most effective method of tracing movers, both in terms of cost and success rate. Willingness to give a contact name may also be an indicator of how co-operative the respondent is and whether they are prepared to commit themselves to future participation. And as the years go by we have in many cases accumulated several different contact names, increasing our chances of successfully tracing movers. The tracking process is time consuming and requires a commitment of resources in terms of staff time. We estimate it costs around £10 per household in staff time and other resources such as telephone costs to carry out the tracing process. However, we successfully track 50 per cent of households for which interviewers can find no new address, which amounts to some 125 households per wave. In the context of a longitudinal panel, this relatively small number of households represents approaching 2.5 per cent of our issued households,

making a critical contribution to maintaining contact rates and minimising cumulative losses to the sample over time. When this cost is spread across the whole 5,500 households in the sample, the additional cost per household is less than fifty pence, an amount well worth spending in the context of what are relatively high survey costs overall.

4.3 Refusal Conversion

Minimising the level of refusals is the other key area to maintain high response rates. While interviewer training on how to approach the doorstep is the first and most important element in countering refusals, it is inevitable that some respondents will refuse to the interviewer. Implementing a refusal conversion programme is therefore an important element in reducing the potential losses from refusals. In our experience, many refusals are wave specific, that is the respondent refuses to take part for one year because of immediate circumstances which at the following year may have passed. The second type of refusal is those who decide they want to withdraw from the survey altogether. The most difficult question for a panel survey is assessing at what point a refusal becomes an adamant refusal to take part and when it would become unethical to attempt a conversion. In many cases we have found that we can ease the interview situation or organisation to encourage the respondent to change their mind and take part. This may simply be through talking to them and explaining the purpose of the survey more fully, or by pointing out their importance as an individual and irreplaceable sample member. Alternatively, we try and accommodate the respondent's needs where we are able. For example, they may request that the interviewer call on a particular day or time of day and, as far as possible, we attempt to meet these requests. Since Wave 2 of the survey a refusal conversion programme has been in place as a standard part of our fieldwork procedures. On receiving a refusal coversheet from field each case is assessed as follows:

Fig 1: BHPS Tracking Procedures



- reasons for refusal assessed
- review any historical information/contacts with the respondent
- decision taken on whether a conversion should be attempted
- if conversion attempted, experienced interviewer approaches by telephone
- if agreement to be interviewed is achieved, re-issue to field for interview
    - reissue to a senior interviewer or area supervisor unless respondent requests particular interviewer
    - interviewer to make the call-back within seven days of the conversion being re-issued
    - interviewer bonus payment for all re-issued conversions where interview achieved
- if telephone conversion attempt fails to gain agreement for the interviewer to return, a short telephone interview is collected
- if no conversion or telephone interview, re-assess before issuing at following wave

The telephone interview was introduced at Wave 3 as a mechanism to keep respondents who may otherwise be lost altogether in the interviewed sample. This approach has proved quite successful, with 50 per cent of the 252 respondents who completed a telephone interview at Wave 3 being converted back to a full individual interview at Wave 4. In the region of 300 households per wave go through refusal conversion. And in approaching 60 per cent of households where a conversion is attempted, either a full interview or a telephone interview is achieved. As with the tracking procedures, this relatively small number of households converted represents between two and three per cent of households in the sample, a significant proportion in the context of longitudinal response rates.

In order to deal with refusals appropriately we have found it necessary to collect as much information about the reasons for refusal as possible and to maintain an ongoing history of contacts with respondents who may have been reluctant to take part or had some problem at an earlier wave of the survey. At all waves interviewers have been asked to record the reason given for refusal at the doorstep. From Wave 3 onwards these responses have been office coded. At Wave 3 there were 719 household refusals with the main reason for refusal being that they 'Couldn't be bothered' (24% n=173). These responses relate to households where at least one interview has been achieved in the past and many of these respondents probably felt that 'I've done my share' and so couldn't be bothered by the third wave. Other common reasons given were that the respondent(s) was busy or rarely home (15% n=108) or that they were too ill or elderly (15% n=108). And in 20 per cent (n=144) of cases no reason for refusal was given. At Wave 4 the number of refusals fell by about one third compared with Wave 3 to a total of 475 households, a reduction which is partly due to the decisions made between waves about how to treat previous wave refusals. Where a respondent refuses adamantly to take part in the survey, they are withdrawn from the sample which means that each year a proportion of resolute refusals are removed, tending to make the sample increasingly co-operative over time. At Wave 4 the proportion of those initially refusing because they 'Couldn't be bothered' was 45 per cent (n=214), almost double that at Wave 3, although the absolute numbers of such refusals is only slightly higher at Wave 4 when compared with Wave 3. By Wave 4 interviewers had become more successful at eliciting reasons for refusal from respondents and the number of 'No reason given' cases was reduced by almost two thirds. It is stressed at interviewer briefings how important this refusal information is in order to tailor conversion procedures. Similar proportions of personal or family reasons such as being too busy, rarely at home, too ill or elderly were found at both waves.

A further question to examine is whether some refusals are easier to convert to interview than others. Table 2 gives the percentage of Wave 4 households who went through conversion by the original reason for refusing and the final household outcome after the conversion attempt. At Wave 4 a total of 276 households went through refusal conversion. In 26 per cent of these households at least one full individual interview was achieved, in a further 31 per cent at least one telephone interview was achieved

while 43 per cent refused once again. A conversion to telephone interview is, overall, more likely than a conversion to a full interview. However, this does vary depending on the type of reason for refusal given by the respondent. A telephone interview was more likely to be achieved where the refusal reason was survey related rather than a personal or family reason. In 69 per cent of cases where a telephone conversion was gained, the respondent had objected to something about the survey process itself or said they couldn't be bothered anymore. In contrast, a full interview rather than a telephone interview was more likely where the original reason for refusal was personal or family related Of those who were converted to a full interview, 37 per cent had given a personal reason for refusing compared with 22 per cent of those who did a telephone interview. In addition, a second refusal was more likely where the refusal reason was survey related rather than personal. To counter survey related reasons for refusal, the presentation of the survey to respondents is of prime importance so that they do not object to taking part on the grounds of length or confidentiality for example. To help counter panel fatigue respondents also need to feel that the survey is covering issues which are relevant and important in their own lives to maintain their commitment to taking part. In addition considerable care in trying to respond to respondents needs and circumstances when making contact should be taken so that sample members are not lost simply through a lack of flexibility in fieldwork procedures and arrangements

Table 2: Household reasons for refusal by conversion outcome - Wave 4, BHPS

	Conversion outcome		
	Full int.	**Tel int.**	**2nd Ref**
Reason for 1st refusal	**%**	**%**	**%**
Survey related reasons	43.7 (31)	69.4 (59)	51.7 (62)
Personal/family reasons	36.6 (26)	22.4 (19)	38.3 (46)
Other/No reason given	19.7 (14)	8.2 (7)	10.0 (12)
Total	25.7 (71)	30.8 (85)	43.5 (120)

sig <.01

5. INTERVIEWER CONTINUITY

Among the range of fieldwork procedures used on the BHPS, one of the principles adopted throughout the panel survey has been to use the same interviewers wherever possible. Anecdotal evidence suggests that having the same interviewer return every year is preferred by both respondents and interviewers. Respondents are able to build up a rapport with the interviewer, developing a relationship of trust between them. From the interviewer's perspective, they are able to maintain contact with people and families for whom they have a genuine concern. We can examine the interviewer continuity effect in relation to respondents' propensity to co-operate over the course of the panel. Since the survey began, 97 per cent of respondents have had the same interviewer for at least two of the first four waves. Of respondents who have been at the same address over the life of the panel, 46 per cent have had the same interviewer for all four waves and 18 per cent for three of the four waves. Table 3 gives the response rates for achieving a full individual interview at each wave for all Wave 1 respondents by whether they have moved address between interview points and whether they had the same interviewer at the previous year of the survey. Moving address will in many cases mean that a different interviewer calls the following year, especially where the move is non-local and we might expect that the change of interviewer could be a contributory factor in people refusing to take part. What is clear from the table below however, is that the strongest negative effect of a change of interviewer from one year to the next is within the non-mover population. The response rates for non-movers who keep the same interviewer wave on wave are without exception higher than where there has been a change of interviewer. For those who have moved, having a different interviewer does not have this effect. The expectation of those who move address may well be that their regular interviewer will not necessarily be able to make to call. In contrast, the non-mover population

may have an expectation that, as they are in the same place, the same interviewer will make the call the following year.

Table 3: Percentage of Wave 1 respondents with a full individual interview by mover status and whether same interviewer as previous year - BHPS, 1992 - 94 (all eligible R's located)

		Whether same interviewer as previous year		
		Same	Different	All
		%	%	%
Wave 1/2	Non-mover	97.0 (4784)	96.5 (3040)	96.8 (7824)
	Mover	96.6 (313)	98.2 (431)	97.5 (744)
Wave 2/3	Non-mover	92.7 (5360)	90.6 (1737)	92.2 (7097)
	Mover	95.1 (386)	93.9 (355)	94.5 (741)
Wave 3/4	Non-mover	95.9 (5616)	84.9 (1184)	93.8 (6800)
	Mover	96.6 (453)	97.6 (324)	97.0 (777)

Apart from the wave on wave effect of interviewer continuity, we can also predict the odds of a Wave 1 respondent being in a non-response household at Wave 4 by a number of key variables, including interviewer continuity over the life of the panel. In the model described in Table 4 we have included some basic demographic and socio-economic characteristics of respondents as well as information reported by the interviewer about respondent reactions to the interview. Interviewers are required to complete a series of interviewer observations after each individual interview describing whether or not the respondent was co-operative, whether they had any health problems which affected the interview and whether or not the respondent was willing to provide a contact name for tracking purposes in case they moved between waves. In terms of this model we can see that having more than one interviewer over the four years of the survey is a significant predictor of being in a non-response household at Wave 4. Respondents with different interviewers over the life of the panel were 58 per cent more likely than those with the same interviewer to be in a non-response household at Wave 4, supporting the contention that keeping the same interviewer is a positive strategy for maintaining longitudinal response rates. The interviewer observations from Wave 1 also provide predictors of non-response at Wave 4. Those coded as being 'fair to poor' respondents at Wave 1, those who did not give a contact name for tracking purposes, and those with health or other problems affecting the interview were all more likely to be in a non-response household at Wave 4. This suggests that the policy of feeding information about respondents forward to interviewers at the next fieldwork period is important so that they can tailor their approach to the household and increase the chances of maintaining contact. In addition, the propensity of a respondent to co-operate is clearly affected by the attitude of other household members to the survey. Where the household had a within household refusal or non-contact at Wave 1, respondents had a greater likelihood of being in a non-response household at Wave 4. In terms of interviewer training, the importance of attempting to gain the co-operation of all household members must be stressed, as well as having implications for refusal conversion procedures for within household refusals.

Table 4: Logistic regression predicting odds of Wave 1 respondent being in Wave 4 non-response household (category in brackets is the reference category)

	B	Sig	Odds in non-response household at W4
Response variables			
Partially co-operating hhold @ Wave 1 (Complete co-op @ W1)	.3177	.0000**	1.3739
One plus interviewers (Same interviewer all waves)	.4630	.0000**	1.5889
Interviewer observations			
Fair to poor R Wave 1 (Good to very good R W1)	.2234	.0013*	1.2504
No contact name @ Wave 1 (Contact name @ W1)	.1750	.0022*	1.1913
Health/language problems @ Wave 1 (No problems @ W1)	.1620	.0266	1.1759
Socio-Demographic/economic			
Accommodation owned (Rented accom)	-.4306	.0000**	.6501
Monthly hhold income	-9.OE-05	.0217	.9999

* sig <.01
** sig <.001

n=7,123 W1 respondents eligible at all waves with no missing information/non-mover population
Non-significant variables entered in the model: whether dependent children in household;highest educational qualification;employment status @ Wave 1 and whether member of household in the Wave 2 Interpenetrating Sample experiment.

6. CONCLUSION

Maintaining high response rates in the context of a longitudinal panel survey requires a fairly complex mix of procedures and survey systems, only some of which have been discussed here. Some of these procedures are fieldwork related and implemented directly by interviewers, while others are office based activities such as the panel maintenance, tracking and refusal conversion procedures used on the BHPS. What is clear from the BHPS experience to date, is that the additional effort expended in keeping track of panel members, the refusal conversion process, and in implementing fieldwork procedures geared specifically to the needs of the panel sample is justified. The combination of these procedures have a significant overall impact on minimising attrition and maintaining response rates at a level which ensures the continuing viability of the sample and the collection of high quality data for substantive analysis.

The support of both the Economic and Social Research Council (UK) and the University of Essex is gratefully acknowledged. The work reported in this paper is part of the scientific program of the ESRC Research Centre on Micro-Social Change in Britain.

Acknowledgements to Graham Kalton and Nick Buck for their helpful comments on earlier drafts.

7. REFERENCES

Sudman, S. and Bradburn, N.M. (1977) *Response Effects in Surveys* Aldine Publishing Co., Chicago

Burgess,R.D.(1989) 'Major issues and implications of tracing survey respondents' in Kasprzyk et al (eds) *Panel Surveys* Wiley, New York.

Corti,L. and Campanelli,C. (1992) 'The Utility of Feeding Forward Earlier Wave data for Panel Studies' in Westlake et al *Survey and Statistical Computing* North Holland, London.

Duncan,G.J.and Hill,M.S.(1985) 'Conceptions of Longitudinal Households. Fertile or Futile?' *Journal of Economic and Social Measurement* 13:361-375

Kalton,G. and Lepowski,J.(1985) 'Following rules in SIPP' *Journal of Economic and Social Measurement* 13:319-329

Kasprzyk,D.,Duncan,G.,Kalton,G.and Singh,M.P.(eds)(1989) *Panel Surveys* Wiley, New York.

Rose,D. et al (1991) 'Micro-Social Change in Britain', *Working Paper No 1*, ESRC Research Centre on Micro-Social Change, University of Essex.

Schupp, J.P. and Wagner, G.G. (1996) 'Maintenance of Long-On-Going Panel Studies - The case of the German Socio-Economic Panel Study (GSOEP)' Paper presented at the Essex '96 Fourth International Conference on Social Science Methodology, University of Essex, July 1996. Mimeo.

Scott, J., Brynin, M.and Smith, R.(1994) 'Interviewing Children in the British Household Panel Survey' Paper presented at the 5th Symposium of the Research Committee on Empirical Family Research Amsterdam, December 1994 (mimeo)

Taylor, A. (1993) 'Sample characteristics, attrition and weighting' in Buck,N. et al (eds) *Changing Households: The BHPS 1990 to 1992* ESRC Research Centre on Micro-Social Change, University of Essex.

SESSION 4

QUESTIONNAIRE DESIGN AND QUALITY MONITORING

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

COGNITIVE RESEARCH IN REDUCING NONSAMPLING ERRORS IN THE CURRENT POPULATION SURVEY SUPPLEMENT ON RACE AND ETHNICITY

Ruth B. McKay[1]

ABSTRACT

Cognitive research played an important role in the May, 1995 Current Population Survey Supplement on Race and Ethnicity. Research interviews testing successive versions of the Supplement questions made it possible to reduce survey error by identifying and correcting problems in the survey instrument. Monitoring survey interviews during the May, 1995 CPS collection week, along with content analysis of responses to open-ended multiracial and ancestry/ethnic origin questions, provided the basis for interpreting some ambiguous findings from the statistical analysis of the CPS supplement data. Thus, according to the definition of multiracial used in this research, it was possible to identify "true multiracials," and "indeterminate (false positive) multiracials." The largest group of "indeterminate multiracials" was made up of persons reporting two or more ethnicities, e.g., Scottish and Italian. The results of the cognitive research indicate that it will be difficult to eliminate nonsampling error in surveys of race and ethnicity until agreement between survey designers and respondents is reached on the meanings assigned to the terms "race" and "ethnicity/ethnic origin."

KEY WORDS: Measurement errors; Cognitive research; Racial terms.

1. INTRODUCTION

Lessler and Sirken divide nonsampling errors into errors associated with the sample frame, errors resulting from missing data for members of the sample, and measurement errors, sometimes referred to as "observational errors," which occur when the recorded measurements do not reflect the true values of the variables they represent. Measurement errors include errors caused by factors residing in: the survey questionnaire; the mode of collecting the data; the interviewer; and the respondent (Lessler and Sirken, 1985).

Cognitive research conducted by the Bureau of the Census following the 1980 and 1990 decennial censuses identified sources of measurement errors in both surveys (Martin, et al, 1990; Elias-Olivares and Farr, 1991; Kissam et al, 1993). In the 1990's, both the Bureau of Labor Statistics and the Bureau of the Census have begun to use cognitive research to reduce nonsampling errors, particularly measurement errors, in their demographic surveys (McKay and de la Puente, 1996; Bates et al, 1994).

Background

The Federal standards for Race and Ethnicity classifications, set forth in a 1977 Office of Management and Budget (OMB) Directive, include the following categories: (Race) White, Black, American Indian or Alaskan Native; Asian or Pacific Islander; (Ethnicity) Hispanic origin, and Not of Hispanic origin. Over the years since the standards were adopted, citizens who report information about themselves, as well as users of Federal statistical data, had indicated that the categories were becoming less useful in reflecting the diversity of the nation's population. In 1994, OMB convened a two-day workshop at the National Academy of Sciences to assess the adequacy of the racial and ethnic standards.

[1]Ruth B. McKay, Bureau of Labor Statistics, 2 Massachussetts Ave., N.E., Washington, DC 20212, USA.

U.S. Census figures showed an increase in multiracial children in the U.S. from 460,302 in 1970, to 1,937,496 in 1990 (Bennett, McKenney, and Harrison, 1995.) In addition, research by the Bureau of the Census had indicated that two possible changes in the racial reporting categories might reduce Hispanic undercoverage in the decennial census. Adding "Hispanic" as a racial category, and reversing the order in which respondents are asked about their race and Hispanic ethnicity, might significantly reduce the numbers of Hispanics who report themselves as "Other" on the initial race question, and then skip the Hispanic ethnicity question, on the decennial census (Bates, et al, 1993). The increasing cultural diversity of the population has also raised additional concerns about the range of interpretation of survey categories of race, ethnicity, ancestry, and national origin, especially among Hispanics (Elias-Olivares and Farr, 1991; Kissam, Herrera and Nakamoto, 1993).

The 1994 OMB Workshop led to the formation of an Interagency Advisory Committee, representing over 30 federal agencies, for the Review of the OMB Racial and Ethnic Standards. A Research Working Group was formed to assess new racial and ethnic reporting categories as well as the potential effects which changes in the categories would have on the statistics collected by the Federal government. Public input into the review of the racial and ethnic categories came from public meetings held in major regions of the US in July of 1994. By late July, the first research project was proposed: a Supplement to the Current Population Survey (CPS) that would collect information on several key issues under review.

THE CPS SUPPLEMENT ON RACE AND ETHNICITY

The CPS, a monthly (BLS/CENSUS) survey of 60,000 households representative of the civilian noninstitutional population of the U.S., routinely collects information on the race and ethnic origin of household members during the first month's interview. The opportunity to try out new versions of race and ethnicity questions in this population would provide comparative data on how respondents would identify themselves under current and modified wording conditions. For example, if the modified wording conditions included a "multiracial" reporting category, the comparison would yield a measure of the potential shift from current racial groups, e.g., Black, American Indian, to "Multiracial."

Among the research issues considered for inclusion in the CPS supplement were: (1) The effect of having a multiracial category on the list of racial categories; (2) The effect of adding "Hispanic" to the list of racial categories; (3) The effect of placing a question on Hispanic origin before the question on race; (4) Questions designed to explore the acceptability of alternative racial terms, e.g., African American for Black, Latino/a for Hispanic; (5) Conceptual questions designed to learn about the criteria which respondents use to categorize themselves and others into racial, ethnic, and ancestral groups.

The Supplement questionnaire was written by the Research Working Group with input from panels of Questionnaire Design Experts, and Subject Matter Experts. The latter panel was composed of academics who were authorities on the major racial and ethnic groups within the United States.

The Supplement was organized into four panels, or versions:

Panel I: Separate race and Hispanic origin questions; no multiracial category;
Panel II: Separate race and Hispanic origin questions, with a multiracial category;
Panel III: A combined race and Hispanic origin question; no multiracial category:
Panel IV: A combined race and Hispanic origin question; with a multiracial category.

Apart from the differences in the initial race and Hispanic origin questions, all four panels would ask the same questions on other aspects of race and ethnicity, such as ancestry and preferred racial terms. All of the panels also had the same conceptual questions in the concluding section of the Supplement.

2. METHODS

The research was conducted by a multi-racial, multi-ethnic team from several governmental agencies and from academia (1.). The research plan called for matching race and Hispanic origin group of respondent and researcher. This was achieved for all groups except for the American Indian respondents who were interviewed by an Asian-American researcher. The team was multidisciplinary as well, representing anthropology, psychology, and sociology. The research was carried out from November, 1994 through April, 1996.

2.1. Developing the Survey Questionnaire

The research protocol to evaluate the Supplement called for individual, face-to-face interviews in which the respondent was to answer all of the questions on one of the four panels to be tested. After responding to each question, the respondent would be asked to paraphrase the question, i.e., to tell the interviewer what the question meant in his or her own words. For questions containing terms of special interest to the research, e.g., race, ethnicity, Latino, the respondent would also be asked to provide a definition of these terms.

All of the materials used in the cognitive research, including the four panels of the supplement to be tested, as well as the research interview protocols, were translated into Spanish for use in the research with Hispanic respondents. The protocol for developing the Spanish translation of the Supplement called for independent Spanish translations of the English questionnaire by translators of Mexican origin, Puerto Rican origin, and Cuban origin. Differences among the three translations were negotiated in a reconciliation conference attended by all three translators.

The research plan called for testing the supplement with members of the following groups: White; Black; American Indian and Alaskan Native; Hispanic; Asian or Pacific Islander; Multiracial. Respondents for the cognitive interviews were recruited by community centers which served the various ethnic and racial populations to be included in the research. The cognitive research interviews were conducted in three phases.

Phase One Cognitive Interviews

Twenty cognitive interviews were conducted with respondents in the Washington Metropolitan area in mid-November, 1994. This number allowed for a trial of the four panels of the supplement across all race/ethnicity groups except the American Indian and Alaskan Native category which was included in Phase two interviews. The Phase One interviews revealed many problems in the four panels of the Supplement. The problems included: 1) vague or imprecise questions; 2) sensitive questions; 3) abstract questions; 4) vocabulary problems; 5) order effects; and 6) perceived redundancy. Some of these problems will be discussed below.

Vague or imprecise questions

Q.4a: You selected [fill: race] from the list I read to you. Do you also have a more specific group that you belong to?

If yes)

Q.4b: What is the name of your specific group?

Since the question did not indicate that we were asking for a specific *ethnic* group within that race, responses to Q4b ranged from "Christian", "Masons", "Black Muslims", to "Rebellious teen-ager."

Q.8a: Earlier I read you a list of groups with the following categories: White; Black; American Indian, Eskimo, or Aleut; Asian or Pacific Islander; and Something else. Would you have liked to have had a multiracial category on the list?

Respondents interpreted this as a general question, and tended to give "politically correct" answers, e.g., "Sure, it should be added to the list." Respondents did not realize that we were asking if they would have liked a multiracial category for reporting *their* racial classification.

Abstract questions

Q.5: Please tell me what you think is the most important characteristic that defines race.
Q.6: Do you think there is any difference between race, ethnicity, and ancestry?

The conceptual questions were found to be too difficult for most of the respondents, including those with some years of college. Respondents perceived these questions as too being abstract and believed that the questions were a test of their intelligence. Several respondents said that the questions were particularly difficult because they did not offer any "clues," or lists of items from which to choose an answer.

Vocabulary problems Very few of the respondents knew the meaning of the word "ethnicity," but several respondents thought that the question was asking about the ethical character of races. One woman thought that the word "characteristic" meant that we were asking her about her character.

Order effects In Panels one and two, the question asking if the respondent would have liked to have been offered a multiracial category, came after the question on the respondent's ancestry or ethnic origin.

Q.5: Now, what is your ancestry or ethnic origin?
Q.8a: Earlier I read you a list of racial groups with the following categories: White; Black; American Indian, Eskimo, or Aleut; Asian or Pacific Islander; and Something else. Would you have liked to have had a multiracial category on the list?

Respondents who had recalled a remote ancestor of a different race on the ancestry question wondered aloud about whether they should say they were multiracial, in light of how they had answered the ancestry question. This did not happen on Panels 3 and 4, where the ancestry question came *after* the multiracial question.

Redundancy/ Sensitivity For most of the respondents, such terms as race, ethnicity, ancestry, and national origin draw on the same semantic domain. To be asked different questions about what was perceived to be a single concept, led respondents to find many of the questions redundant. In its most extreme form, some respondents thought that the same questions were being asked in different ways in an attempt to "trick" them into revealing their covert racial attitudes.

Phase Two Cognitive Interviews

Following Phase One, major revisions were made to the Supplement questionnaire to correct the problems identified by the cognitive research. The revised instrument was tested in Phase Two. Fifty-four cognitive interviews were conducted in the second phase of the research. The larger number of respondents for each race and ethnicity category allowed us to include respondents with less than high school education and those with one or more years of college. The following locations and racial and ethnic groups were included in Phase Two: Albuquerque (American Indians); Chicago (Blacks); Houston (Hispanics, Whites); New Orleans (Creoles); New York (Hispanics, Whites); Rural California (Hispanics); Rural Mississippi (Blacks); Rural West Virginia (Whites); San Francisco (Asians and Pacific Islanders, Hispanics, Multiracials)

The Phase Two cognitive interviews revealed that many of the problems identified in Phase One had been corrected in the revised instrument. An example of a revision that served to correct the problem is seen in revised Q.8a.

Q.8a: Earlier I read you a list with the following groups. Would you have liked to have had a multiracial category on the list to better describe (yourself) (proxy)?

The revised question, which now asked if the respondent would have liked to have had a multiracial category specifically for *self* or *proxy* reporting, no longer elicited "yes" responses for "politically correct" or other non-relevant reasons. Other questions that had been revised continued to present problems.

> Q2: People sometimes think of customs, or language, or physical appearance, or country of birth when they think of race. What comes to your mind as most important when you think of a person's race?
> Q4: And finally, what comes to your mind as most important when you think of a person's ancestry?

Even though the revised conceptual questions now offered anchors for the respondent's answers, respondents still found them too abstract. Thus, Q.2 elicited such answers as "It's important that everyone should get along" or "Race shouldn't matter." For Q.4, respondents' answers included "People should know their medical history," and "It's important to know if all your ancestors were saved."

The key findings of the Phase Two interviews were that respondents found some of the questions to be confusing or redundant, and that the Supplement continued to evoke a negative emotional response. We also realized that the intermingling of self-descriptive questions, e.g., "What is your race?" with preference questions, e.g., "Which term do you prefer for your racial group?" contributed to the apparent redundancy.

Phase Three Cognitive Interviews

The questionnaire was revised again in light of the findings of the Phase Two research. The research team's cognitive anthropologist had suggested making the conceptual questions less abstract by asking about the concepts in relation to a specific situation. The entire questionnaire was reorganized to group the questions in each of the panels into three distinct sections: 1) self-identification questions; 2) preference questions; 3) conceptual questions.

The new version was tested in small samples of rural Whites and suburban Hispanics. Respondents in both groups had much less difficulty with this version of the Supplement. For the Whites, there was none of the community's previous suspicion that we were trying to trick them by asking the same questions in different ways. However, some problems remained.

> Q2. People sometimes think of customs, or language, or physical appearance, or country of birth, or other things when they think of the different races. What comes to your mind as most important when trying to decide what a person's race is?

Even when asked to apply the concept of race to a specific situation, i.e., characteristics/traits used in assigning a person to a specific racial group, respondents, some with college degrees, continued to have great difficulty with the conceptual questions. After the Phase 3 interviews, the decision was made to abandon any attempt to retain the conceptual questions in the supplement.

Unresolved Problems Identified by the Cognitive Interviews

The cognitive interviews revealed other problems in the survey instrument that did not lend themselves to ready solutions. These included respondents' comprehension of the term "multiracial." Although all of the respondents in the cognitive interviews offered some variant of "more than one race" in defining the term "multiracial," there were problems in their use of the term as a self-referent.

Multiracial respondents recruited for the cognitive research interviews, were selected for having parents of different racial backgrounds. However, two respondents who had been so identified as "multiracial" did not identify themselves as "multiracial." One young man of Hispanic and Black parentage chose the "Black" racial category. In cognitive debriefing, he stated that he selected that category because he is recognized as "Black" in his community. Another young man, of American Indian and Hispanic parentage, self-identified as American Indian. Cognitive debriefing revealed that he had a poor

relationship with his father and therefore wished only to be associated with his mother's group. Thus, persons who had been classified as "multiracial" by observers did not necessarily self-identify as multiracial.

The cognitive interview findings also revealed the opposite situation: persons classified by observers as members of a single racial group who self-identified as "multiracial." A college-educated White woman in a Washington suburb selected the "multiracial" category when responding to the question on race. Cognitive debriefing revealed that she chose the "multiracial" category because she was "half-Irish and half-Italian."

From the latter respondent, and from information gathered from probing respondents for their definitions of "race" and "ethnicity/ethnic group," and "ancestry or national origin," we learned that these are overlapping concepts for some non-Hispanic as well as Hispanic individuals. The relative lack of differentiation among these concepts was also evident in the answers given to less abstract questions which asked respondents to name their ancestry or ethnic origin. In some cases, respondents gave the same answer, e.g., "White," to questions about their race and their ancestry or ethnic origin. Some groups who have been in the United States for several generations, such as the rural Whites in West Virginia, could not answer the question about their ancestry or national origin; although a few recalled that they had a remote American Indian ancestor. (All of the respondents interviewed in this community had British surnames.).

2.3. Implementing the Survey

Once the final Supplement instrument was constructed, additional cognitive research was conducted to help interpret the results from the Supplement. During the CPS collection week of May 14th through May 20th, cognitive researchers monitored Supplement interviews in the Hagerstown and Tucson Computer-Assisted Telephone Interviewing (CATI) facilities on one day, and conducted focus groups with interviewers in both facilities on the following day to learn about their experiences in conducting the Supplement interviews. Also during that week, researchers accompanied CPS interviewers in Tucson and in Miami to observe how the interview was conducted in the field. Four hundred CATI interviews, two hundred in each CATI facility, were taped for subsequent behavior coding. The findings of this research helped in interpreting the findings of the statistical analyses of the CPS Supplement data.

Monitoring CATI interviews

Monitoring revealed that multiracial identification could change over the generations, and could change in *either* direction. One young woman in Panel 2, reported her mother as "multiracial," (Black, American Indian, Hispanic), her father as "Black," and her own race as "Black." A mother who reported her race as Black, and ancestry as American Indian, reported her daughter as "multiracial" (Black and American Indian). Occasionally, interviewers were asked to explain the "ancestry or national origin" question.

Field observation of CAPI interviews

Cognitive researchers observed May CPS interviews conducted in Miami and Tucson to learn about "Hispanic" and "multiracial" reporting. These observations pointed to problems with the interpretation of concepts and terms relating to these categories. The following is a summary of some of these observations.

In Miami, a Mexican-American man reluctantly chose "White" from the list of racial categories in Panel 1, gave an affirmative response to the question about wanting a multiracial reporting category, and said "Mexican-American," when asked which additional racial groups he would like to add to "White." In another Panel 1 household, a non-Hispanic White woman married to a Mexican-American, said that her daughter is of Hispanic origin, is multiracial (on race question), and listed White and Hispanic as her daughter's racial groups. In a third household, which received Panel 3, an elderly man who listed his race as White, and declined the multiracial option, included "American Indian" in his ancestry/ethnic origin response.

Focus group interviews with CATI interviewers

Interviewers reported no problem in respondents' comprehension of the Hispanic origin question on Panels 1 and 2. A number of the non-Hispanic respondents seemed to be surprised at being asked this question first, that is, it was not what they were expecting. Some non-Hispanics also seemed either confused or offended to be asked this question first, saying, "Oh, no. Nothing like that.". Some Hispanics felt they were being singled out, and that Hispanic was being emphasized too much. (Some non-Hispanics also asked why there were so many questions on Hispanic origin.)

Behavior coding of questions on multiracial status in taped CATI interviews

The questions on multiracial status in the taped CATI interviews were behavior coded for a range of interviewer and respondent behaviors The interviewer behaviors coded in this analysis included: major or minor changes in the wording of questions; verifying vague answers; probing for incomplete answers, and correct or incorrect coding of responses. Respondent behaviors which were coded included: requests for clarification, interrupting the reading of a question to give a response, offering "don't know" or refusal responses, providing inadequate or adequate answers, and any comments regarding the difficulty or sensitivity of the question.

There were relatively few problems in the way the interviewers read and coded the questions, other than occasionally not reading the entire list of races for the race questions. Interviewers would not always read the term "multiracial" accurately, occasionally substituting "multicultural," or "multinational." The coding data revealed that respondents were sometimes uncertain about how to report ancestry.

One of the problems in an interview mode is that the respondents cannot see the layout of the race and ancestry/origin questions, and therefore cannot anticipate that they would have a separate question on their ethnic origin. This might have led some respondents to include ethnicity in their answers to the race question. Some respondents said that everyone would want a "multiracial" category so that they could report everything they were, e.g., Dutch, German, etc.

3. POST-SURVEY ACTIVITIES

Following data collection, responses to open-ended questions on multiracial status, and ancestry or ethnic origin, were analyzed for content. The results of the content analysis, together with the findings of the earlier cognitive research, provided the basis for greater accuracy in interpreting the results of the statistical analysis of the supplement data, especially the ambiguous findings relating to multiracial reporting.

In the May, 1995 CPS Supplement, the multiracial response category was listed on the race question for Panels 2 and 4. In Panel 2, the race question was preceded by a separate question on Hispanic origin. In Panel 4, "Hispanic, Latino, or of Spanish origin" was a response option on a combined race/Hispanic origin question.

The breakdown of racial identities for those who chose the multiracial response option on Panel 2 and on Panel 4 may be seen in Table 1 (from Tucker et al, 1996).

We see that some respondents who identified as "multiracial" on Panel 2 and Panel 4 selected only one race on the follow-up question asking to check all races that apply. An initial concern in the statistical analysis was that the cognitive research interviews had missed the fact that some CPS respondents would not know the meaning of "multiracial". The finding of a large group of "one-race multiracials" led us to develop the categories of "Multiracial" (True) and "Indeterminate Multiracial" (False) for analyzing the CPS Supplement data. For purposes of this analysis, the following definitions were used:

Multiracial: persons who report belonging to 2 or more of the racial groups listed on the race question for that Panel;

Indeterminate Multiracial: persons who do not report belonging to 2 or more of the racial groups listed on the race question for that Panel.

Table 1. Multiracial Breakdown

	Panel			
	1	2	3	4
	%	%	%	%
Total Multiracial	-	1.65	-	1.55
no race / DK / NA	-	0.02	-	0.00
"Se" as only race	-	0.51	-	0.22
Only 1 race	-	0.53	-	0.15
WB / BW	-	0.09	-	0.16
AmerInd + 1 race	-	0.20	-	0.28
A/PI + 1 race	-	0.07	-	0.28
1 race + "Se"	-	0.16	-	0.07
Other 2 races	-	-	-	0.20
3 or more	-	0.08	-	0.21

A very large proportion of the "1-race multiracials" on Table 1 selected "Something else" as their single race, while a portion of the 2+ multiracials chose a single racial category and "Something else." From the cognitive research interviews, it had been observed that multiracial respondents who were asked to respond to race questions that did not offer the multiracial category most often chose "Something else," and listed their several racial identities in the open-ended follow-up question. An analysis of the "open-ended" responses to the follow-up question for those who identified as "Something else," under Multiracial, revealed a range of racial and ethnic designations.

"One-race = 'Something else' Multiracials"
The open-ended answers for the "1-race = 'Something else' multiracial" respondents included such diverse entries as: Creole; Eurasian; Chinese and White; Cape Verdian; German and Irish. Some of the entries, such as Creole and Eurasian, although single terms, do represent multiracial groups. Reporting two ethnicities, e.g., German and Irish, presents the overlapping of the semantic categories of race and ethnicity observed during the cognitive research interviews.

"Two Races = Race + 'Something else' Multiracials"
A wide range of open-ended answers was found for the "Something else" follow-up question for multiracials who identified as belonging to one of the races on the list, e.g., Black, White, and as "Something else." The following "Something else" entries are preceded by the first letter of the race chosen from the list of racial response options: (W) Mexican, American Indian and German; (B) Puerto Rican and German and African American; (W) Armenian; (W) Italian and Dutch and Irish. Thus, the pattern of equating ethnicity and racial groups was common in these open-ended responses as well.

From the analysis of the "Something else" entries it became apparent that some of the "One-Race = 'Something else' Multiracials," e.g., Creole, Chinese and White, *did fit* the definition of Multiracial constructed for the statistical analysis. The analysis also revealed that some of the "Two Races = Race + 'Something else' Multiracials" *did not fit* this definition, and should be classified as "Indeterminate Multiracials."

Following the analysis of the open-ended entries for one-race and two-race multiracials who identified as "Something else," the percentages of "Multiracials" (M) and "Indeterminate Multiracials" (IM) were calculated. Table 2 presents the counts of both types of multiracials for Hispanic and for non-Hispanics. Persons who reported two or more of the racial categories other than "Something else" are also included in the "Multiracial" category.

Table 2. Percentage "Multiracials" and "Indeterminate Multiracials"

	Panel 2		Panel 4	
	M	IM	M	IM
	%	%	%	%
Name 1 race				
Hispanic	2.24	10.74	0.0	0.71
Non-Hispanic	4.88	45.78	5.15	17.02
Name 2+ races				
Hispanic	3.58	4.61	22.79	0.0
Non Hispanic	26.03	2.34	52.46	1.88
Totals:	36.73	63.47	80.40	19.60

The initial racial breakdown of multiracials displayed previously in Table I indicated that about half of the one-race multiracials on both Panels had identified themselves as belonging to a racial group other than "Something else." Observations made in the course of monitoring CPS Supplement interviews at the Hagerstown CATI facility provided a lead for the investigation of factors contributing to one-race reporting by some multiracial reporters. It had been observed that occasionally, a CPS respondent who identified as "multiracial" and listed only one race, e.g., White, would include a second race in answering the later ancestry/ethnic origin question. The wording and placement of the ancestry/ethnic origin question were identical across Panels. The question reads as follows:

Now, what is (your/name's) ancestry or ethnic origin?

An analysis of entries to the ancestry/ethnic origin question for the 1-race "Indeterminate Multiracials" who had listed a single race other than "Something else" revealed that 54% of the 152 individuals in this group had listed a second race under ancestry. To learn how widespread this phenomenon was, an analysis of entries under ancestry was carried out for a random sample of 2000[2] Panel 2 and Panel 4 respondents, drawn proportional to the racial distribution for those two Panels, who had not identified as multiracial and had selected a racial response option other than "Something else." This latter group was designated "Single Race Respondents." For the Single Race Respondents, only 7% of the entries under ancestry included a second race not previously-named. These results are displayed in Table 3.

While the sizes of the two groups are too small to estimate significance of the difference between the groups, there is a basis to suggest that the existence of a second racial group in their heritage contributed to the selection of the "multiracial" designation for many of the "1-race Indeterminate Multiracials." The fact that some individuals with two racial heritages chose to self-identify as "multiracial," while another group with two racial strains self-identified as a "single race" led to further demographic investigation of these two groups. (See McKay and de la Puente, 1996.)

[2] (118 of the cases selected in the random sample of 2000 Single Race Respondents had entries under ancestry that could not be coded for a second race. These entries included such items as: "Heinz 57," "American," and "A little bit of everything.")

Table 3. Additional Races under ancestry for 1-Race "Indeterminate Multiracials and Single Race Respondents

	Additional Races	No Additional Races	Totals
(non-Hispanic) 1-Race "Indeterminate Multiracials"	82 (54%)	70 (46%)	152
(non-Hispanic) Single Race respondents	132 (07%)	1750 (94%)	1882
Totals	214	1820	2034

4. CONCLUSIONS

Cognitive research findings reduced measurement errors during the development of the instrument, during survey collection, as well as in interpreting the findings of the statistical analysis of survey data. We identified characteristics of **respondents** (understanding of concepts, social desirability/concern about revealing racist attitudes); and of the **instrument** (vocabulary, order effects, organization of questions, in both English and Spanish); which reduced these effects in the final questionnaire and informed the decision that the CPS survey not an appropriate vehicle for asking questions about race and ethnicity.

Interviewer effects were reduced by interviewer training which stressed need to read all of the response options on the race question; a special training tape for Spanish-speaking interviewers to facilitate use of the Spanish instrument. In addition, the observed interviewer problem in pronouncing "multiracial" correctly, led us to consider possible underreporting of multiracial by respondents who had not been read the correct term. The possibility of underreporting of "multiracial," and "Hispanic," as racial categories was also supported by **respondent** characteristics observed during survey implementation. These included the tendency to report a second race under ancestry, frustration at having to listen to long list of response options which might have led to "tuning out" new response options, such as "multiracial," which occurred after their familiar response option was read. We identified possible **mode** effects in an interview survey in which respondent could not anticipate a later question on ancestry or national origin, and therefore listed ethnicity under race question.

1. (118 of the cases selected in the random sample of 2000 Single Race Respondents had entries under ancestry that could not be coded for a second race. These entries included such items as: "Heinz 57," "American," and "A little bit of everything.")

Cognitive research findings were critical in accurately interpreting ambiguous and counterintuitive findings from the statistical analysis of data on multiracial reporting. Respondents' confusion of race, ethnicity, and ancestry during the cognitive interviews led to a content analysis of "open-ended" answers for those reporting "something else" as a race under multiracial. This led to the discovery of the "one race true multiracial," e.g., Creole, Eurasian; and the "two race indeterminate multiracial," e.g., White and German, Black and African American.

Thus, the findings from the cognitive research activities carried out in conjunction with the development and implementation of the CPS Supplement on Race and Ethnicity also helped to reduce measurement error associated with the statistical findings of the Supplement, including the fact that over half of respondents reporting multiracial status did not fit the survey definition of multiracial, and that some non-multiracial reporters traced descent from more than one race.

ACKNOWLEDGMENTS
I would like to acknowledge the valuable contributions to this research by the other members of the Cognitive Research Team: Adalberto Aguirre (U. of California, Riverside), Patricia Bell (Oklahoma State U.), Ada Costa-Cash (CENSUS), Manuel de la Puente and Eleanor Gerber (CENSUS), LuAnn Moy

(GAO), Jorge Nakamoto (Aguirre International), and Jaki Stanley (NASS/DOA); as well as Clyde Tucker, Brian Kojetin, Shail Butani, and Linda Stinson (BLS) and Betsy Martin (CENSUS).

NOTES
The views expressed are those of the author and do not necessarily reflect those of the Bureau of Labor Statistics.

5. REFERENCES

BATES, Nancy A, M. DE LA PUENTE, T. J. DE MAIO, and E.A. MARTIN, "Research On Race And Ethnicity: Results From Questionnaire Design Tests," Proceedings of the Census Bureau's 1994 Annual Research Conference, Rosslyn, pp. 107-136.

BENNETT, Claudette, N. McKENNEY and R. HARRISON, "Racial Classification Issues Concerning Children in Mixed Race Households," Paper presented at the annual meeting of the Population Association of America, San Francisco, California, 1995.

BUREAU OF THE CENSUS, Statistical Abstract of the United States: 1993 (113th edition.), Washington, DC, 1993.

DE LA PUENTE, MANUEL and RUTH B. McKAY , "Developing and Testing Race and Ethnic Origin Questions for the Current Population Survey Supplement on Race and Ethnic Origin," Proceedings of the 1995 Annual Meeting of the American Statistical Association, Section on Government Statistics, pp. 19-28.

ELIAS-OLIVARES, L. and M. FARR, Sociolinguistic Analysis of Mexican-American Patterns of Non-Response to Census Questionnaires. Report Submitted to the Census Bureau under Joint Statistical Agreement 88-25, Ethnographic Exploratory Research Report #16, 1991.

KISSAM, Edward, E. HERRERA, and J.M. NAKAMOTO, Hispanic Response to Census Enumeration: Forms and Procedures, Report submitted to the Census Bureau under Contract No. 50-YABC-2-66027, Task Order No. 46-YABC-2-0001, March, 1993.

LESSLER, J. T. and M.G. SIRKEN, "Laboratory-Based Research on the Cognitive Aspects of Survey Methodology," Milbank Memorial Fund Quarterly/Health and Society, Vol. 63, 1985, pp. 565-581.

McKAY, RUTH B. and M. DE LA PUENTE, "Cognitive Testing of racial and ethnic origin questions for the CPS Supplement," Monthly Labor Review, Vol. 119, No. 9, 8-12., Sept., 1996.

SNIPP, MATTHEW C., "Who Are American Indians? Some Observations About the Perils and Pitfalls of Data for Race and Ethnicity." Population Research and Policy Review 5: 237-252, 1986.

TUCKER, N. CLYDE, R.B. McKAY, B.A. KOJETIN, R. HARRISON, M. DE LA PUENTE, L. STINSON and E. ROBISON, "Testing Methods Of Collecting Racial And Ethnic Information: Results Of The Current Population Survey Supplement On Race And Ethnicity," Bureau of Labor Statistics Statistical Note Series, No. 40, June, 1996.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

QUALITY MEASUREMENT IN SURVEY PROCESSING

Kathryn Williams, Connie Denyes, Mary March, Walter Mudryk[1]

ABSTRACT

While avoidance of nonsampling errors is built into the design of surveys, quality control techniques are needed to ensure quality standards are met when carrying out surveys. The processing of survey data involves many complex procedures where mistakes can occur: data may be captured or coded incorrectly, manual editing procedures may be misapplied. Acceptance sampling and control, statistical process control and pareto analysis are statistical techniques used to control these errors during production. This paper will describe how these statistical methods are incorporated into a framework for managing the quality of survey processing within Operations and Integration Division at Statistics Canada. Examples of existing quality control applications will be used to demonstrate this structured program of quality measurement and continuous improvement.

KEY WORDS: Acceptance sampling; Acceptance control; Statistical process control; Quality standards; Quality management framework; Continuous improvement.

1. INTRODUCTION

Statistics Canada considers quality a high priority, and with good reason. Our reputation depends on it. In introducing Symposium 90, Dr. Fellegi remarked that "our reputation for producing reliable statistical information is . . . a direct determinant of the usefulness of our output [as] . . . few users have an opportunity to replicate or otherwise directly assess the quality of our output: they must rely, in the final analysis, on the reputation of the producing agency." This reputation has been gained through constant attention to quality at all stages of a survey, whether it is deciding on the content of the survey, reducing sampling error through improved designs, increasing response rates through computer assisted interviewing, reducing respondent errors with editing or reducing the errors introduced during data capture, coding or manual editing.

This paper focuses on the initiatives of the Operations and Integration Division (O&ID) for ensuring the quality of the processing services they provide for many of the agency's programs. These initiatives are consistent with practices common in industry such as setting quality standards and using statistical quality control methods to ensure the work performed meets or exceeds clients' expectations. These practices are grounded in a strong foundation laid by the pioneers of statistical methods for quality management: Juran (1988), Deming (1986), Shewhart (1931). This paper describes the statistical quality control methods used at Statistics Canada and how they are incorporated into a framework for managing the quality of processing operations. Examples are provided to illustrate how these methods and the framework are used to prevent or reduce nonsampling errors in survey processing.

2. ISSUES RELATING TO NONSAMPLING ERRORS

The nonsampling errors occurring in survey processing are usually associated with collection, capture, coding and editing. The following issues are faced when addressing these errors:

- What quality standards must be met?
- How do we measure the nonsampling error in data capture, coding or other processes? Measurement

[1]Kathryn Williams, Mary March, Walter Mudryk, Business Survey Methods Division, Connie Denyes, Operations Research and Development Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

comes first, because as Deming said "If you can't measure it, you can't improve it."

- How do we reduce this error?
- How do we ensure that continuous improvement occurs? Monitoring quality on an ongoing basis and providing feedback is essential.
- How do we identify the root causes of errors and prevent them from happening again?

These five issues can be illustrated with the following true example. We measured a level of error which was higher than the quality standard for data capture in the case of one code on a questionnaire. We reduced the error through inspection and correction. Then we analyzed the quality control results and the root causes of the errors. We found that the codes to be captured were circled with a green pen and the keyer could not see them because the questionnaire is also green. So a red pen was used in the future. Simple, but effective prevention.

The final issue faced relates to the removal of the stigma of Quality Control. We must actively work to ensure that everyone involved understands that we are not the "quality police". We are measuring errors to identify and to feed back what needs to be improved.

3. STATISTICAL QUALITY CONTROL METHODS

Statistical quality control methods involve determining what we need to control, setting the quality standards, selecting appropriate statistical techniques, measuring actual performance, analysing the difference (between the actual and the standard) and finding and correcting the cause for the difference.

The three statistical techniques used most commonly at Statistics Canada are acceptance sampling, statistical process control and pareto analysis. These are usually used within the framework of Acceptance Control (Schilling, 1982). The refinement and application of these tools for survey processing and the development of the associated support systems have been documented in numerous internal reports and papers which include Mudryk (1988), Mudryk, Croal and Bougie (1994) and Mudryk, Burgess and Xiao (1996).

3.1 Acceptance Sampling

An important aspect of acceptance control is the underlying theory of acceptance sampling. In the survey operations context, acceptance sampling involves dividing work to be performed by an employee into homogeneous batches and randomly selecting a sample of items of work from each batch to be checked for errors (and corrected). If the total error count for a sample is less than a specified acceptance number, the batch is accepted. If not, the batch is rejected and inspected completely and the remaining errors are corrected.

Sample sizes and acceptance numbers (n, c) are based on a specified quality standard. This standard, the Average Outgoing Quality Limit (AOQL), minimizes inspection at an expected incoming error rate and assumes certain levels of risk for the producer and the client (Dodge & Romig, 1959). The standard on which an acceptance sampling scheme is based cannot be chosen arbitrarily since the tighter the standard, the greater the probability that batches will be rejected and the higher the rate of inspection and therefore the cost.

Figures 1 and 2 illustrate the relationship between the incoming error rate, the outgoing error rate, the amount of inspection required for a single sampling acceptance plan $(n, c) = (4, 5)$ for a batch size $N=30$ and the quality standard or Average Outgoing Quality Limit (AOQL) of 70 defects per hundred units assuming rectifying inspection. This example is a data capture operation for a questionnaire with 200 fields where the maximum level of error was set at 70 fields in error per 100 questionnaires or less than 0.4% of the questionnaire fields in error (Duddek, 1996). If the total number of fields in errors in the sample (4 questionnaires from a batch of 30) is less than or equal to the acceptance number ($c=5$) the batch is deemed to be acceptable. If the total is greater than the acceptance number, the batch is rejected and the rest of the batch is inspected.

Figure 1

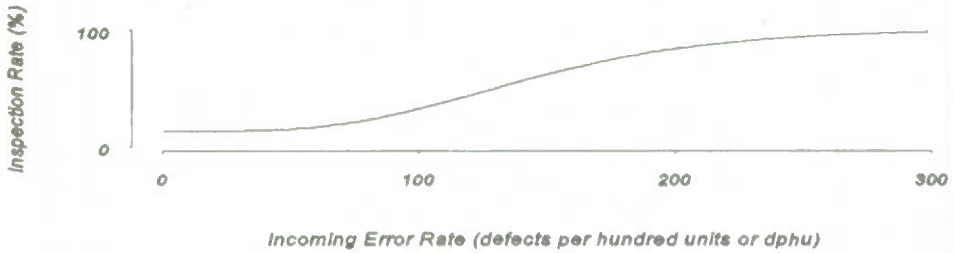


Figure 2

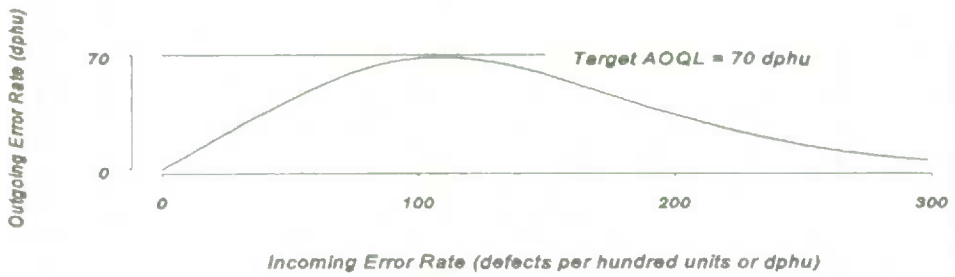


Figure 1 shows that as the incoming error rate increases, the level of inspection also increases. This occurs because the more errors in a batch, the greater the chances that the batch will be rejected. Figure 2 shows what happens to the average outgoing quality (AOQ) of the batches under rectifying inspection. Rectifying inspection requires that identified defects be removed through a correction process. As the incoming error increases, the outgoing error also increases to a maximum called the AOQL. After this point, the average outgoing error rate decreases because more and more of the batches will be rejected and defects corrected. The AOQ takes the form

$$\mathrm{AOQ} = \frac{P_a\, p\, (N-n)}{N}$$

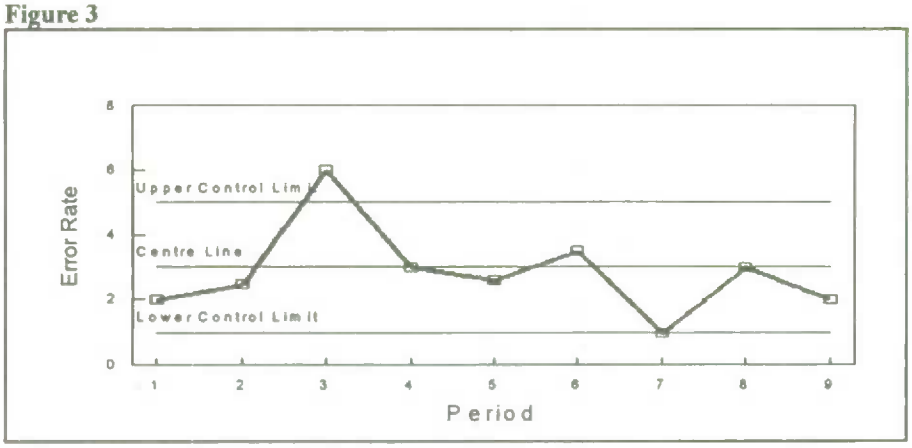
where p is the fraction of incoming errors in a batch and P_a is the probability that the number of errors, d, is less than or equal to the acceptance number, c. P_a takes the form

$$P_a = P\{d \le c\} = \sum_{d=0}^{c} \frac{n!}{d!(n-d)!} p^{d} (1-p)^{n-d}$$

3.2 Statistical Process Control

Another statistical technique that is being used more and more at Statistics Canada is Statistical Process Control (SPC). It can easily be described using a control chart for a typical case. As Figure 3 illustrates, the control chart consists of a centre horizontal line representing the expected value of a process parameter such as the average error rate and two parallel lines known as upper and lower control limits. The two lines are normally set at three standard deviations from the centre line. The chart may also contain upper and lower warning limits set at two standard deviations from the centre line.

Figure 3



The control chart generally works as follows. Random samples taken at regular intervals of time are used to obtain estimates of error rates. These estimates are then plotted on the control chart showing their relationship to the expected value and control limits. If a plotted point falls outside the control limits or the data shows a trend, the process is stopped and action is initiated to find the cause for this "out-of-control" situation and to eliminate it. With Statistical Process Control, process stability is managed objectively using the statistical limits that are on the control chart (Shewhart, 1931).

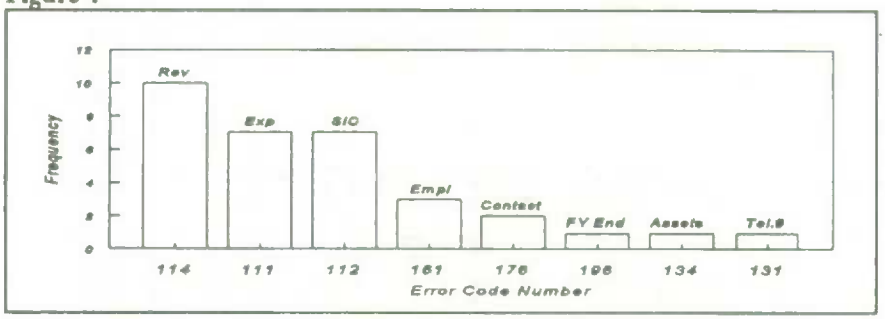
Statistical Process Control is most appropriate in situations where a process has been running for some time and is considered to be relatively stable. It provides a mechanism to reduce costs since sampling rates are generally much lower than with acceptance sampling. Statistical Process Control is more than a statistical technique. By requiring the commitment to finding the root causes of errors in "out-of-control" situations, it allows for continuous improvement.

3.3 Pareto Analysis

Finally, the Pareto Chart is an important tool for analysis and feedback. It is a special case of a frequency distribution where the error categories are presented in order from the most to the least frequent. It is based on the Pareto Principle which suggests that most of the errors usually result from only a few of the possible causes. The categories in the chart normally represent different types of errors. In practice, the first few categories often account for a very significant proportion of the errors occurring as an operation begins. The Pareto distribution focuses attention on these errors quickly enabling cause and effect analysis and early corrective action.

An example of a Pareto chart is shown below in Figure 4. This example presents the Pareto distribution of data capture errors for fields on a questionnaire. The revenue field had the highest frequency of errors, followed by expenses and then the Standard Industrial Classification (SIC) code. Investigating the causes of the errors in these three fields will significantly reduce the overall data capture error.

Figure 4



3.4 Acceptance Control

At Statistics Canada, many of the processing operations are short in duration and therefore the process stability needed for Statistical Process Control is not possible. In this case, acceptance sampling plans are used, but within a framework called Acceptance Control. Acceptance Control involves "a continuing strategy of selection, application and modification of acceptance sampling procedures to changing inspection environment" (Schilling, 1982). In longer life-cycle operations, initial instability in the process will require acceptance sampling. As the process stabilizes, transitional acceptance sampling methods such as reduced sample inspection and skip-lot sampling plans are used. The ultimate goal is to eventually phase out acceptance sampling inspection and replace it with process control methods through continuous improvement efforts. Through the use of Acceptance Control, it is possible to prevent or reduce many nonsampling errors.

4. QUALITY MANAGEMENT FRAMEWORK

The statistical methods just described are used as part of a structured program for managing quality in Statistics Canada's central survey processing area. Operations and Integration Division provides processing services such as collection, capture, coding and editing for many of Statistics Canada's surveys. Their quality management framework has several objectives:

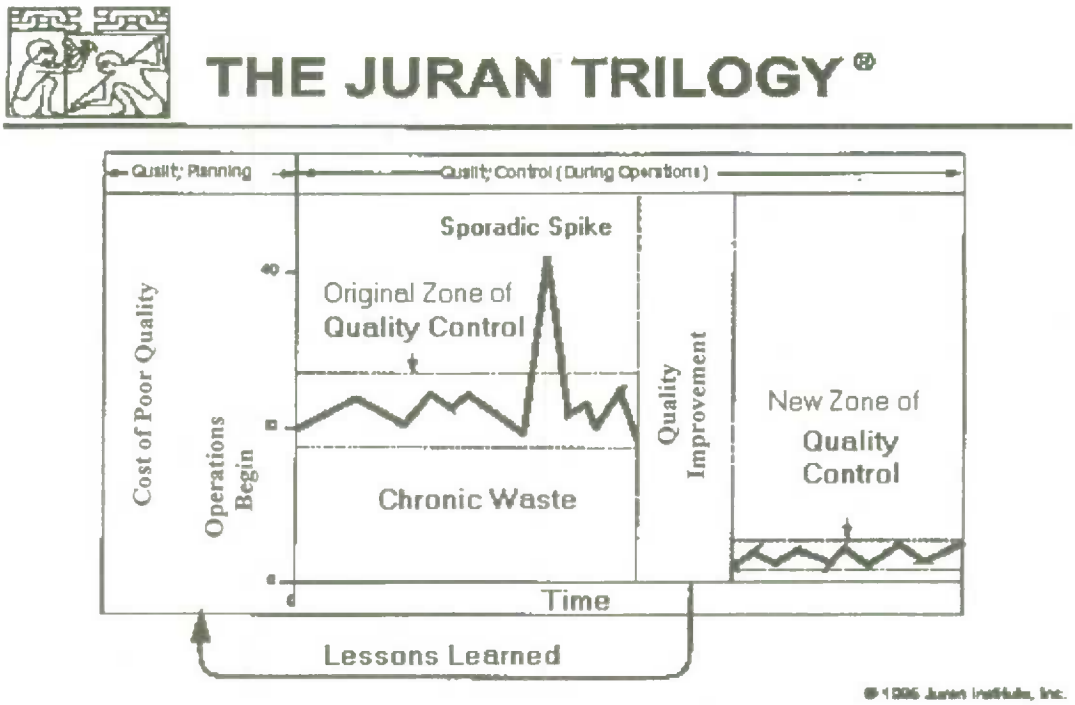
- To ensure that the products provided to clients meet or exceed their quality expectations;
- To assure the Division's management that operations are performing as expected and specified standards are being met;
- To continuously improve the operations by improving quality and reducing costs.

This framework facilitates continuous improvement of survey operations by addressing quality planning, quality control and quality improvement. It follows the model of quality management used in industry which was developed by Juran (1988) in his Trilogy chart by applying an ongoing measurement, feedback and improvement cycle. This model has been adapted and documented for survey operations by Mudryk (1991).

Managing for quality involves three standard managerial processes: planning, control and improvement. As shown in Figure 5, at the **quality planning** stage, products and processes are strategically designed to meet specific needs and the quality goals are set. In the example chart, the initial cost of poor quality is shown to be 20% of the units defective.

Once in production, **quality control** ensures that quality goals established at the planning stage are met. Statistical methods, such as those already described, are used to identify errors or quality deficiencies. Quality control ensures that any sporadic increase in poor quality - such as the spike to 40% on the chart - are quickly returned to specified levels. But there is an ongoing level of error that remains, implying chronic waste. This is associated with the cost of rework of rejected lots or dealing with errors and corrections later in the process. Causes for this chronic level of error can often be traced back to the initial design in the planning stage.

Figure 5



Quality improvement activities involve investigating the ongoing level of error (chronic waste) to determine causes and identify lasting corrective action which makes the process better over the long term. After the efforts for quality improvement, a new zone of quality cost is established at 3% on this chart. The lessons learned about the process and the causes of poor quality are then used in quality planning for the future.

This quality management framework can best be described by following an application through the process. It begins with identification of a need for quality control by either the operations area or the survey manager. A team of people from operations, the survey division and methodology are involved and their first step is to review the operation considering such aspects as process, procedures, constraints, volume, resources and time schedules. Once they have determined where quality control is needed, they set the quality standards and determine the statistical technique to use. Setting the standard involves balancing the level of quality required with the costs and the risks. The statistical quality control methods used at Statistics Canada are generally Statistical Process Control (SPC) for computer assisted telephone interviewing (CATI) operations and Acceptance Control for most other types of processing (e.g., data capture, coding, etc.).

To have an effective process as well as an effective quality control application, clear processing specifications are required and the team ensures that these exist. Error codes are defined from the specifications and the different types of errors may be weighted if some types of errors are deemed to be more critical than others. Once all the various parameters are determined and loaded into the appropriate support systems, everything is tested. Procedures are written for quality control verifiers or monitors who also need training in error identification, recording for feedback and actions to be taken.

Once production has started, the quality control verifiers redo a portion of the work or in the case of CATI, monitors part of selected telephone interviews, all according to predetermined sampling plans. Errors are identified and corrected, with immediate feedback provided for those which are critical, so that the process remains in control. All errors are recorded and after the results have been collected and summarized, reports are provided to the team for analysis. They look for root causes of errors and corrective actions which will have lasting results such

as changes to processes, methods, procedures, manuals, or training. Feedback is also given to the operators / interviewers in order to eliminate some errors and to inspire self-improvement. Finally, the lessons learned from this analysis are considered in the planning stage so that continuous improvement to the process can occur.

This quality management framework has many benefits:

- It verifies that the process is "in control".
- It guarantees a level of quality in products delivered to clients.
- The most frequent errors can be pinpointed so that action can be taken first where there will be the most benefit.
- The results and feedback provide an excellent tool for training and coaching and result in more knowledgeable and skilled employees processing surveys.
- It provides a structure for continual improvement, and costs decrease.

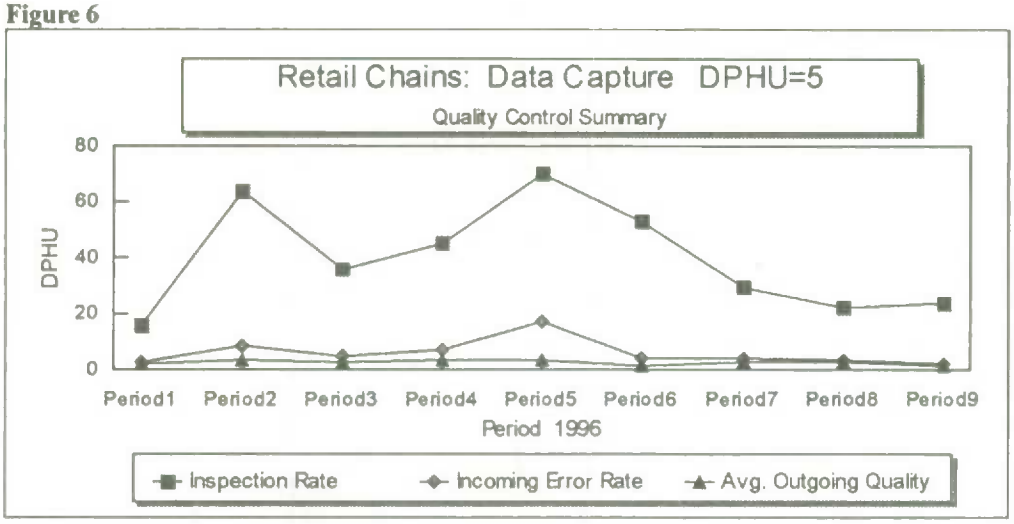
5. QUALITY CONTROL APPLICATIONS

The use of the statistical methods and quality management framework can be best illustrated with actual examples implemented in Operations and Integration Division.

The first example involves the use of Statistics Canada's generalized collection and capture software, DC-2. This software, which has a built-in Quality Control Module, allows the recapture and on-line comparison of data to determine the quality of data capture. One survey currently using this quality control technology is the Annual Retail Chain and Department Store Survey. A standard was set for the acceptable level of error, in this case, 5 defects per hundred units (AOQL=5 dphu). This means that for approximately 1,000 fields on one hundred questionnaires, less than 5 fields should have errors. The acceptance sampling plans ensure that the level of outgoing quality is always less than this standard. The quality control results are reviewed by the quality team to ensure the quality standards are being met and to identify improvements that could be made in the data capture process. The results and improvements are fed back to the keyers.

As Figure 6 illustrates, when the incoming error rate is higher the inspection rate is also higher. With this example, some errors were occurring because keyers tended to miss a field since it was not usually reported. A review of the quality control results with the individual keyers helps to draw attention to the fields with the most errors. This figure shows a number of fluctuations in the incoming error rates indicating an increase and then a decrease as steps are taken to find the root cause of the error. With this example, the work was started by an experienced keyer, then the two peaks in the incoming error occurred because new staff began working on the capture of this survey.

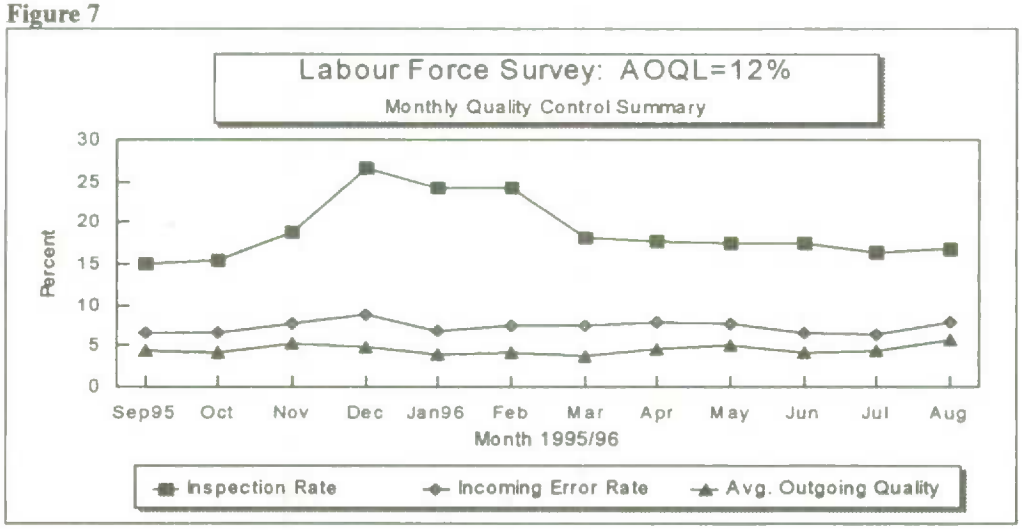
Figure 6



A second example involves quality control techniques developed for the Labour Force Survey Industry and Occupation coding operation. This application was developed to ensure a quality standard for records requiring manual coding because the assigning of codes is highly subjective and complex. In this case, the quality standard or AOQL was set to allow 12% of the manually coded questionnaires to have errors. This fully automated quality control application is unique because it uses a majority-rule strategy which requires a second level of independent verification which involves a third independent coding when the first two codes are different. Whichever two out of three of the codes are the same, identifies the correct code. With this methodology it is possible to monitor individual Standard Industrial Classification (SIC) codes and/or Standard Occupational Classification (SOC) codes and therefore, provide quality control by code. The results are fed back to the coders so that they can focus on understanding the codes with the most errors.

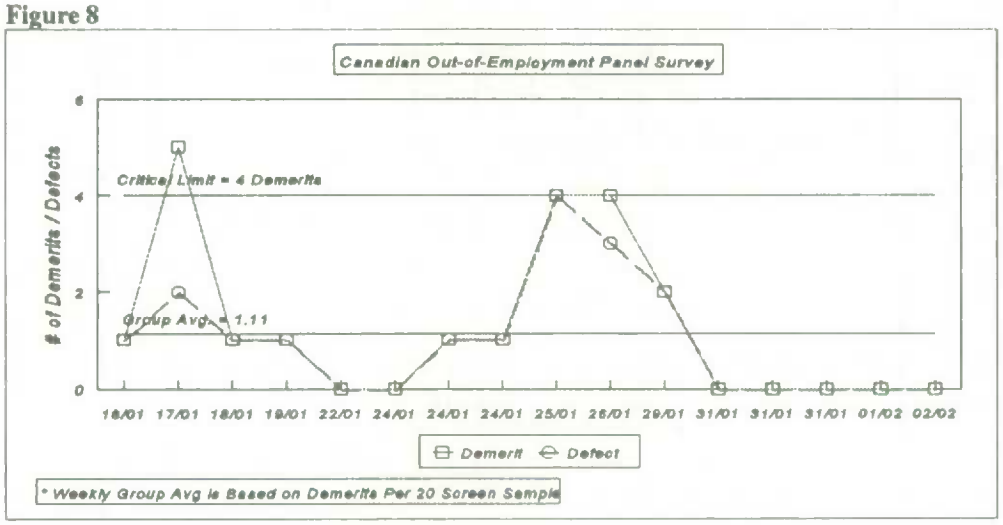
Figure 7 shows that the average outgoing quality is in the range of 4% to 6% for the twelve months. This is 2 to 4 percentage points lower than the incoming error rate, with generally about 15% to 20% inspection required. This lowering of the error rate by as much as 50% in March 96 for example, illustrates the benefits of quality control.

Figure 7



The third example involves the Computer Assisted Telephone Interviewing (CATI) operations which use Statistical Process Control (SPC) to evaluate process stability by comparing the quality results of monitored interviews against preestablished control limits. Figure 8 shows some results from the Canadian Out-of-Employment Panel Survey. The upper control limit is user-specified at four demerits (weighted errors) and the lower limit is set to zero. Samples of interviewing sessions are monitored and any problems with the interviewers' performance such as deviations from the script are recorded as demerits. If the total demerits fall within the control limits, no corrective action is taken. If, on the other hand, the observation falls on or above the upper limit, then demerits are fed back immediately to the interviewer for corrective action. Both positive and negative feedback is given to the interviewers, as appropriate. These efforts of dealing with process problems early in the operation ensure continuous quality improvement for each CATI operation.

Figure 8



The above examples represent a small sample of the many quality control applications that have been implemented in Operations and Integration Division.

One final example involving Statistical Process Control is worth mentioning. A quality control application is being developed for the Automated Data Entry System (ADES) which uses Electronic Imaging and Intelligent Character Recognition. Once implemented, both machine and operator errors will be monitored and corrective action will be taken when the error level is outside of the control limits. It is essential that in perfecting and promoting this new innovative cost-effective system, the quality of the output is measured and controlled.

6. CONCLUSIONS

This paper has shown how statistical quality control methods and the quality management framework help to prevent or reduce nonsampling errors. Deming stated that "A standard (as of performance, safety or capability) to have meaning must be defined in statistical terms." Setting such a performance standard and monitoring and improving performance against this standard using statistical methods should be a part of any survey development. The consideration of quality control requirements should be as integral a part of survey development as the consideration of editing. It is possible for these methods to be applied to both automated (CATI, ADES) and manual processes. This structured framework for managing and continuously improving process quality may complement re-engineering efforts or major redesigns of survey processing.

These good practices of Operations and Integration Division not only ensure that they deliver the quality standards agreed upon with their immediate client, but also help to build Statistics Canada's reputation.

7. REFERENCES

Deming, W.E. (1986), Out of the Crisis, Cambridge, MA: MIT Press.

Dodge F., Romig H., (1959) Sampling Inspection Tables, 2nd ed., John Wiley & Sons, New York.

Duddek, C. (1996), Modelling Quality Control Strategies for the 1996 Census of Agriculture, 1996 ASA Proceedings.

Early, J.F. (1990), Managing Quality in National Statistics Programs, Proceedings of Statistics Canada Symposium 90 Measurement and Improvement of Data Quality, Ministry of Industry, Science and Technology.

Fellegi, I. (1990), Opening Remarks, Proceedings of Statistics Canada Symposium 90 Measurement and Improvement of Data Quality, Ministry of Industry, Science and Technology.

Juran, J.M. (1988), Juran's Quality Control Handbook, 4th ed., McGraw-Hill, New York.

Montgomery, D. C., Introduction to Statistical Quality Control, 2nd ed., John Wiley & Sons.

Mudryk, W. (1988), Quality Control Processing System for Survey Operations, Survey Methodology, Vol. 14, No. 2, December 1988, pp 309-316, Statistics Canada.

Mudryk, W., Burgess, M.J., Xiao, P. (1996), Quality Control of CATI Operations in Statistics Canada, 1996 ASA Proceedings, Section on Survey Research Methods, Chicago, Illinois.

Mudryk, W., Croal, J., Bougie, B. (1994), Generalized Data Collection and Capture (DC-2) Release 2.5.1, Sample Verification (SV) Quality Control Methodology Manual, Statistics Canada.

Mudryk, W. (1991), Quality Management: A Framework, Internal Report, Statistics Canada.

Schilling, E. (1982), Acceptance Sampling in Quality Control, Chapter 19, Marcel Dekker Inc., New York.

Shewhart, W.A. (1931), Economic Control of Quality of Manufacturing Processes, ASQC, Milwaukee, WI.

Proceedings of Statisics Canada Symposium 96
Nonsampling errors
November 1996

MONITORING COMPUTER-ASSISTED TELEPHONE INTERVIEWING AT THE U.S. BUREAU OF THE CENSUS

Mary Ellen Beach, Jane Woods and Geraldine Burt[1]

ABSTRACT

Traditionally, the U.S. Bureau of the Census has placed uppermost value on providing high quality data while achieving high response rates in its surveys. Interviewer monitoring is one of the primary tools used by the Census Bureau for quality assurance in its telephone centers. Monitoring provides a means of observing interviewer behaviors that contribute to interviewer-induced nonsampling error and/or detract from eliciting respondent cooperation and obtaining accurate and complete data. The purpose of monitoring is to ensure that: 1) proper procedures are followed with regard to survey guidelines, 2) proper telephone etiquette and techniques are observed, and 3) the survey instrument is working in accordance with the sponsor's requirements and needs. This paper describes the system of monitoring developed for the three Computer-Assisted Telephone Interviewing (CATI) centers operated by the U.S. Bureau of the Census. The current system evolved over a period of ten years and provides both immediate feedback on interviewing behavior and a "score" that becomes part of the formal performance measurement system.

KEY WORDS: Monitoring; CATI; Interviewers; Quality Assurance; Feedback; Evaluation.

1. INTRODUCTION

1.1. Overview

The U.S. Bureau of the Census operates three Computer-Assisted Telephone Interviewing (CATI) centers, each employing a staff of 80-110 interviewers to collect data on a wide variety of demographic and economic surveys. The Hagerstown Telephone Center (HTC) in Hagerstown, MD, about 75 miles from Washington, DC, opened in 1985. A second CATI center opened in Tucson, AZ, in 1992. The Tucson Telephone Center (TTC) offers additional Spanish language interviewing capability, as well as better time zone coverage and back-up capacity, increased capacity, disaster recovery and options for back-up capability. In 1994, the Census Bureau established a "multi-functional" CAI (Computer Assisted Interviewing) facility in Jeffersonville, IN. The Jeffersonville Telephone Center (JTC) was designed to support CATI and other computer-assisted interviewing technologies, such as computer-assisted data entry (CADE), touchtone data entry (TDE), and voice recognition entry (VRE). In this way, the investment in facilities and staffing could be shared by various sponsors using the widest range of available technologies.

Depending on the survey, CATI may (1) share the workload with field interviewing, (2) offer low-cost nonresponse follow-up to mail-out/mail-back surveys or censuses, or (3) be the sole means of data collection. The CATI Centers currently have a total of 250 workstations.

1.2. Technical Aspects

Interviewers at the three facilities conduct a wide range of surveys, using either MicroCATI or CASES software. Surveys that have been done by CATI for a long time, such as the Current Population Survey (labor force data), the National Crime Victimization Survey (crime data), and the Quarterly Apparel Survey (industry data) are written in microCATI. Surveys that are new to CATI administration, such as National Survey of Fishing, Hunting, and Wildlife Associated Recreation (recreational data), the American Community Survey (census long-form content),

---

[1]Mary Ellen Beach, Geraldine Burt, U.S. Bureau of the Census, Washington, DC 20223; Jane Woods, U.S. Bureau of the Census, Jeffersonville, IN 47132.

and an Employers survey were developed using CASES. MicroCATI software was developed and is maintained by the U.S. Census Bureau. It is gradually being replaced by a commercial product, CASES, or Computer-Assisted Survey Execution System, from the University of California at Berkeley.

2. CATI MONITORING AT THE BUREAU

2.1. Why Monitor?

One of the major sources of nonsampling error in personal and telephone surveys is the error introduced by the interviewer. At the Census Bureau's three CATI facilities, interviewer monitoring is used to prevent and identify interviewer-induced error. Monitoring at the Bureau is used to: (1) evaluate an interviewer's ability to apply survey concepts and procedures during the actual work situation; (2) assess a new interviewer's job performance; and (3) identify ways to reduce nonsampling error by improving the survey process. An interviewer can be monitored at any time, and each center trys to monitor current interviewers about 2.5-5.0% of their login time. If an interviewer is working multiple surveys, attempts are made to monitor him/her on all surveys worked.

2.2. Who Monitors?

In the Bureau's three telephone centers, monitoring is conducted by supervisors or specially trained "coaches". Two facilities use supervisors and the third site hires and trains non-supervisory "coaches" to monitor the interviewing staff. The approaches used by the various sites reflect differences in philosophy regarding monitoring. The Jeffersonville, Indiana site wanted to remove supervision from monitoring in order to promote a mentoring/coaching approach, rather than a evaluative approach. The hope is that an interviewer will be more at ease and receptive to constructive feedback from a person that is not in the direct supervisory chain of command. There has been no evaluation to determine which method works best; however, at least one of the other sites is taking the "coach" approach under consideration.

2.3. When is Monitoring Done?

The Bureau uses three types of monitoring: initial, systematic and/or special needs monitoring. The specific type of monitoring used is based on the interviewer's experience and/or current performance.

INITIAL monitoring is used during the first three interviewing sessions for newly-hired interviewers. It provides a transition between classroom training and actual production work for recurring surveys. An interviewer is not permitted to continue working on a survey if he/she does not successfully complete initial monitoring. In addition, all experienced interviewers who have already successfully completed the initial sessions must go through one initial monitoring session for each new survey that is assigned to the facility

SYSTEMATIC monitoring is used for those interviewers who have successfully completed initial monitoring. An interviewer will stay on systematic monitoring unless he/she receives a less than satisfactory score on a particular session and/or performs in an unsatisfactory way on one category on the monitoring form. Traditionally, systematic monitoring was used as a coaching and training tool. More recently it has been approved for use in performance evaluations.

SPECIAL NEEDS monitoring is usually conducted as a result of interviewer performance problems. A person on special needs is monitored more frequently, given comprehensive feedback, and may be re-trained if necessary. Special needs interviewers, like new interviewers, must achieve a minimum of three consecutive, satisfactory sessions in a survey period to be placed back on systematic monitoring. If the interviewer improves he/she goes back to systematic monitoring. If he/she does not improve after re-training and intense coaching he/she is removed from the staff or the survey.

2.4. What Is Monitored?

The monitoring form covers six basic categories. They are as follows:

1. MANNER AND VOICE...The coach/supervisor listens to the manner and voice of the interviewer to ensure that he/she is conducting the interview using a pleasant, professional voice and appropriate speech rate. The interviewer must also adapt to the speech rate of the respondent, be friendly, articulate, polite, and convincing.

2. READING SKILLS...Questions must be asked exactly as worded in survey instruments. This is a major requirement for all surveys to avoid bias through interviewer interpretation of specific questions. The coach/supervisor will watch carefully to see that questions are read as worded and ensure that questions are not omitted from the interview. In addition the coach/supervisor makes sure the interviewer uses the proper emphasis on specific words and verifies responses as required by the survey.

3. PROBING SKILLS...Interviewers must recognize unclear responses from the respondent and obtain clarification as appropriate. Many times this requires probing questions which must be carefully worded to avoid leading the respondent. The coach/supervisor listens to make sure probes are used appropriately.

4. RESPONSE ENTRIES...The coach/supervisor watches (on screen) to ensure that the interviewer properly classifies the respondent response to each question and enters the data appropriately. The interviewer should also be familiar enough with the instrument to maneuver back and forth as necessary and detect instrument problems. In addition the interviewer must always enter clear, concise notes in the information section of the instrument to prepare the next interviewer who may be assigned to finish or close out the case.

5. DIFFICULT SITUATIONS...Each interviewer must be ready to effectively deal with difficult respondents. The coach/supervisor listens for any difficult situations and ensures that they are handled properly. The interviewer should be able to convert reluctant respondents, handle respondents with language barriers, deal with hostile respondents, and remain calm and professional.

6. SURVEY CONCEPTS...Each interviewer should be able to introduce him/herself and the survey and be prepared to answer questions about the survey from the respondent. The coach/supervisor listens to the conversation to assess the interviewer's knowledge of the survey and ability and willingness to respond confidently with legitimate answers.

2.5 Monitoring Procedures

Until recently, while consistency existed with regard to specific skills to be monitored (Manner and Voice, Reading Skills, etc.), each telephone center made its own decisions concerning the administration and use of the monitoring results. Today all three centers are the same scoring system for monitoring operations. (S+, S, S- for initial monitoring and O, C, FS, M, and U for systematic and special needs monitoring). Each center also monitors at basically the same rate.

At any given time, the centers may have up to seven coaches/supervisors monitoring the interviewing staff. The monitor conducts audio and visual monitoring using a telephone headset and personal computer. At each monitoring station, a site display on the monitoring system highlights where the selected interviewer is sitting. The monitor then keys the appropriate phone extension to get the audio portion of the interview. The monitor is able to hear the interviewer and the respondent. In addition, he/she can see exactly what the interviewer sees on the screen. This enables the monitor to observe the interviewer's performance, the respondent's reaction, the path of the survey instrument, and the data keyed into the instrument The monitor can make specific entries to the on-screen monitoring form to record performance information for use in providing feedback. Multiple monitors can observe an individual interviewer simultaneously. Monitoring can be done both on-site, from monitoring "stations" in the supervisory area, and remotely from Census Headquarters in Suitland, MD.

Consistency with regard to monitoring procedures and subjective evaluations on each of the six categories is critical to success. Periodically, the coaches/supervisors get together and use a conference phone to monitor one individual. Each coach/supervisor rates the session and makes notations as usual (on paper). A meeting is then held to discuss each person's evaluation of the monitoring session. This discussion quickly identifies inconsistencies and enables the coaches/supervisors to come to consensus on how to resolve them. The intended result is consistent feedback to interviewers regardless of the person monitoring.

2.6 Feedback Procedures

Monitoring for purposes of quality and performance would be useless without feedback. The feedback must be very timely. Delaying the feedback would limit its utility, especially for periodic or one time surveys. Formal feedback is given to interviewers who have a better than average session or a less than satisfactory session. Feedback is given just after the monitoring session, where possible, and the discussion is held in a private office setting. Informal feedback is given for those monitoring sessions that do not significantly deviate from the average expected performance. For those on systematic monitoring, this feedback is often given at the interviewer's workstation since there is not a lot of information to give to the interviewer. However, those on initial monitoring are always given formal feedback since they are new to the job.

The supervisor/coach uses the "sandwich" approach when conducting a feedback session. This is accomplished by beginning the session with a compliment concerning what the interviewer did well, followed by specifics concerning what was done incorrectly, and ending once again on a positive note. This leaves the interviewer with some degree of self-esteem and hopefully gives him/her the desire to become a better interviewer. The positive feedback on both sides of the "sandwich" also helps the interviewer to accept help from the coach/supervisor concerning those things that were done improperly.

During the feedback session the supervisor/coach reviews the paper copy of the monitoring record with the interviewer. The coach/supervisor can identify specific cases and screens with detailed documentation concerning what the individual did well or poorly. Once the feedback is given both the supervisor/coach and the interviewer sign the form, and it is then entered into a database and the paper copy is filed. Copies are available to interviewers upon request.

3. ANALYSIS RESULTS

For this paper, we examined monitoring data for the Current Population Survey (CPS) from two of the three telephone facilities (Tucson and Jeffersonville) for the period November 1995 through May 1996.[2] CPS is the national survey of labor force characteristics. It is an ongoing monthly CATI/CAPI[3] survey which generally runs Sunday though Tuesday of the following week. Except for the month of December, CPS is always conducted during the week containing the 19th. At the CATI centers, CPS is usually closed out on Wednesday of the first week, with recycled cases sent back to the field. The major portion of CPS CATI interviews are completed during the first two days of the survey period. The average labor force interview is under 10 minutes. Monthly supplements vary in length and add to the average interview time. The average interviewer works 20 or fewer hours per month on CPS. The general monitoring guidelines for CPS are that each interviewer must be monitored one 30 minutes session during the survey period which amounts to somewhere between 2.5% and 5.0% of each interviewers login hours.

---

[2]Between November 1, 1995 and January 30, 1996, the U.S. government experienced several furloughs of workers. These furloughs affected the amount of interviewing and monitoring conducted during this period.

[3]CAPI is the acronym for computer assisted personal interviewing. At the Bureau, CAPI interviewing is conducted either at the respondent's home or by telephone from the field representative's home using a laptop computer.

3.1 Frequency of Monitoring

Table 3.1 displays data on the relative frequency of the three types of monitoring sessions by month for CPS at the Tucson and Jeffersonville Telephone Centers (TTC and JTC, respectively). Chi Square tests reveal there is a statistically significant correlation between facility and overall frequency of type of monitoring session. Clearly, for the period November 1995 through May 1996, month also is correlated with the proportion of different types of monitoring sessions conducted for CPS in both Tucson and Jeffersonville A much larger proportion of the CPS monitoring sessions in Jeffersonville during the period under study were initial CPS monitoring sessions. This was due to the fact that Jeffersonville trained 50 new interviewers on CPS in February in preparation for the March CPS. These 50 interviewers went through initial monitoring in February and March 1996.

TABLE 3.1. FREQUENCY OF MONITORING SESSIONS BY TYPE OF SESSION

	TTC			JTC		
	% I	% S	% SN	% I	% S	% SN
November 1995	73.1	26.9	0.0	35.7	64.3	0.0
December 1995	28.6	68.3	3.2	40.0	60.0	0.0
January 1996	22.5	74.2	3.4	21.4	78.6	0.0
February 1996	39.8	59.3	0.9	54.8	45.2	0.0
March 1996	17.5	79.0	3.5	51.1	48.9	0.0
April 1996	5.3	91.7	3.0	42.1	57.9	0.0
May 1996	6.3	91.8	1.9	2.9	97.1	0.0

3.2. Monitoring Outcomes

Tables 3.2a--3.2f show the proportion of interviewers receiving various initial monitoring ratings by facility. S+ is excellent, S is good or very good and S- is needs improvement There was no statistically significant[4] difference in the percentage of interviewers rated satisfactory or unsatisfactory in any initial monitoring category (e.g., manner/voice, reading skills) within facilities or across facilities. In Tucson, initial CPS monitoring session results indicate the largest percentage of new interviewers need improvement in reading questions exactly as worded, probing skills and survey concepts. In Jeffersonville, initial monitoring results indicate that the largest percentage of new interviewers need to work on survey concepts, response entries, and probing skills. This suggests that more emphasis should be place on probing skills, survey concepts, and the importance of reading questions exactly as worded during all initial training sessions in both facilities.

[4]Statistical tests for significance were done at the .05 level.

TABLE 3.2a. MANNER/VOICE

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=129	JTC n=80
S+	3.9	1.3
S	93.8	93.8
S-	2.3	5.0

TABLE 3.2b. READING SKILLS

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=127	JTC n=80
S+	6.3	2.5
S	75.6	82.5
S-	18.1	15.0

TABLE 3.2c. PROBING SKILLS

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=79	JTC n=74
S+	2.5	0.0
S	82.3	86.5
S-	15.2	13.5

TABLE 3.2d. RESPONSE ENTRIES

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=135	JTC n=80
S+	4.4	0.0
S	85.2	83.8
S-	10.4	16.3

TABLE 3.2e. DIFFICULT SITUATIONS

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=12	JTC n=28
S+	33.3	0.0
S	66.7	92.9
S-	0.0	7.1

TABLE 3.2f. SURVEY CONCEPTS

INITIAL RATING	% INTERVIEWERS RECVING RATING	
	TTC n=98	JTC n=73
S+	3.1	0.0
S	84.7	79.5
S-	12.2	20.5

Tables 3.3a--3.3f show the proportion of interviewers receiving various levels of systematic ratings in each of the CATI centers. O is outstanding, C is commendable, FS is fully successful, M is marginal and U is unsatisfactory. There was a statistically significant correlation between ratings on probing skills, response entries and survey concepts and facility. These data suggest that reading questions exactly as worded is the largest problem area encountered during systematic monitoring in Tucson. In Jeffersonville, the largest problem areas are response entries, probing skills and reading questions as worded. As mentioned earlier, Jeffersonville trained over 50 new interviewers during this period and keying in responses appears to be such a big problem (11.7% marginal or unsatisfactory ratings) because new interviewers take a while to get accustomed to keying on the computer

Table 3.3a. MANNER/VOICE

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=560	JTC n=137
O	3.0	3.6
C	13.0	6.6
FS	82.1	88.3
M	1.1	1.5
U	0.7	0.0

TABLE 3.3b. READING SKILLS

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=541	JTC n=135
O	2.6	1.5
C	11.8	9.6
FS	79.9	83.7
M	2.8	5.2
U	3.0	0.0

TABLE 3.3c. PROBING SKILLS

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=408	JTC n=126
O	1.7	0.0
C	10.3	15.1
FS	86.0	78.6
M	1.2	5.6
U	0.7	0.8

TABLE 3.3d. RESPONSE ENTRIES

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=569	JTC n=137
O	1.2	0.0
C	8.3	5.8
FS	88.4	82.5
M	1.1	8.8
U	1.1	2.9

TABLE 3.3e. DIFFICULT SITUATIONS

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=58	JTC n=35
O	25.9	8.6
C	34.5	28.6
FS	39.7	60.0
M	0.0	2.9
U	0.0	0.0

TABLE 3.3f. SURVEY CONCEPTS

SYSTEMATIC RATING	% INTERV'WERS RECVNG RATING	
	TTC n=504	JTC n=122
O	6.2	0.8
C	5.0	13.1
FS	85.9	84.4
M	2.2	1.6
U	0.8	0.0

Table 3.4 shows the mean monitoring session ratings (and standard deviations) by monitoring category. Using the SAS Proc TTEST to test for differences in means by telephone center, the results suggest differences for selected monitoring categories. For initial monitoring sessions, there were statistically significant differences in mean ratings between facilities for difficult situations. For systematic monitoring sessions, there was a statistically significant difference in ratings on reading skills, response entries, and session averages. The data in this table also show that after going through initial monitoring and receiving feedback, by the time interviewers get to systematic monitoring they are doing much better, suggesting the overall effectiveness of monitoring in both centers.

TABLE 3.4. MEAN MONITORING RATINGS BY SURVEY

CATEGORY	INITIAL		SYSTEMATIC	
	TTC	JTC	TTC	JTC
MANNER/VOICE	2.02 (0.25)	1.96 (0.25)	3.17 (0.51)	3.12 (0.46)
READING SKILLS	1.88 (0.48)	1.88 (0.40)	3.08 (0.60)	3.07 (0.45)
PROBING SKILLS	1.87 (0.40)	1.86 (0.34)	3.11 (0.45)	3.08 (0.48)
RESPONSE ENTRIES	1.94 (0.38)	1.84 (0.37)	3.08 (0.42)	2.91 (0.51)
DIFFICULT SITUATIONS	2.33 (0.49)	1.93 (0.26)	3.86 (0.80)	3.43 (0.70)
SURVEY CONCEPTS	1.91 (0.38)	1.79 (0.41)	3.13 (0.58)	3.13 (0.41)
SESSION AVERAGE	---	---	3.13 (0.40)	3.07 (0.25)

3.3 Inter-Rater Monitoring Results

One constant concern regarding CATI monitoring is consistency in ratings between monitors. Because of the use of monitoring results for performance evaluation, inter-rater reliability takes on heightened meaning at the Bureau. Tables 3.6a and 3.6b show the average systematic ratings given by monitors over the time period under study in the two telephone centers. Using SAS Proc GLM, the results indicate that in Jeffersonville, there was a significant difference in ratings for probing skills, response entries and survey concepts. Overall, the statistical results suggest more monitor variance in ratings in Tucson than in Jeffersonville.

TABLE 3.6A. AVERAGE SYSTEMATIC RATINGS BY MONITOR BY CATEGORY- TTC

MON-ITOR	MANNER/ VOICE	READING SKILLS	PROBING SKILLS	RESPONSE ENTRIES	DIFFICULT SITUATION	SURVEY CONCEPT	OVER-ALL
W02	3.02	2.96	3.00	2.95	4.50	3.04	3.00
W09	2.96	2.96	3.00	2.96	3.67	2.96	2.99
W10	3.16	3.15	3.06	3.07	3.29	3.04	3.09
W12	2.94	2.88	2.86	2.97	5.00	2.90	2.93
W13	3.59	3.33	3.52	3.49	4.50	3.79	3.54
W14	3.10	3.19	2.96	2.95	3.57	2.90	3.02
W16	3.02	2.86	3.06	2.91	3.50	2.97	2.97
W20	3.12	3.17	3.47	3.06	4.33	3.29	3.22
W21	3.65	3.55	4.00	3.43	4.75	3.48	3.60
W24	3.00	3.00	3.00	3.00	----	3.00	3.00
W26	3.12	2.92	3.03	3.00	3.50	3.00	3.01
W27	3.03	2.90	3.03	3.00	4.20	3.08	3.02
W29	3.15	3.10	3.03	3.05	4.00	3.07	3.08

TABLE 3.6B. AVERAGE SYSTEMATIC RATINGS BY MONITOR BY CATEGORY- JTC

MON-ITOR	MANNER/VOICE	READING SKILLS	PROBING SKILLS	RESPONSE ENTRIES	DIFFICULT SITUATION	SURVEY CONCEPT	OVER-ALL
SW12	3.03	3.04	3.04	2.94	3.75	3.07	3.02
SW15	3.00	2.50	1.00	1.50	----	3.00	2.44
SW17	3.32	3.12	3.27	2.71	3.46	3.18	3.11
SW18	3.17	3.17	3.33	3.00	4.50	3.00	3.21
SW19	3.71	3.57	3.25	3.57	3.67	3.86	3.63
SW20	3.00	3.00	3.00	3.00	3.08	3.00	3.02
SW21	3.00	3.00	3.00	3.00	---	4.00	3.17

4. CONCLUSION

Monitoring in the Census Bureau's CATI centers is a valuable tool for survey management. It provides a mechanism to assess the performance of both the interviewing staff, and the survey instrument and procedures. As a performance feedback tool, monitoring allows the supervisor or coach to observe interviewing unobtrusively and to give immediate, specific feedback. It also provides an accurate picture of the behavior of the survey instrument in actual practice, and offers the opportunity to make improvements in present and future questionnaires.

Information that is learned through monitoring is used to modify interviewer training and it serves to reduce the nonsampling error in the survey. This paper has highlighted some differences across CATI centers in how monitoring is done and how the data are used. We suggest that additional research be done to examine these differences so that all centers can benefit from what is learned.

5. REFERENCES

Biemer, Paul (1996). "Evaluator Error in the Assessment of Interviewer Performance", Bureau of the Census, July 10.

Cannell, C.F. and Oksenberg, L. (1988). "Observation of Behavior in Telephone Interviews," in R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg (eds), TELEPHONE SURVEY METHODS, pp. 475-495, New York: Wiley.

Couper, Mick, Groves, Robert M. and Smith, Timothy K. (1990). "Developing Systematic Procedures for CATI Monitoring", Paper Presented at the International Conference on Measurement Errors, Tucson, AZ. November.

Couper, M.P., Holland, L. and Groves, R.M. (1992). "Developing Systematic Procedures for Monitoring in a Centralized Telephone Facility," JOURNAL OF OFFICIAL STATISTICS, Vol, 8, No. 1, 63-76.

SESSION 5

ISSUES IN NEW COLLECTION AND CAPTURE METHODS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

STATISTICS CANADA'S EXPERIENCES WITH AUTOMATED DATA ENTRY

Suzanne M. Vézina[1]

ABSTRACT

In recent years, Statistics Canada has carried out research projects in the field of automated data entry. Approximately two years were spent on research into the technology, and the past year has been devoted to developing and implementing automated data entry systems for various surveys conducted by the Agency.

The purpose of this paper is to share some of the results obtained. These results highlight the importance of well-designed forms and questionnaires in ensuring that the data are of high quality. The form is the essential element of any data entry system. A well-designed form facilitates data collection and capture and is easy to fill out. The quality of automated data entry is optimized when the areas reserved for responses are clearly defined. Since the automated procedure uses an image of the form as the capturing agent, the paper and ink used must produce a high-quality image.

When forms are designed to facilitate automated collection and entry, fewer errors are introduced during these two processes. Automated capture serves to speed up the process while producing data that meet the expected standards of quality.

KEY WORDS: Automated data entry; Forms design; Image.

1. INTRODUCTION

In the "traditional" data entry process, responses are entirely keyed in by an operator reading from a form. This manual process is long, repetitive and monotonous. It therefore lends itself to the introduction of errors. Quality control procedures serve to reduce and almost eliminate keystroke errors. Quality control is an additional component related to data entry. Also associated with this component are the time and cost elements.

The goal is to minimize errors introduced into the data during collection and capture. Automated data entry from forms designed according to pre-determined standards is a promising approach.

2. AUTOMATED DATA ENTRY

In automated data entry the interpretation of the data, and hence the storage of the data on computer, is done mostly by computer and is then completed or corrected by an operator. The entire process is based on images of forms.

[1]Suzanne M. Vézina, Statistics Canada, Ottawa, Ont, K1A 0T6.

The functions of an automated data entry system are as follows:

- Scanning or import — The insertion of forms into an optical scanner that reads (scans) and registers the image in a computer-usable format. Existing images can also be imported into an automated data entry system.
- Form identification — Automatic identification of forms (pages) subjected to automated capture.
- Exception — Automatic detection of pages that are not part of the form and pages that have not been properly identified by the scanner.
- Extraction — Automatic recognition of machine-printed characters, hand-printed characters, bar codes and marks.
- Verification — Manual verification and correction of suspect data and data that do not satisfy the validation rules. These data, along with the image of the form, are displayed on operators' workstations. The operators complete the data using these images.

3. FACTORS THAT FACILITATE DATA COLLECTION AND CAPTURE

The form is the key component of any data capture system. If a form is designed to take account of the factors that affect collection and hence capture, fewer errors are introduced into the data during collection and also during the subsequent capture of the data. This is true regardless of whether the capture is manual or automated.

Manual data capture is done by a human being, and the latter tries to decipher the data to be captured when the responses are hard to read. The responses may simply be hard to read or they may be entered wholly or partially outside the areas or zones reserved for responses; they may even go against what was expected as a response to a question. Utilizing the human eye's ability to adapt, data entry clerks perform their work as best they can.

If a form is designed according to strict parameters and standards, fewer errors are introduced during data capture. Under these circumstances, automated capture is therefore a practical alternative. Designing forms so as to facilitate collection and hence automated capture is key to the success of the automated capture process. Since the capture is mostly done by machine, the process is carried out at high speed and the results are reliable. The possibility of defining confidence levels for the results of the recognition helps to reduce the errors introduced during capture. In an automated environment, the operators do not intervene unless the machine queries the recognition results, or unless the results go against the validation rules defined in the application. Thus we have seen a number of factors that reduce the number of errors introduced during the collection and capture of data: forms designed according to standards that facilitate collection; good images; a recognized and effective data capture system; and validation rules applied to the data extracted.

Factors related to questionnaire design that facilitate collection and hence automated data entry are paper characteristics, the colours of ink used to print the form, page identifiers, registration points, the definition of the zones containing the data to be captured, and finally the printing of the forms.

4. PAPER CHARACTERISTICS

All scanners, scanning software and automated data entry software impose restrictions on the paper used. Texture, fibre content, surface quality, thickness or weight, reflectivity, opacity, colour and size are all important factors.

- Texture, fibre content and surface quality affect the sharpness and precision of the image produced.

- Thickness or weight have a direct impact on the functioning of the scanner. Continuous, smooth and uninterrupted scanning is desirable. If the paper is too thick or too thin, the scanning will be constantly interrupted and may even be impossible. When the paper is too thick, the scanner cannot accept it. When the paper is too thin but the scanner manages to read it, the images are often blurred because of the waviness created when the form is fed into the scanner.

- Non-glossy paper should be used. Glossy paper causes reflection during scanning, and this blurs parts of the image.

- The paper used should be opaque, so that what appears on one side of a page will not show through to the image on the other side.

- The paper should be white or a pastel colour, with enough reflectivity for the images to appear blank when viewed on screen.

Plain paper made from non-recycled fibres as well as most papers made from recycled fibres may normally be used for imaging.

5. INK COLOURS

The use of colours and areas without colour enhances the appearance of forms and consequently the appearance of the images taken of these forms. Colours serve to guide respondents, thus reducing ambiguities. This means that fewer errors are introduced during the collection process. The colours of ink used on forms may be either scannable or non-scannable; in the latter case they are referred to as "drop out" colours. Automated data entry is facilitated by the use of drop out ink colours around response areas.

- Drop out ink colours are in subtle shades. They are used to guide the respondent through the form but are invisible to the scanner. They are not perceived by the scanner because they are highly reflective. These colours are appropriate for use in framing the response zones and for any other non-essential information such as backgrounds and graphics.

  The use of drop out ink colours in the appropriate places improves data recognition, reduces the size of the images (thereby cutting down on storage space requirements), speeds up the imaging process and produces images that are clearer when viewed on screen.

- Scannable ink colours are in dark shades. They are used for the following:
  - to print page identifiers and registration points;
  - for any information that must be extracted by the recognition system;
  - for any information that must be visible when the image is viewed on the screen, such as the questions.

Black is the most commonly used colour for printing information that must be discerned when the forms are scanned. Information printed with scannable ink colours should not obstruct the response zones.

It is also necessary to define areas without colour when designing a form. Extraction zones and the area immediately surrounding them should not be obstructed or blurred by partially or totally scannable colours.

Another point to consider is the form's fold area. When a form is folded, the image produced through scanning

will be greyish in the fold area. Therefore when designing the form, it is important not to locate response zones in this part of the form.

6. PAGE IDENTIFIERS

Page identifiers permit automatic recognition of forms subjected to the automated data entry system. The system examines the image and determines which form and which page within the form are to be processed. Page identifiers also make it possible to detect pages that are not part of the form being processed, when such pages have been introduced inadvertently.

A page identifier should be unique within a multi-page form. A page identifier should be defined for each side of every page of a form. Page identifiers should be printed with an ink colour that is discernable by the scanner. Examples of page identifiers are: EERH, E1, E2.

7. REGISTRATION POINTS

Registration points enable the system to properly align the images and eliminate skew during scanning. Alignment of the images is essential in order to obtain good results in data extraction or recognition.

Registration points should be determined on each side of each page of a form. They should be printed with an ink colour that is discernable by the scanner. Registration points should be printed at various places on the page. The automated data capture system used determines the type and possible locations of registration points on a page. Examples of registration points used by various automated data entry systems are small black triangles or black circles printed in each corner of the page.

8. DEFINITION OF ZONES OF DATA TO BE CAPTURED

A zone reserved for data is designed according to the type of data that the survey seeks to obtain. The table below lists the types of data that can be recognized by an automated data capture system. It also specifies, in English and in French, the types of technology used to recognize each.

8.1 Bar codes

Bar codes correspond to alphanumeric data, and they use a series of vertical lines of varying width, with spaces between them:

Following the table are several concepts used in designing a form according to the type of data sought.

TYPES OF DATA	TECHNOLOGY - English	TECHNOLOGY - French
Bar codes	BCR Bar Code Recognition	RCB Reconnaissance de Codes à Barres
Characters printed with magnetic ink characters	MICR Magnetic Ink Recognition	RCEM Reconnaissance de Caractères à Encre Magnétique
Marks	OMR Optical Mark Recognition	ROM Reconnaissance Optique de Marques
Machine-printed characters	OCR Optical Character Recognition	ROC Reconnaissance Optique de Caractères
Hand-printed characters	ICR Intelligent Character Recognition	RIC Reconnaissance Intelligente de Caractères

Bar codes should be printed with a scannable ink colour. The area surrounding the bar code should be blank; in other words, no information printed with scannable ink should obstruct the code itself.

8.2 Characters printed with magnetic ink characters

These characters are formed with a highly specialized character set, and they are printed with magnetic ink:

This type of characters should be printed with a colour of ink that is discernable by the scanner. The characters should be surrounded by a blank space, so that no information printed with scannable ink obstructs the characters themselves.

8.3 Marks

Mark-sensed zones or fields are delimited areas in which the respondent indicates a choice by making a mark:

Mark-sensed zones must be delimited by a box, circle or ellipse. Depending which automated data capture system is used, they will be printed with either a scannable ink colour or a drop out ink colour. The ideal size for a mark-sensed zone is 1/8 inch square. It may also be larger. It is important to locate the zones sufficiently far apart that a choice indicated by the respondent does not extend into another response zone.

8.4 Machine-printed characters

Machine-printed characters include all standard character sets as well as typographic characters and characters generated by laser printers and typewriters:

There should be a blank area measuring at least 1/4 inch all around the characters, so that no other information printed with scannable ink obstructs the characters themselves.

Spacing between characters, spacing between rows of characters and the density of the characters all affect recognition performance. If the characters are touching each other, or are not far enough away from any other writing or diagram, recognition will be difficult.

8.5 Hand-printed characters

As the name indicates, hand-printed characters are characters printed by a person.

Hand-printed characters are the least conducive to automatic recognition, since they are quite varied and inconsistent in terms of style and form. Hand-printed characters are sharper and therefore easier to read when the zones for hand-printed responses are constrained. The clearer the boxes on a form, the more respondents are spontaneously inclined to print the information in the right place and to do so clearly and consistently. The best way to receive hand-printed data is to provide a set of boxes. Boxes encourage respondents to print the characters in capital letters instead of writing them in manuscript or cursive style.

- Provide enough boxes to receive the maximum number of characters for a response.

- The minimum box size to enable the respondent to properly form the characters is 3/16 inch wide and 1/4 inch high.

- The boxes may be contiguous (semi-constrained), that is, without space between them. In this case, a single median line separates two boxes and is therefore shared by them.

- The boxes may be juxtaposed (fully constrained), that is, close together but slightly separated from each other.

- When there are several rows of boxes, there should be a minimum of 1/8 inch separating the rows..

- There should be no information obstructing the space reserved for characters

9. PRINTING OF FORMS

Since the image is the medium by which automated data entry is carried out, the print quality of the image directly affects the performance of the system.

The precise positioning of page identifiers, registration points and response zones, *on all copies of a given form*, are key factors in achieving effective and ideally error-free capture. The more clearly and precisely delimited the zones are on all copies of a form, the better the capture will be.

Lines and boxes that surround characters to be extracted that are printed with scannable ink colours should be continuous and of consistent thickness. A number of automated capture software programs use a technique that consists in removing the boxes before extracting the characters. When the boxes are removed, they are less likely to interfere with the characters. When the lines and boxes are not printed continuously and with consistent thickness, these software programs do not recognize them as effectively. They therefore do not remove them, since if they did, they would risk also removing characters to be extracted. Such situations do not arise when the lines and boxes are printed with drop out ink colours.

10. CONCLUSION

Automated data entry is a viable technology that can be used with positive results. The design of forms according to the standards cited in the article is an essential and integral part of the operation of such a system. Motivated respondents write more clearly, with greater regularity and consistency. As a result, fewer errors are introduced during the collection process. The automation of data entry becomes possible with these forms, and this automation results in fewer errors than when data entry is done entirely manually.

The automated approach offers several advantages: capture time is shorter, data quality is improved, and the cost of the operation is lower than if the capture had been done manually:

- Computerized data entry greatly reduces the time required for the process. Computer speeds are constantly increasing. It is increasingly advantageous to use them for data processing, including data entry.

- The fact that data entry clerks intervene only when the system cannot extract the data with certainty also helps reduce the time required to complete the process. With fewer manual interventions, fewer errors are introduced than when the process is entirely manual.

- When data are captured automatically from a form designed in accordance with the standards cited in this article and when validation rules are applied, the data contain less than 1% errors. This percentage is based on the assumption that data entry clerks do not introduce errors during the verification process. But since verification is a manual process, the data entry clerks sometimes make mistakes. The validations carried out on extracted data serve to minimize the number of errors introduced during the verification process. Quality

control techniques are also included in various automated data entry software programs. These techniques are used to detect and directly correct data capture errors.

- Time saved is also money saved. The initial investment in software and computer equipment is recovered as the system is used. Depending on which automated data entry system is used, at least 40% of the data entry clerks can be assigned to other tasks. This figure may even be as high as 60% when advanced automated data entry systems are used.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

NONSAMPLING ERROR, CAN ELECTRONIC REPORTING HELP?

Laurie Hill[1]

ABSTRACT

Electronic data reporting is still relatively new in Statistics Canada, but early results from a number of trials suggest that for business surveys, these techniques may be able to improve timeliness, reduce costs, and improve the quality of the data by reducing nonsampling errors.

KEY WORDS: Non-response; Error; Electronic; Reporting.

1. INTRODUCTION

1.1 What is electronic data reporting?

For years, the predominant form of data collection used by Statistics Canada for business surveys was paper questionnaires, with mailout – mailback as the primary delivery mechanism. More recently, we have introduced a number of alternative approaches, including telephone follow-up of mailed questionnaires, pure telephone collection with interviewers taking down the information on paper for subsequent data capture, direct interviewing by Statistics Canada staff, and computer-assisted telephone interviewing (CATI). We have also experimented with touch-tone data collection, computer-assisted personal interviewing, and other innovative techniques. We are now involved in what we normally call "mixed-mode" data collection. We continue to make use of a number of these techniques, mixing and matching to obtain the most effective combination of methods to collect the required data in an efficient and timely manner while minimizing the reporting burden on respondents.

Electronic data reporting is the newest of the techniques we are experimenting with in a continuing effort to reduce the perceived reporting burden on business respondents as we improve the speed and efficiency of our data collection efforts.

We have experimented with a number of methods to deliver questionnaires and retrieve data responses in electronic form. The most frequently used method to date to present the survey questions and to specify the form and content of the responses has been to mail out the survey on diskettes. Respondents use computers to complete electronic questionnaires and normally send the data back to Statistics Canada on a diskette as well. We are developing alternatives to offer respondents more choice in transmission methods, including print-and-FAX, electronic mail with file attachments, modem-to-modem transfer, and (coming sooon) secure Internet Mail transfer. We are also looking into the possibility of extracting information from respondents' administrative records in electronic form (with prior permission), or obtaining copies of electronic transactions from the "value-added networks" which are transporting those transactions between "trading partners".

Our objectives in all of these efforts are the same: to reduce the real or perceived burden on business respondents and at the same time obtain better quality responses to our surveys. We are offering electronic reporting as another option for respondents; we are not insisting on it.

---

[1] Laurie Hill, Assistant Director, Enterprise Survey Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

1.2 Common features of these techniques

- They are all designed to make the respondent's job easier.
- We specify the data we wish to receive in electronic form, **and** we specify the format in which we wish to receive the data.
- Edits are built into the applications, so they are applied as the respondent enters the data into the program, not when we receive the data in Statistics Canada.
- All of the EDR efforts to date have been survey-based, i.e. they are not generalized systems; they are designed for a single survey, and most of the design work must be repeated if we wish to apply the technique to a second survey.

1.3 The expected benefits to Statistics Canada

- Improved timeliness, if we work at it!
- Lower cost surveys, if the applications are well-designed, **and** only after an introductory period.
- Improved data quality, if the correct edits are built into the application.

1.4 The results

On the questions of **timeliness**, the jury is still out. There is no question that some of the reports have come back faster in electronic form than they did when the same respondents filled in paper questionnaires. However, there are also cases where respondents have taken longer to complete the electronic return than they did with paper. Generally we can say that follow-up is still required when we use electronic reporting techniques. During the introductory period, a "help line" is essential, because no matter how complete you think the instructions are, respondents are sure to run into problems. They will need lots of hand-holding, and if they don't have a "friend to turn to," the chances are they will give up. If you want electronic reporting to work, you need to be prepared to invest time and effort not only in building a good application, but also in helping the respondents to get it working in their environment, and to become familiar with it. The **payoff** in terms of improved timeliness should come with reprtition.

When it comes to **cost**, we **know** there should be savings! After all, the data capture is done by the respondents, so there whould be measurable savings in that function at least. For most of the "good" electronic reporting applications, the fact that the respondent performs the edits on the data should also lead to lower costs. This will be true **if** the application is well-designed, **if** edits are incorporated in the application, and **if** the respondents understand what we want from them. In short, there is very little doubt that electronic data reporting will lead to reduced costs, but only after an investment in proper application design and well-planned introduction of the new methods.

However, the real question for this symposium is: "Do electronic data reporting applications lead to improved data quality? If they do, the improvements will undoubtedly come as reductions to non-sampling error. There are no differences in sampling when we use EDR techniques, nor are there any significant frame differences. The improvements, if any, will come from removal of interviewer-induced error, and a reduction in respondent interpretation error if the application design meets the conditions outlined above.

It is too soon to say with certainty that data quality has been improved through the use of electronic data reporting techniques. The applications used to date have not been designed with evaluation in mind, so it is difficult to draw firm conclusions. We would like to have comparable statistics on edit failures in records from samples of respondents before and after EDR techniques were introduced. Unfortunately, in the applications introduced so far, no such records were kept. However, results from the PERQS (Personalized

Electronic Reporting Questionnaire System) application in the Annual Retail Chain and Department Store Survey are promising. EDR techniques were tested in this survey with a sample of 25 large respondents during the 1994 survey, and the application was extended to 280 responding firms in the 1995 survey. These 280 respondents account for more than 50% of the retail locations covered by this survey. The supervisors involved in our Operations and Integration Division are certain that quality has improved. They say **the proportion of clean records is up** from 45% using earlier techniques (a mixture of mailout-mailback and CATI collection) to roughly 75% using EDR diskette collection. Further, they suggest that many of the remaining 25% of the EDR records which fail edits are caused by respondents legitimately over-riding the edits in the EDR application, but not providing a reason or explanation. Comments like these are comforting, but it would be more reassuring to have some solid statistical evidence that quality has definitely improved. Methods to assess the quality change should definitely be a part of the next EDR test application.

1.5 Future directions

Statistics Canada has recently decided to initiate a new respondent relations program for large enterprises. These large businesses are included in many of our surveys, and their operations are so large that it is essential that they respond, and that they respond correctly. These large businesses are looked upon as "key data providers", and Statistics Canada is setting up a Key Provider Management program to deal with them on a more concentrated basis. Each of the selected enterprises will be assigned a Key Provider Manager (KPM), a primary contact in Statistics Canada, who will focus his/her attention on a small number of such important respondents. The KPM will be responsible for providing the enterprise with a consolidated view of our data requirements, for coordinating data collection across all business surveys collected by the agency, and for negotiating data reporting arrangements with the enterprise to reduce response burden. One of the techniques KPM's are expected to promote in an effort to reduce the reporting burden is Electronic Data Reporting. The KPM's will also be responsible for coordinating data confrontation with Statistics Canada for all reports received from the enterprises for which they are responsible. In this way, we expect to improve the consistency, coverage and quality of the responses we receive from these large enterprises.

For small and medium sized enterprises, we have no similar program at this time. However, some ideas have been suggested which, if pursued, could lead to improvements. We could begin working to develop a more consolidated set of data requirements on a monthly, quarterly and annual basis. We could then relate these requirements as closely as possible to data normally available from accounting records. We might then work closely with the relatively small number of companies which provide the accounting software used by many small and medium enterprises in Canada. We should be able to help them create one or more statistical reporting modules which could pull information from the account records of the business, package the information in an electronic file, and ship it off to Statistics Canada, via toll-free telephone lines, on a regular basis. It is possible that some of these ideas will be put to the test in the years ahead.

Proceedings of Statistics Canada Symposium 1996
Nonsampling errors
November 1996

THE WORLD WIDE WEB AS A DATA COLLECTION METHODOLOGY: DESIGNING THE SURVEY OPERATIONS OF THE FUTURE

Richard L. Clayton [1]

ABSTRACT

The Internet has become the symbol of the Information Society. This technology tool for accessing and sharing information is now reaching throughout our culture. The Internet and the Universal Resource Locator (URL) of the World Wide Web (WWW) are linked to virtually every other type of communications whether newspaper, magazines, TV or radio. This technology promises to provide people with any information needed in any format, the ability to answer any question, to replace much of the print media. The information superhighway has now opened up many new dynamic areas of electronic collection research for establishment surveys. The results from this research will ultimately position establishment surveys as the most timely, accurate, and cost effective of the survey operating environments.

KEY WORDS: Internet; Data Collection; Technology; E-mail; World Wide Web.

1. INTRODUCTION

The promise of widespread access and instant access via Internet and World Wide Web (WWW) have already spurred research into the potential for improving surveys. This paper reviews current Internet/WWW features relevant to data collection, reviews some existing literature, identifies considerations in the development of a Web data collection system, and profiles the strengths and weaknesses of the E-mail/Internet against other Computer Assisted Survey Information Collection (CASIC) methods in terms of quality, timeliness and costs, and discusses issues relating to its future use for surveys including confidentiality. Also described are the early research results from an integrated pilot E-mail/WWW methodology at the Bureau of Labor Statistics (BLS) in the Current Employment Statistics (CES) program.

1.1 The Internet and World Wide Web

The Internet is a global network of computers linked by a standard communications protocol. It was originally introduced by the U.S. Department of Defense in 1969. The Defense Advanced Research Project Agency provided funds for the development of ARPANET, whose intent was to have a highly resilient network for military communications. The WWW is a graphical interface to the Internet. The WWW was developed at CERN, the European physics laboratory. The first version of Web software was released in 1991. The popularity of the Web dramatically increased in 1993 with the release of Mosaic. Now, the WWW is the most popular and promising Internet application.

[1] Richard L. Clayton, U.S. Bureau of Labor Statistics, Room 4860, 2 Massachusetts Avenue, N.E., Washington, D.C. 20212.

1.2 Evolution of CASIC Methods

Over the last two decades, there has been a rapid development of alternative automated collection methods beginning with CATI in the 1970s, then with the availability of the microcomputer came CAPI, TDE and VR. Each of these methods required little of the respondent except a touchtone telephone.

In the late 1980s came Computerized Self Administered Questionnaires (CSAQ). Also known as CASI (Computer Assisted Self-Interviewing), this method was the first to take advantage of the growing access of advanced microcomputers in the hands of our respondents. Using CSAQ, respondents load the provided software on their PCs, and use the system for entering and editing their own data. Thus, CSAQ methods are much like CATI except the software acts as the interviewer. Like in CATI, CSAQ includes branching and on-line editing (OMB, 1990).

Much of the drive for CASIC has been to improve the quality of data collected and edited at the source, while controlling interviewer error through computer-driven branching (Nicholls, Baker and Martin, 1995). These methods have offered improvements in data quality, improved timeliness and/or reduced costs. As the World Wide Web provides an inexpensive, easy-to-use communications framework, and building on widespread availability of high-powered desktop computers, WWW reporting is the next logical step in CASIC evolution.

1.3 Advantages of WWW for Data Collection

Web Versatility: Unlike the telephone collection methods of the 1980s, WWW collection can accommodate a wide range of surveys and survey operations. The use of telephone collection procedures were often limited by the length and complexity of the questionnaire, the frequency of the collection cycle, and the range of survey operations for which the telephone could be used in a cost effective manner.

Questionnaire Length: CATI is generally limited to surveys which could be conducted within a few minute session and is problematic if respondents needed to refer to their records. TDE and VR are correspondingly limited to the number of items for which a respondent was willing to push buttons and for the number of questions a respondent was willing to answer to a machine. WWW, however, has the ability to accommodate structured questionnaires of any form or length including "form-layout" designs or traditional "question-by-question" designs. The respondent has the ability to refer to records as frequently as needed or to partially complete the questionnaire and return to it at a later time.

Survey Frequency: Ongoing WWW surveys are easy to maintain if a file of contact information, including the E-mail address is used. One-time multi-mode surveys are more difficult to implement as a complete file of E-mail, mail, and phone addresses may be difficult to obtain.

Altering Questionnaire Content: The WWW has the same broad flexibility as traditional mail in easily accommodating content changes (e.g., adding new data items) or conducting periodic survey supplements. Under the WWW environment, respondents access the server and use the software residing there. Thus, the system can be modified and loaded at a single point and maintained. Once loaded, then, all respondents have immediate access to the modified software.

Product and Customer Service Improvements: WWW systems can provide any array of survey products or other facilitation features for respondents. Surveys have worked diligently to associate the utility and importance of the published data with the reporting of microdata. These efforts have taken the form of booklets, brochures, or press releases. The WWW interface will take this effort to a new level by being able to profile the data provided by each respondent against, for example, their industry, in their state, and against the nation. Also, multi-media capabilities could provide on-line "clippings" showing

the data in use, whether from the print media, radio or TV, further cementing the utility and worth of their reporting efforts.

Accuracy: For our surveys, accuracy will also be improved in a number of ways. The microdata from employers is based increasingly on direct computer generated tabulations rather than secondary sources. The respondent can respond directly to all edit queries.

Timeliness: Our customers will benefit from more timely data. For some surveys this will mean "final" estimates will be quickly available and thus will eliminate the need for "preliminary" estimate surveys, or for others, a reduction in the size of revisions between preliminary and final estimates.

1.4 The Current Employment Statistics Program

The CES is a monthly survey of about 390,000 non-agricultural establishments. Collecting a small number of payroll related data items, the CES is well-suited for CASIC methods. Over the past 14 years, the CES has researched and used CATI, TDE, Voice Recognition (VR), and Electronic Data Interchange (EDI) before addressing the cost and error reduction properties of the WWW. A driving force behind the CES implementation of CASIC methods is the very short collection window, only 2 to 2.5 weeks of data collection for the preliminary estimates. Under this array of methods, the average response rates for delinquent reporters were raised by 30 percentage points and average monthly revisions to preliminary estimates have been reduced by 38 percent. At this writing, over 230,000 respondents report monthly using TDE with thousands more planned for transition to TDE.

2. WWW METHODOLOGY

2.1 Combining Two CASIC Methods in One

Developing a WWW survey methodology involves using tools developed under two other methods. These tools include the respondent contact procedures, and the automated self-interviewing techniques familiar from Computerized Self-Administered Questionnaires (CSAQ). The combination of these features comprises the likely direction of WWW collection methodology.

The WWW survey collection cycle, begins with a sample control file containing the respondent's E-mail address in addition to the normal respondent contact information of name, address, and phone number. The collection form is a standard "web page" containing an image of the questionnaire, survey instructions, definitions, and hypertext links to definitions. An E-mail address is provided for problem reporting and inquiries. As the collection cycle approaches, the respondent opens their E-mail to find a reminder, "surfs" the net to the CES homepage, accesses the data collection screen, and fills in the requested data. The moment the respondent clicks the "submit data" icon, the data are transferred to the survey agency. Schedules are electronically checked-in and, at predetermined time periods, E-mail nonresponse reminder messages are automatically sent.

Existing TDE methods largely eliminate labor-intensive activities for mail-out and mail-back and data entry. However, neither method yet directly addresses another expensive activity: data editing and reconciliation. Current labor-intensive edit and reconciliation operations can also be directly handled under WWW collection; allowing the respondent, as in CSAQ, to directly review perceived edit failures and correct them as necessary.

2.2 Respondent Contact Methods

Traditional mailed surveys relied on the arrival of the form to spur the respondent's reply. If a response was not received at certain times, nonresponse prompts were sent. The TDE methodology developed at BLS incorporates these same three message types in a carefully timed approach which can be mirrored in a WWW methodology. Under TDE, the essential steps are 1) an advance notice message via postcard or FAX which replaces the mailed form as a notice to take action, 2) the self-initiated call to a toll-free number for entering the numerical answers to the pre-recorded questions, and 3) a non-response prompt message by CATI or FAX (Werking and Clayton 1994). Under WWW, the advance notices and nonresponse prompt messages and timing are mirrored by E-mail messages, see Figure 1 below.

Under this view, the entire methodology can be automated, from the timing and content of the E-mail messages to firm-specific nonresponse prompts based on the known availability of needed records. With all data entered and edited on-line, a truly "peopleless- paperless" methodology is possible (Werking 1994).

2.3 The CSAQ Interviewing Method

Computerized Self-Administered Questionnaires (CSAQ), also known as Computer Assisted Self-Interviewing (CASI), essentially places the CATI or CAPI instrument in the hands of the respondent reading questions and providing the answers themselves. The essential difference versus CATI/CAPI is that the CSAQ instrument must incorporate all of the knowledge and training of the interviewer, including probes. Of course, the most powerful tools, automated branching and on-line edits, are essential for making this technique more than remote data entry. CSAQ involves two key features, the questionnaire instrument and the transmission vehicle such as mailing diskettes, printouts or sending E-mail files. The particular CSAQ issues which are translated directly to WWW collection are screen design and navigation among screens, or "pages".

Figure 1: Comparison of Respondent Message Types : TDE versus WWW

Methodology Feature	TDE	World Wide Web	Fully Developed, Integrated System
Monthly Advance Notice	Postcard or FAX	E-Mail	Automatically generated
Data Reporting	Call 800 number to TDE system	https://www.ces.bls.gov	Incoming data are instantly received, edited, and stored
Nonresponse prompting	Phone call, Fax, or Postcard	E-Mail	Automatically generated according to monthly automated calendar.

WWW methods can be viewed as very similar and an extension of CSAQ, with the added benefits of instant receipt of data, central updating of the instrument and the elimination of mailing or lengthy downloading of large instruments. To achieve these benefits, WWW design must be designed for rapid and multiple downloads of individual screens.

2.4 WWW Design Issues

The WWW poses some difficult issues for designers. Maintenance of security, accounting for transmissions and interface design are key issues. Federal surveys are required to maintain strict confidentiality about participants and their data, requiring encryption technologies, discussed in a later section.

There are several issues regarding the design of the system as it appears to the respondent, including the approach used for screens, the design of each screen and navigation within and among screens. Researchers in the basic GUI environment include Ben Shneiderman at the University of Maryland's Human Computer Interaction Laboratory and more specifically for WWW site design, Jakob Nielsen at Sun Microsystems.

The Shneiderman approach focuses on building systems which serve user needs. The approach is that good design retains and satisfies users to the benefit of the system. Good interfaces should be simple and intuitive and provide some form of feedback or reward (Shneiderman, 1993).

Specifically for WWW development, Jakob Nielsen also focuses on the intuitiveness of the design. In most cases, users should not have to read documentation. He also recommends "think aloud" sessions with users as an integral part of the development process (Nielsen, 1996)

Scrolling versus "pages" is the first design consideration. Many WWW sites use scrolling capability, but these are mostly text-based sites. Neilsen emphasized use of page-based sites a more intuitive and less confusing. Respondents may not realize that more questions are hidden, while page based systems should be designed to follow branching as is routine in CATI, CAPI and CSAQ. The tradeoff may be one of the number of downloads, and the time these may take, versus the clarity and control of a page-based design.

Providing the respondent with the ability to knowledgeably navigate within and among screens is key to user satisfaction. The screen design must make it clear to the respondent how to move about within the questionnaire. Commonly used buttons, labeled for their exact purpose, aid in this design. Also, familiar browser tools for "back" or "forward", or buttons with similar functionality, would allow the user to move within the instrument and should be integrated to make such movement easy. It is also possible to disable certain features to prevent respondents getting lost in trying to see all of the branches, that is to only open certain branches.

Page composition is also crucial. WWW survey researchers should be aware of their respondent's environment. Currently, standard modem connections are usually either 14.4 or 28.8 kb. Therefore, to minimize respondent burden as measured by the time needed to download the page, heavy use of graphics should be avoided. The problem of lengthy and numerous downloads is euphemistically called the "World Wide Wait".

2.5 The WWW - CSAQ Continuum

One essential difference between WWW and CSAQ is the location of the automated survey instrument. CSAQ places the instrument on the respondent's PC, whereas WWW in its pure sense relies on central access of the instrument. Variations of WWW design may place some features on the respondent's PC. For example, given the current difficulty of securely storing historical microdata outside of the firewall for use in longitudinal editing, we may store it on the respondent's PC. As other features shift to and from the respondent's PC or the central WWW server, at some point, the use of the respondent's PC becomes more CSAQ than WWW. The continuum between these two methods will be faced by each designer based on each survey's needs.

2.6 Total Design Method On-line

The eventual replacement of traditional methods with WWW will require a careful review of all mail-based research. The results serve as reasonable starting points for WWW methodology. Under TDE, for example, very high response rates have been attained using a combination of advance notices, easy to use data entry interfaces, and carefully-timed nonresponse prompts. Will WWW work the same? The Total Design Method (TDM) offers a rigorous approach to maximizing response rates (see Dillman 1978).

Under the TDM, each survey feature (prenotification message, the survey instrument, reminders and the timing of each) carries potential for improving response rates. Figure 2 illustrates how some of the features of the TDM can be implemented on the WWW. WWW design should be intuitive and accessible to respondents. A good design facilitates higher response rates.

Figure 2: Translating the Total Design Method to the WWW

Survey Function	TDM Recommendation	WWW Application
Survey Correspondence	Correspondence should be personalized	Advance notice and nonresponse prompt E-mail messages are addressed to a specific respondent. The questionnaire is always available on the Web site.
Postage	Use actual stamp on return envelope	"Registered" E-mail could be used for the advance notice and nonresponse prompts.
Survey Form	Simplify form: -white space -booklet approach -attach definitions	The screen can be designed to maximize white space. Definitions can be included as hypertext links.
Question Lists	Prioritize questions: Shorter, respondent friendly questionnaires get higher response rates that longer respondent friendly questionnaires.	Respondent friendly design is easily implemented on the WWW. The perceived length of the questionnaire can be minimized through the use of hypertext links.

2.7 Respondent Access to E-mail and the WWW

Employers have responded to increasing international competitiveness pressures by downsizing and flattening their organizations, increasing their productivity and controlling their wage and price structures. However, perhaps more importantly, employers responded by investing heavily in computing technology and communications during the 1980s to boost productivity, to link their national and international operations and to provide instantaneous access to critical management information on inventories, personnel, and cash-flow transactions. In 1991, U.S. companies for the first time spent more on computing and communications gear, than on industrial, mining, farm, and construction machines (Werking 1994). These investments should pay off in increasing availability of technology, and Internet access for some worker groups.

In a 1995 survey of 404 randomly selected Chief Information Officers (CIO) of Fortune 2000 companies (Spanski, 1996), several key indicators of the current and future potential for E-mail were outlined. Most significantly, 89% of CIOs had E-mail within their companies. About half of the remainder expect E-mail access within the next two years. However, about 60% of their employees had access. Less than half, 44%, had a link to the Internet. Most of these large businesses initially establish E-mail linkages to improve internal communications and internal decision making. Importantly, those with Internet access point to it as a means for further improving decision making, indicating that those without Internet access are likely to follow.

The number of US households with Internet access is estimated at 14.7 million, having doubled in just one year (Wall Street Journal, 1996). Daily access is estimated at 9 million adults. A 1996 Nielsen poll estimated that 30 percent of all working adults have Internet access (Hoffman, Kalsbeek and Novak, 1996). Using the WWW for messages to households may face some difficulties. For example, we tend to think that people will retain Internet service once in place. This has been largely true for telephones and FAX machines. However, there is some evidence that people drop and add Internet service over time, and that they will shift among providers as prices and services differ, referred to as the "churn rate" in the

industry, changing their E-mail address. Perhaps the trend in the telephone industry of a single number covering a person at home, office and portable will also take over in Cyberspace. For establishments, the churn rate may tend to be lower, but keeping track of individual respondent's E-mail addresses will pose the same difficulties as telephone numbers when respondent turnover takes place. A hopeful note is that there have been no changes in E-mail addresses for the 50 respondents now reporting to the CES using WWW, see below.

These statistics are hopeful depending on the target respondent for a particular survey. Questionnaires targeted for CIOs may consider the WWW collection as the primary vehicle. However, these signs may be less promising for most establishment respondents. In a 1995 unscientific review of 1300 respondents to the CES survey, typically payroll staff, only 6 percent had E-mail access. Under a live WWW prototype begun in 1996, the CES found that 10.7 percent of existing respondents already reporting via TDE met the eligibility criteria of Internet access, Netscape 2.0 browser or better, and were willing to participate in the pilot. These units were concentrated in industries thought likely to have high access rates, specifically Computer and Data Processing Services and State and Local Government. There were no respondents meeting our criteria in other service industries, see Table 1.

3. RESULTS OF THE CES WWW PILOT

Beginning in March 1996, the CES launched a pilot test of WWW collection. The pilot is designed to prove feasibility and to provide a platform for learning design and integration issues and solutions. The evolution of the system will depend on our ability to use new software to address system features. Our view is to get even a rudimentary system in small scale use to work out problems rather than wait for a completely developed system to be ready. The mail-based CES form, collects 5 or 6 data items each month from a fixed panel of respondents. The monthly data are entered on consecutive lines covering an entire year in a set of rows and columns. Column headings state the data item requested.

The WWW pilot replicated this basic row of data for only the current month. Multiple months were not shown due to the current difficulty of guaranteeing the security of previously reported microdata. The column headings are shown as hypertext links to complete definitions of the data items. To enter the system, the respondent must enter their unique report number and password to proceed to a home page providing the Office of Management and Budget-mandated statements on confidentiality and respondent burden. The report number and password are also used to identify the correct industry form to present to the respondent. The respondent enters the number for the reference month and their data. When complete, a button labeled "Send data to BLS" is pressed which encrypts the data and transmits to the BLS server. At this writing prototype interactive editing has been developed and will be provided to WWW respondents in the next few months.

Currently 52 CES respondents report their employment, payroll and hours data by WWW. Each formerly reported via TDE and know the monthly reporting cycle. Each month just prior to the reporting timeframe, a E-mail message is sent as an advance notice, just as advance notice postcards and Faxes are now sent to TDE respondents. Those which do not report within our specific timeframe receive another E-mail as a nonresponse prompt.

This research has yielded two critically important results. First, the basic response rates for our preliminary estimates for these units are the same under WWW, 76% they reported under TDE. Thus, we see no reason at this early juncture why WWW collection will not be able to match the same high response rates seen under similar telephone-based self-reporting CASIC methods: TDE and VR. Second, the E-mail prompts seem to be as effective as other prompting methods. The proportion of WWW units needing E-mail nonresponse prompts, about 25 percent, is very nearly the same for the overall TDE

sample. Thus the overall package of messaging as described in Table 1 is proving effective and comparable to the successful TDE methods.

These two critical methodological features pave the way for more elaborate tests and refinements in methodology, including the packaging of the E-mails such as voice files or graphics.

Table 1: CES WWW Sample Solicitation Results: March-September 1996

	Current Totals	Computer and Data Processing Services (SIC=737) n=313	Other Service SICs n=121	State and Local Government n=264
E-mail only	4.1%	6.4%	0%	2.4%
Company has capabilities, respondent does not use	10.5%	12.8%	7.0%	8.7%
Old/ Incompatible browser	4.5%	6.4%	5.8%	0.6%
E-mail and Web, not on desktop	2.1%	3.4%	0%	1.2%
Compatible browser, E-mail/Web on desktop	***10.7%***	***14.3%***	***0%***	***10.4%***
Prefer TDE	2.1%	3.4%	0%	1.2%
No E-mail or WWW capabilities	62.5%	49.4%	79.1%	74.9%
Out of Business/ Out of Scope	3.5%	3.9%	8.1%	0.6%
Total	100%	100%	100%	100%

The rather low penetration of WWW in both household and establishment environments, regardless of its growth, points directly to the inevitability of mixed mode collection for some years to come. Developing and integrating two or more modes requires multiple systems for control and makes tracking and integration more difficult, extending available systems support and leading to potential for overlooking some important detail.

4. COSTS

Over the decades we have invested large sums of money to develop and refine the labor-intensive centralized and decentralized operations which help ensure the quality of our estimates, these operations include: collection and collection control, multiple stages and modes of nonresponse follow-up, key entry with verification, and editing with reconciliation. However, under WWW reporting, all collection activities can be fully automated and centralized on a dedicated LAN system. Messages are electronically sent at predetermined dates and information checked-in on a flow basis. Edits are implemented as part of the WWW data collection session.

One of the major thrusts of CASIC development is to change the costs structure by replacing the traditional labor-intensive mail-based processes with increasing portions of capital-intensive factors. Labor costs tend to increase while technology costs have tended to decline over time. Under the traditional mail-based methods, clerical or semi-professional staff produce, fold, stuff envelopes at mailout and then open, often pre-edit and key enter the returned questionnaires. CATI offers reduced mail handling and new costs for technology, but retains high labor costs for interviewers. Under TDE, most of the large labor-intensive processes are replaced by respondent entry of data, and telephone costs drive

overall units costs. WWW collection offers essentially free data entry and transmission costs when the survey organization's Internet access costs are spread over the agency and large numbers of respondents. Such access is currently also declining in price as competition increases.

Table 2: Typical CES Monthly Unit Costs of Data Transmission

Function	Mail – postage	TDE/FAX	Internet/WWW
Outbound	$.32	$.08	$.00
Inbound	$.32	$.16	$.00
Non-response Prompting	$.10	$.04	$.00
Total	$.74	$.28	$.00

For all but the smallest surveys, the growing disparity between postage and telephone costs will drive more surveys to CASIC as transmission cost savings overwhelm systems development and maintenance costs. CES unit costs of data transmission are shown in the Table 2 illustrating the dramatic cost reducing opportunity offered by a switch from TDE and Internet collection. For organizations purchasing unlimited Web access, the average cost of a session should approach zero.

5. ORGANIZATIONAL EFFECTS

Our organizations will, as they have over the past two decades, continue to evolve as significant technological improvements are implemented. As discussed above, the large scale implementation of CASIC methods, and WWW collection in particular, will result in the flow of clean, edited data directly to the survey organization. Thus, both labor and non-labor resource allocation will be shifted among the remaining and new factors of production. The view of WWW collection as a "people-less, paperless" methodology (Werking, 1994) has profound implications for our organizations. While full-scale use of WWW for collection is some distance off, we can already see the resources shifting under other methods. Certainly methods such as TDE and VR rely on respondent data entry and decreasing dependence on interviewing staffs.

The specific organizational affects encountered in any survey will depend upon where on the CASIC scale they currently exist. Surveys conducted by mail will be dramatically reorganized, while surveys currently using CATI or TDE will see smaller effects and in different resource categories. For example, a mail or TDE-based survey will see current editing workload dramatically reduced. Surveys using CATI already will see labor costs decline. Surveys using TDE, having already captured the majority of labor savings from mailings and key entry will see spending on editing and telephone charges virtually disappear.

The new costs of systems development and maintenance will receive new resources, likely drawn from the reductions in interviewers. New expenditures will be necessary for network management, new and changing servers and software, including encryption. It is just as important that we fully invest in the infrastructure supporting the new survey environment as it is that we carefully extract all possible savings from our processes. The role of questionnaire designers will grow, particularly for individuals migrating from paper-based surveys.

6. SECURITY

Perhaps the single most critical technological feature of the Internet infrastructure is the security of the transmitted and stored information. This limitation is repeated by every student of the Web and is drawing the attention of much of the computer community. Some characteristics of the Web security profile are:

- Authentication of the respondent.
- Protection against snooping during transmission (Packet data security).
- Protection of the session (hijacking).
- Protection of confidential data once it has arrived at the server.
- Prevent non-user access to the agency LAN.

This field is rapidly changing, new tools and software are on the horizon for changing our current approach. These issues must be addressed by all WWW systems and procedures. The most difficult, and as yet unsolved in the CES pilot system is storing historical data for longitudinal editing outside the firewall. Even when encrypted, sufficiently motivated persons could eventually break encryption codes.

The security issues around WWW collection often seem to be far more exhaustive than for our other CASIC methods. The differences lay in two aspects, the Internet is currently open to persons wishing to look around and to technological snooping, and the Internet is too new to have a body of both statutory and case law acting as restraints on peoples' behavior. The postal system has statutory laws prohibiting interference with the mail, a wide range of case law supporting those statutes, and even a special police force dedicated to catching and prosecuting criminals. The telephone system also has statutory and case law prohibiting wiretapping. For both the mails and telephone systems, the public at large are aware of these laws and criminals are routinely publicized. Laws already exist in the U.S. which may be found to cover the Internet. The Electronic Communications Privacy Act of 1986 and the Computer Fraud and Abuse Act of 1986 may both cover the Internet. Once a formal opinion is available on the applicability of these laws, they can be cited on home pages as a deterrent.

7. THE FUTURE

Towards The Virtual Interviewer: While any student of CASIC methods has their own blue sky view of the future set of methods and technologies, the fascinating research of the two decades will likely reshape our survey processes. The set of WWW-related features built and tested will depend on the individual survey's issues. New technologies are becoming available which provide fascinating opportunities including multi-media audio and video. For example, we could send e-mail messages with audio attachments. Upon clicking on the icon, any usual message such as advance notice or nonresponse prompt would be "Please report your data by the 28th". Or we could include radio of TV clips showing the uses of the data. Also currently, inexpensive cameras exist that can be attached to PCs for on-line video conferencing. It is conceivable that these cameras will be part of the packaging for sale just as CDs are now. Alternatively, surveys could use this camera technology for "On-line help desks". When a respondent spends too much time on a particular item, a pop-up screen could offer more information or examples of the desired data item. The pop-up screen could also display an icon which when clicked, links to a live interviewer whose picture and voice are carried to the respondent's screen. The interviewer would see the same screen and other historical information on the respondent, and be able to talk them through the difficult item or teach them how to use the system. The results of such interviews would lead to improvements preventing such problems.

The role of the interviewer will continue to be shaped by CASIC methods. The WWW provides sufficient computing power and easy user interface to both replace the interviewer and continue to offer the respondent all of the human interfacing that Cyberspace will allow. Based on the currently available technology, we can foresee the "Virtual Interviewer" built on prerecorded video/audio clips analogous to the interactive games now available. A portion of the screen provides a visual representation of the interviewer speaking the questions in any language needed with the questions or form also shown. Response-driven branching, by keyboard or voice recognition, directs the presentation of visual and audio queues. Keyed or spoken responses will drive branching and menuing will offer the respondent more

information in various levels. The Virtual Interviewer vision could be fielded now using interactive CDs, like those our children are using for learning and killing untold numbers of villains.

On-line Questionnaire Development and Debugging: This vision offers survey practitioners the best monitoring and questionnaire refinement capability imaginable. Monitoring can be conducted by key stroke capture or remotely by simply watching the self-response unfold. The Virtual Interviewer will have already requested and stored audio and video consent. Under this view, the WWW system would provide audio prompts and video representation of the interviewer in a small window in a corner of the screen. The interviewer can speak and respond to either keyed or spoken responses.

Internet II: The existing Internet will not fully support the vision outlined above. However, the next generation on the Internet, called Internet II, is on the drawing board and discussions on its design focus on providing exactly the types of features which would make the Virtual Interviewer possible.

8. CONCLUSIONS

The development of strong, fully automated data collection via the WWW is inevitable. The very basic WWW research conducted to date, while slowed by the newness of the technology, is supported by the results of the previous two decades of CASIC research. The methodological issues, again supported by previous CASIC research and implementation, point directly to neutral or improved data quality as compared to other methods, improved timeliness versus the traditional methods and equal to CATI and TDE, at much reduced costs.

Thus, the future is fairly clear, just as it is equally clear that our mail survey operations are remnants of the past. The information superhighway has now opened up many new dynamic areas of electronic collection research for establishment surveys. The results from this research will ultimately position establishment and household surveys as the most timely, accurate, and cost effective of the survey operating environments.

Acknowledgment: The author would like to acknowledge the superior contributions of Louis J. Harrell and Christopher Manning for their untiring and groundbreaking work developing the CES WWW prototype and in analyzing the results.

9. REFERENCES

Dillman, D. A. (1978), "Mail and Telephone Surveys: The Total Design Method", New York: Wiley-Interscience.

Hoffman, D. L., W. D. Kalsbeek, and T. Novak. "Internet Use in the United States: 1995 Baseline Estimates and Preliminary Market Segments". Project 2000 Working Paper, Owen Graduate School of Management, Vanderbilt University, April 12, 1996.

Nielsen, Jakob (1996),"Usability Testing of WWW Designs", http://www.sun.com/sun-on-net/udesign /usabilitytest.html.

Spanski, R, L. Wickham, (1995), "Connecting the Workplace: Electronic Commerce in Business and Government", Study 944013, Louis Harris and Associates, Inc.

Shneiderman, Ben (1993), "Designing the User Interface, Strategies for Effective Human-Computer Interaction", Adison Wesley.

Statistical Policy Working Paper 19 (1990); *Computer Assisted Survey Information Collection*, Office of Management and Budget.

Wall Street Journal, "U.S. Households With Internet Access Doubled to 14.7 Million in Past Year", October 11, 1996, B11.

Werking, George S. (1994), "Designing the Survey Operations of the Future: A Paperless and People-less Collection Environment." Proceeding of the American Statistical Association Joint Statistical Meetings, Invited Panel on the Future of Establishment Surveys, in print.

Werking, G.S. and R.L. Clayton (1995); "Automated Telephone Methods for Business Surveys", Business Survey Methods, Wiley, 317-338.

SESSION 6

RESPONSE ERRORS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

MICRODATA MATCHING: A TOOL FOR EVALUATING AND IMPROVING THE QUALITY OF SURVEY DATA

L. Gervais-Simard[1]

ABSTRACT

This article describes the recent experiments of the Survey Co-ordination and Quality Assurance Section in the area of microdata matching. It deals with three separate microdata matching exercises involving data from Statistics Canada surveys and Revenue Canada administrative files. The purpose of these studies was to evaluate the quality of the survey data, identify non-sampling errors and then establish measures to improve data quality, all with a view to raising the quality level of employment and payroll estimates in the surveys conducted by the Labour Division.

KEY WORDS: Administrative data; Revenue Canada; Survey data quality; Microdata matching; Quality improvement measures.

1. INTRODUCTION

1.1 Context

In 1993, at the request of Statistics Canada, Revenue Canada added two questions to the statement of account for source deductions (PD form) for enterprises. One question dealt with employment and the other with gross payroll. In 1994, the Labour Division began incorporating the data from these two questions (referred to as administrative data) into its monthly estimates. This new process continued to expand, with the result that administrative data now account for more than seventy percent of the estimates in the Survey of Employment, Payrolls and Hours (SEPH). Since the administrative data from aggregated PD forms had already been the object of comparisons with the survey data, we were aware of their positive contribution to the quality of employment and payroll estimates. Starting from this premise, the matching of microdata from these two data sources seemed an appropriate tool for evaluating the quality of the survey data, identifying non-sampling errors, and establishing measures to improve the quality of the survey data. Since the administrative data provided limited information, they will never entirely replace the data obtained from surveys. More limited surveys will still be used to attribute missing characteristics to the administrative data.

In the past year, the Survey Coordination and Quality Assurance Section (SCQA) has conducted three separate microdata matching experiments. These involved matching the following data:

- Annual data from the Survey of Employment, Payrolls and Hours with data from the Summary of Remuneration Paid (T-4) of Revenue Canada;
- Monthly data from the Survey of Employment, Payrolls and Hours with data from the Statement of Account for Source Deductions PD7A(TM)[2] of Revenue Canada;
- Monthly data from the Small Business Payroll Survey with data from the Statement of Account for Current Source Deductions PD7A[3] of Revenue Canada.

---

[1] L. Gervais-Simard, Labour Division, Statistics Canada, Ottawa, Ontario, K1A OT6.

[2] The term PD7A(TM) indicates that the enterprise makes more than one remittance per month to Revenue Canada.

2. MICRODATA MATCHING EXPERIMENTS

2.1 Matching of annual data (1993) from the Survey of Employment, Payrolls and Hours[4] (SEPH) with data from the Summary of Remuneration Paid (T-4 Supplementary form) [5] of Revenue Canada.

2.1.1 Basic information

The data quality to be evaluated was that of the Survey of Employment, Payrolls and Hours, which obtained data on a monthly basis from approximately 43,000 enterprises in 1993. The matching of data from this monthly survey with those from Revenue Canada's Statement of Remuneration Paid (T-4 Supplementary form) was done for 1993. Employers are required to provide Revenue Canada with information on remuneration paid once a year, in this case on February 28, 1994. Since the present study was conducted in summer 1995, the most recent administrative data available were those for 1993. The time lag between the reference period and the time of our study is due to the processing that Revenue Canada must carry out on these data.

Since there is no direct link between the enterprises surveyed by SEPH and the Revenue Canada PD account numbers, it was necessary to use Statistics Canada's Business Register to establish a link between these two data sources. Since the SEPH is a survey of enterprises ,[6] the enterprise was chosen as the basic unit for matching the microdata, with the payroll deduction (PD) account number nevertheless being the common denominator.

2.1.2 Goals and constraints

The goals of this study were to:

- determine the actual discrepancy between the two data sources,
- find the reasons for the discrepancies by contacting SEPH respondents,
- correct the weakness of the survey by establishing quality measures.

At the outset, our team received instructions to match data for which discrepancies in employment figures had already been noted between SEPH and the Labour Force Survey (LFS). To comply with these instructions, enterprises in the Non-Durable goods sector were targeted for the purposes of our study. The second constraint was one of time; it was therefore necessary to limit the enterprises selected to those identified as having only one PD number, thereby reducing the number of enterprises to be contacted.

2.1.3 Universe

We thus identified 1,246 enterprises in the Non-Durable Goods Sector that satisfied the above-mentioned condition, namely: 1 enterprise = 1 PD number = 1 or more establishments.[7] These enterprises represent more than 30% of all the enterprises (4,066) in the SEPH for which at least one establishment was operating in this economic sector.

---

[3] The term PD7A indicates that the enterprise makes only one remittance per month to Revenue Canada.
[4] See Sources and Definitions at end of article.
[5] See Sources and Definitions at end of article.
[6] See definition of enterprise at end of article.
[7] See definition of establishment at end of article.

For microdata matching purposes, some enterprises had to be excluded, namely:

- all enterprises for which data were not available for all the months of 1993. In actual fact, such enterprises were those with one or more establishments that either entered or left the SEPH sample during the year;
- all enterprises for which at least one month of data was imputed for one or more establishments as part of the SEPH during the 12 months of 1993; and lastly,
- all enterprises with at least one establishment that was not part of industry group 37 (Non-Durable goods).

Following these exclusions, 99 enterprises were retained for microdata matching. Two variables common to the two data sources were analysed: number of employees[8] and gross payroll.[9]

2.1.4 Matching of variables

Since SEPH is a monthly survey, the maximum number of employees reported in any month during the year for each enterprise was used to represent the number of employees for 1993. On the Revenue Canada side, it was the number of unique social insurance numbers reported by each enterprise that represented the number of employees. Consequently, we cannot expect to obtain identical results from the two data sources; at best, the results may be relatively close.

If there was no problem with the quality of the survey data, then the data obtained from the two sources (SEPH and Revenue Canada) for the annual gross payroll variable should be identical. The response guide clearly specifies that the gross payroll amount should be equivalent to the total of all amounts that would appear in Box 14 of the T-4 slip. So that the data for this variable could be matched, the survey data for each of the 12 months of 1993 were totalled.

2.1.5 Gross discrepancy following matching

At this stage, the matching of the 99 enterprises retained shows the gross discrepancy between the two data sources for each of the variables. The actual discrepancy will be calculated at the end of the exercise, after each of the enterprises has been surveyed. For the moment, these results are not significant.

Source	Employees	Gross Payroll $ (000)
T-4 slip (Revenue Canada)	38,834	$ 636, 326
SEPH (Statistics Canada)	30,389	$653,549
Overall discrepancy	8,445 or **27.8%**	- $ 17 222 or **-2,6%**

[8] For definition of employee, see Sources and Definitions at end of article.
[9] For definition of gross payroll, see Sources and Definitions at end of article.

2.1.6 Survey technique and questionnaire

The SEPH respondents were contacted by telephone and asked to answer a questionnaire[10] on the discrepancy between the data from the two sources. Before bringing up this topic, the interviewer was required to verify that the link between the enterprise and the PD number was unique; if not, the enterprise was excluded from the study. After verifying the link, the interviewer explained that in making a comparison between the data reported to Statistics Canada and those supplied to Revenue Canada, discrepancies were discovered, some of them quite sizable. The respondent was then informed of the size of the discrepancy in the two figures relating to the enterprise for two common variables, namely the number of employees and gross payroll. Then questions were asked to determine the causes of the under-reporting or over-reporting on the SEPH questionnaire.

2.1.7 Observations following contact with respondents

The somewhat imperfect match of the microdata for the employee variable became clearer after the interviews with the respondents. The latter explained the discrepancy between the data for that variable as follows: this is an economic sector characterized by very high personnel turnover, accentuated by low wages in the textile and food industries. Since each person has a unique social insurance number (SIN), and since the total number of SINs constitutes the number of employees for the administrative file, it is therefore understandable that this number would be much greater (27.8%) than the number of employees reported to SEPH. Therefore it was impossible to determine what the actual discrepancy for the number of employees was between the two data sources. Therefore the analysis of this variable cannot be carried further.

2.1.8 Factors explaining the gross discrepancy for the gross payroll variable

After the respondents had offered explanations on the causes of the discrepancies between gross payroll reported to Statistics Canada and the figure supplied to Revenue Canada, we were able to eliminate those enterprises that were distorting the results of the matching. Eight enterprises had to be excluded because the link between the enterprise and the PD number was not unique. The information from the Central Frame Data Base (CFDB) that was used to select the enterprises was therefore not accurate for those eight. In addition, two other enterprises were excluded because of the concepts used in SEPH. Their PD numbers included employees whose work activity placed them outside the SEPH target population, namely fishers. Thus, if we subtract the discrepancies represented by the ten enterprises mentioned above, we obtain a net discrepancy, which indicates a positive result of **7.8% as opposed to the gross discrepancy of -2.7%**. A positive result indicates that when microdata from the Non-Durable Goods Sector are matched, the gross payroll amount obtained from Revenue Canada is greater.

[10] The questionnaire may be obtained by contacting the author of this article.

The net discrepancy results from an amalgam of the 89 remaining enterprises for which positive or negative discrepancies were registered. The following table shows how they are distributed.

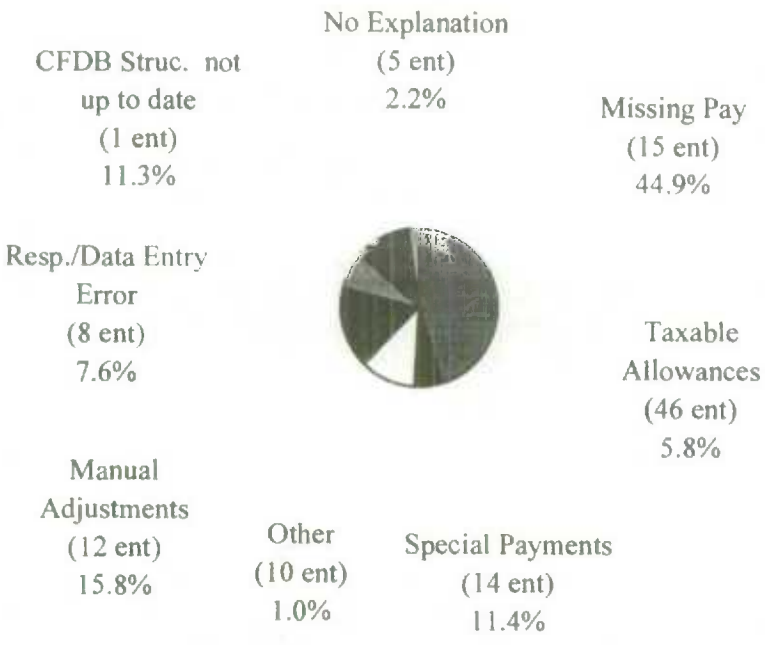
Distribution of the net discrepancy	$ (000)	Number of enterprises
Positive part of net discrepancy	$48,692 or discrepancy of **9.0%**	70 enterprises
Negative part of net discrepancy	- $3,882 or discrepancy of **- 5,4%**	19 enterprises
Net discrepancy	$44,809 or discrepancy of **7.8 %**	89 enterprises

The following chart shows separately the factors contributing to the positive part of the net discrepancy (70 enterprises) observed with respect to enterprises in the Non-Durable Goods Sector.

Chart 1

Distribution of the positive part of the NET DISCREPANCY, by reason
Gross Payroll 1993
(Non-durable Goods Sector)

Discrepancy of $48,8 million or 9.0%
(70 enterprises)



The factors shown in **Chart 1** are explained below:

Missing pay: this is the pay of executives and administrative staff as well as so-called confidential pay that respondents admitted not reporting to SEPH. It is easy for respondents not to include these types of pay, since they are often treated separately and are therefore overlooked.
Manual adjustments: this part of the discrepancy is strictly attributable to adjustments made in the edit or analysis phase of the survey. The respondent is in no way responsible for this situation.
Special payments: these are amounts that are paid to employees for work performed or for other entitlements that do not relate exclusively to the last pay period of the month. These payments are made at any time during the month, and more importantly they are not part of regular wages and salaries. Respondents tend not to include them, owing to their special nature. In addition, there are sometimes problems relating to the accessibility or availability of the information.
Structures in the Central Frame Data Base (CFDB), not up to date: the information on the structures of enterprises is not up to date, and this indirectly affects the SEPH data. If some establishments cannot be linked to a enterprise on the CFDB, they cannot be surveyed.
Respondent error or data entry error: the discrepancy observed was attributed to these factors when uncorrected SEPH data were not equivalent to the data maintained by the respondent in its records.
Taxable allowances: these are taxable federal allowances and benefits recognized by Revenue Canada.
No explanation: neither the respondent nor the interviewer was able to provide a plausible explanation for the discrepancy observed. This factor applied to enterprises where the discrepancies were minimal.
Unable to contact respondent: the telephone number is no longer in service, suggesting that the enterprise has shut its doors.
Other: this includes enterprises that refused to answer the questions and those that had made annual adjustments, which were not reported to SEPH.

As regards the negative part of the net discrepancy, to which 19 enterprises had contributed, it was noted that 77.4% was due to manual adjustments made during the survey.

2.1.9 Immediate actions to improve quality

During the interviews, respondents who were not reporting all the information required were encouraged to do so as soon as possible. They were also told how important good response quality was for Statistics Canada. When new resource persons were identified, their names were referred to the Operations and Integration Division so that it could incorporate them into the survey's master file. In addition, the Business Register was informed of any structural irregularities that came to light during the interview.

Broader quality measures with a more long-term impact were also developed following the first two microdata matching studies described in this article. They are described in Section 2.3 of this article.

2.2 Matching of monthly data from the Survey of Employment, Payrolls and Hours (SEPH) and Revenue Canada's Statement of Account for Current Source Deductions, Form PD7A(TM)

2.2.1 Basic information

In the redesign of the Survey of Employment, Payrolls and Hours, the quality improvement group was asked to contact enterprises for which sizable and inexplicable discrepancies had been noted in the microdata matching process. The Business Survey Methods Division (BSMD), in co-operation with the quality improvement group, had matched microdata from SEPH with those from the Statement of Account for Current Source Deductions (form PD7A(TM)) of Revenue Canada for May 1995. The primary goal of this matching exercise was to test the appropriateness of the rules developed by the Business Survey Methods Division for estimating the number of employees on the basis of administrative data. Those rules were designed to obtain a single figure for the number of employees and gross payroll in view of the fact that several remittances are made for each PD to Revenue Canada every month. However, with interviews of the enterprises, this exercise could become an additional source of information on the quality of SEPH data.

Micro-matching of data from these two sources is possible owing to the fact that two boxes on the payroll deduction form PD7A(TM) ask enterprises to indicate the number of employees in the latest pay period as well as gross payroll. It was at the request of Statistics Canada that these boxes were added to the payroll deduction form in 1993.

2.2.2 Enterprises selected

As a first step, matching was carried out on enterprises that had been identified as having a single link between the enterprise and the PD number. From this initial matching of 1,759 microdata, it was found that for **44** enterprises there was a discrepancy of at least 200 employees between the two data sources; for 17 of them, the number of employees was greater according to SEPH, while the situation was reversed for the other 27. As a second step, matching was carried out on another group of 852 enterprises with multiple links, meaning that they were linked to either more than one PD number or more than one establishment. When discrepancies in the number of employees at individual enterprises were examined, it was found that **15** enterprises showed a discrepancy of at least 1,000 employees.

2.2.3 Enterprises selected for interviews

Because of time constraints, only a limited number of enterprises could be contacted. We therefore contacted a range of enterprises that was as representative as possible in terms of the province, the economic sector and, where applicable, the number of links. Accordingly, 5 single-link enterprises were contacted in December 1995 and 9 multiple-link enterprises were contacted in January 1996. For that telephone interview there was no official questionnaire; interviewers informed respondents that the figures reported to SEPH regarding the number of employees or gross payroll differed from those indicated on Revenue Canada form PD7A(TM).

2.2.4 Reasons for discrepancies

The following list summarizes the reasons for discrepancies between the two sources:

Defective/missing links or shared PD number: this is a problem resulting from delays in updating Statistics Canada's Central Frame Data Base (CFDB).
PD numbers reporting pension and welfare data: here the data reported to SEPH are not the problem. However, these PD numbers should have been identified so that they would not be used; this is a problem relating to the management of the CFDB.
Missing pay: vacation pay should have been reported to SEPH with the amounts of special payments. Also, confidential pay was not included.
Way of responding to Revenue Canada: deficiencies in the response to Revenue Canada cause estimation problems. In some cases it is not the respondent who is at fault, but rather the banking service that completes the form for the respondent.
Manual changes: in one case, changes made by the analysts lowered the SEPH data.

3.1 Quality improvement measures undertaken following the previous two studies

3.1.2 Actions taken to reduce discrepancies

- In order to collect missing information on certain types of pay, allowances and special payments, the following actions were taken:
  1) remind respondents of the SEPH concepts;
  2) review those same concepts with interviewers in the regional offices;
  3) plan to re-contact SEPH enterprises on an ongoing basis in order to verify their structures and response habits.

- There is some possibility that the introduction of a more effective edit system in the redesign of SEPH has served to reduce the impact of manual adjustments made during surveys.

- As to reducing response and data entry errors that have a major impact on employment and payroll estimates, a detection procedure was put in place to catch these errors and correct them before the data are compiled.

- In addition, so that the Business Register would be more up to date, all structural problems found in it were reported on an ongoing basis.

3.2 Matching of monthly data from the Small Enterprise Payroll Survey with data from the Statement of Account for Current Sources Deductions (form PD7A) of Revenue Canada

3.2.1 Description of the Small Business Payroll Survey

The Small Business Payroll Survey (SBPS), which is based on a subsample of 7,500 units drawn from 86,000 payroll deduction accounts, was set up in March 1994. Those units consist of small enterprises with fewer than 100 employees, for which a single establishment was identified. This survey, which collects information for all the variables in SEPH, is used to attribute monthly the characteristics missing in the administrative data (PD7A), which were already accounting for 30% of the employment and payroll estimates.

3.2.2 Matching of microdata

Since the SBPS subsample is drawn from monthly PD7-A accounts, it was not difficult to match the microdata supplied to Revenue Canada with those sent to Statistics Canada. The task was further facilitated by the fact that the relationship between enterprise, establishment and PD number was unique. In addition, since the data from the two information sources are available monthly, it was decided that the microdata would be matched at that frequency. Establishing a time series on the discrepancies in the matched data offered a useful perspective for analysing the quality of the survey data. This time series was therefore established as soon as the SBPS was created, in March 1994.

As in the previous two experiments, the matched variables were the number of employees and monthly gross payroll. Of the 2,500 enterprises surveyed between March 1994 and April 1995, it was possible to match an average of 2,300. The failure to match some enterprises is due either to missing data, major data entry errors in one of the two sources or reasons such as refusals or the non-availability of data on the survey side. The average distribution of the matched data for this period was as follows: in 68% of cases the number of employees was identical, in 16% the administrative data had been imputed, and in 16% there were discrepancies that may be described as real.

3.2.3 Results of matchings

For the mean of the 391 enterprises (6%) for which discrepancies were registered, here are a few statistics that serve to quantify the discrepancies observed in March 1994 to April 1995:

Variable	Average discrepancy %	Maximum discrepancy %	Minimum discrepancy %
Employees	7.7	13.9 (January 1995)	1.0 (April 1996)
Gross payroll	6.7	14.8 (April 1995)	0.2 (February 1996)

The discrepancies observed are positive 92% of the time for the employee variable and 85% of the time for the gross payroll variable. A positive discrepancy means that the data from the administrative source yield higher figures than those from the SBPS.

3.2.4 Reasons for discrepancies observed

An analysis of the discrepancies according to the SBPS response code revealed that there was a sizable number of enterprises for which the discrepancy between the two sources was 100%, when the enterprises in question were closed (code 8) and/or had nothing to report (code 9). The discrepancies observed between the two sources attributable to these two nonresponse codes contributed to the survey's underestimation of both the number of employees and gross payroll. Hence it was necessary to focus more specifically on the enterprises reporting such results. The tool used was the list of comments written by the interviewers, which had been made compulsory in view of the number of enterprises for which response codes 8 and 9 had been used. An examination of the comments relating to 57 enterprises out of 74 for September 1994 revealed the following:

1) when the nonresponse code stated that the enterprise had closed its doors (code 8),
   - the date of closure was noted in most of the comments examined;
   - on average, comments were provided in only 5 cases out of 8.
2) when the nonresponse code stated that the enterprise had nothing to report (code 9).
   - often the interviewers did not include working owners, partners and other heads of incorporated enterprises, who are covered by the definition of employee. Interviewers noted in the comments that there were no employees, but only working owners.

Another aspect of the matching also deserves to be examined, namely where the number of employees is equal according to the two sources but the gross payroll amount is different. Examining this aspect leads to the discovery of other factors that may result in the SBPS data being low. An analysis of data was therefore undertaken, focussing on 37 of the 95 enterprises to which this situation applied in the reference month of November 1994. The work instrument used here was the complete information from the SBPS questionnaire. By examining that information, we were able to identify the following possible reasons for the discrepancies observed:

1) When special payments are involved, the discrepancy may be due to pay periods that differ, especially when the employees are paid either weekly or biweekly, even though the dates of the pay periods must be identical between the two sources. This has the effect of causing the SBPS data to be high, which, however, is rare.

2) The frequent omission of special payments is definitely a major factor in the SBPS data being low, just as it was in the case of the SEPH data.

3.2.5 Recommendations made to interviewers

These two analysis showed that it was necessary to clarify the concept of active owner, which seemed not to have been well understood by the interviewers. It was also necessary to explain to them the importance of obtaining figures on special payments from respondents, who often believe that reporting regular pay amounts for their business is quite sufficient. In addition, to pursue our studies on the quality of the data, it was considered essential that interviewers always write down comments when the response code is either 8 or 9. The Survey Operations Division therefore made a point of communicating our research findings to interviewers in four regional offices.

3.2.6 Quality indicators

In order to maintain and improve the quality of the survey data, it is essential to clarify certain situations with the interviewers and also to enhance their training. In addition, establishing and disseminating quality measurements on a monthly basis may also ultimately help to raise the quality of employment and payroll estimates. Merely showing interviewers the gap that exists between the two data sources would not provide an accurate picture of the quality of a certain part of the work performed. However, if we were to eliminate from the comparison those records in which the contact between the respondent and the interviewer is not at issue (minor data capture errors perceptible, intentional or inadvertent errors on the respondent's part, conceptual differences), the resulting adjusted comparison could then be used as a quality indicator. Evaluation of the quality of the contact between the interviewer and the respondent is done at head office by examining interviewers' comments, which are now quite thorough. Thus, starting with the reference month of May 1995, two quality indicators – one for the employee variable and the other for the gross payroll variable – were introduced into the monthly production process.

4. CONCLUSION

For the Survey Coordination and Quality Assurance Section, microdata matching has been a tool that has made it possible both to better identify non-sampling errors and then to implement corrective measures. The three studies clearly confirmed that non-sampling errors tended to underestimate employment and payroll. These studies contributed to both the implementation of a major project to improve the quality of the data on complex enterprises and the creation of a quality indicator for SBPS interviewers.

Since the trend is toward replacing a growing number of survey units by administrative data to which it is necessary to impute characteristics drawn from the surveys in question, improving the quality of the survey data will always be a timely topic. Therefore, so long as non-sampling errors exist, the Survey Coordination and Quality Assurance Section will strive to eliminate them.

5. REFERENCES

Paillé, B. *PD7A/SEPH Micro Match Analysis: An Evaluation of Administrative Employment Data*, Labour Division, Statistics Canada, February 1994.

3.2.3 Results of matchings

For the mean of the 391 enterprises (6%) for which discrepancies were registered, here are a few statistics that serve to quantify the discrepancies observed in March 1994 to April 1995:

Variable	Average discrepancy %	Maximum discrepancy %	Minimum discrepancy %
Employees	7.7	13.9 (January 1995)	1.0 (April 1996)
Gross payroll	6.7	14.8 (April 1995)	0.2 (February 1996)

The discrepancies observed are positive 92% of the time for the employee variable and 85% of the time for the gross payroll variable. A positive discrepancy means that the data from the administrative source yield higher figures than those from the SBPS.

3.2.4 Reasons for discrepancies observed

An analysis of the discrepancies according to the SBPS response code revealed that there was a sizable number of enterprises for which the discrepancy between the two sources was 100%, when the enterprises in question were closed (code 8) and/or had nothing to report (code 9). The discrepancies observed between the two sources attributable to these two nonresponse codes contributed to the survey's underestimation of both the number of employees and gross payroll. Hence it was necessary to focus more specifically on the enterprises reporting such results. The tool used was the list of comments written by the interviewers, which had been made compulsory in view of the number of enterprises for which response codes 8 and 9 had been used. An examination of the comments relating to 57 enterprises out of 74 for September 1994 revealed the following:

1) when the nonresponse code stated that the enterprise had closed its doors (code 8),
   - the date of closure was noted in most of the comments examined;
   - on average, comments were provided in only 5 cases out of 8.
2) when the nonresponse code stated that the enterprise had nothing to report (code 9).
   - often the interviewers did not include working owners, partners and other heads of incorporated enterprises, who are covered by the definition of employee. Interviewers noted in the comments that there were no employees, but only working owners.

Another aspect of the matching also deserves to be examined, namely where the number of employees is equal according to the two sources but the gross payroll amount is different. Examining this aspect leads to the discovery of other factors that may result in the SBPS data being low. An analysis of data was therefore undertaken, focussing on 37 of the 95 enterprises to which this situation applied in the reference month of November 1994. The work instrument used here was the complete information from the SBPS questionnaire. By examining that information, we were able to identify the following possible reasons for the discrepancies observed:

1) When special payments are involved, the discrepancy may be due to pay periods that differ, especially when the employees are paid either weekly or biweekly, even though the dates of the pay periods must be identical between the two sources. This has the effect of causing the SBPS data to be high, which, however, is rare.

2) The frequent omission of special payments is definitely a major factor in the SBPS data being low, just as it was in the case of the SEPH data.

3.2.5 Recommendations made to interviewers

These two analysis showed that it was necessary to clarify the concept of active owner, which seemed not to have been well understood by the interviewers. It was also necessary to explain to them the importance of obtaining figures on special payments from respondents, who often believe that reporting regular pay amounts for their business is quite sufficient. In addition, to pursue our studies on the quality of the data, it was considered essential that interviewers always write down comments when the response code is either 8 or 9. The Survey Operations Division therefore made a point of communicating our research findings to interviewers in four regional offices.

3.2.6 Quality indicators

In order to maintain and improve the quality of the survey data, it is essential to clarify certain situations with the interviewers and also to enhance their training. In addition, establishing and disseminating quality measurements on a monthly basis may also ultimately help to raise the quality of employment and payroll estimates. Merely showing interviewers the gap that exists between the two data sources would not provide an accurate picture of the quality of a certain part of the work performed. However, if we were to eliminate from the comparison those records in which the contact between the respondent and the interviewer is not at issue (minor data capture errors perceptible, intentional or inadvertent errors on the respondent's part, conceptual differences), the resulting adjusted comparison could then be used as a quality indicator. Evaluation of the quality of the contact between the interviewer and the respondent is done at head office by examining interviewers' comments, which are now quite thorough. Thus, starting with the reference month of May 1995, two quality indicators – one for the employee variable and the other for the gross payroll variable – were introduced into the monthly production process.

4. CONCLUSION

For the Survey Coordination and Quality Assurance Section, microdata matching has been a tool that has made it possible both to better identify non-sampling errors and then to implement corrective measures. The three studies clearly confirmed that non-sampling errors tended to underestimate employment and payroll. These studies contributed to both the implementation of a major project to improve the quality of the data on complex enterprises and the creation of a quality indicator for SBPS interviewers.

Since the trend is toward replacing a growing number of survey units by administrative data to which it is necessary to impute characteristics drawn from the surveys in question, improving the quality of the survey data will always be a timely topic. Therefore, so long as non-sampling errors exist, the Survey Coordination and Quality Assurance Section will strive to eliminate them.

5. REFERENCES

Paillé, B. *PD7A/SEPH Micro Match Analysis: An Evaluation of Administrative Employment Data*, Labour Division, Statistics Canada, February 1994.

Sources and Definitions

Survey of Employment, Payrolls and Hours: the results of this survey are published by Statistics Canada under the name 'Employment, Earnings and Hours', Catalogue No. 72-002, monthly.

Revenue Canada forms T-4, PD7A, PD7A(TM): for further information, see "Employer's Guide to Payroll Deductions – Basic Information."

Employee: any person receiving pay for services rendered in Canada or for paid absence, and for whom the employer is required to complete a Revenue Canada T-4 Supplementary Form. These persons may work on a full-time, part-time, casual or temporary basis.

Gross payroll: this is the total remuneration paid to employees in the survey reference month, before deductions. The amount should be equivalent to the total of all amounts that would appear in Box 14 of the T-4 slip. It includes: regular wages and salaries; commissions; overtime pay; paid leave; piecework payments; special payments; and taxable federal allowances and benefits that are recognized by Revenue Canada.

Enterprise: An enterprise is any business or institution, whether incorporated or not; included are sole proprietorships, partnerships, companies and other forms of organizations.

Establishment: For statistical purposes, this is the smallest entity capable of reporting statistics of economic production; typically an establishment is a factory, mine, store or similar unit.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

THE VALIDITY OF SELF-REPORTED CHRONIC CONDITIONS IN THE NATIONAL POPULATION HEALTH SURVEY

Gary Catlin, Karen Roberts[1] and Susan Ingram

ABSTRACT

The validity of information reported in a survey is critical to the confidence researchers have in the data. The National Population Health Survey (NPHS) is a longitudinal survey conducted every two years designed to provide information on health status, factors affecting health and use of health services. The longitudinal nature of the survey places a particular emphasis on the accuracy of the change in characteristics over time. One important health measure the NPHS has collected is chronic conditions such as asthma, arthritis, high blood pressure, diabetes, migraine headaches, epilepsy, ulcers and effects of a stroke. These data were collected in 1994/95 and are being collected again in 1996 from the same respondents. This study will utilize the respondents in the first two data collection periods of the 1996 survey, a sample of approximately 7,000 households. In the second survey cycle we will be probing to find out reasons for differences between the two points in time. The reasons for change will include real change due to the onset or disappearance of conditions as well as non-sampling error. This comparison will study the types of conditions which are more accurately reported and some of the factors that affect the accuracy of the reported information.

[1]Karen Roberts, Health Statistics Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

SURVEY ON SMOKING IN CANADA

Lecily Hunter[1]

ABSTRACT

The Survey on Smoking in Canada (April 1994 to February 1995) was designed as a longitudinal survey with four contacts at 3-month intervals. The objective of the survey was to measure changes in smoking rates following a reduction in cigarette taxes. Contrary to expectations, the survey indicated a decrease in smoking prevalence. This raised questions as to whether the results were real, or unduly affected by non-sampling errors.

KEY WORDS: Smoking; Non-response bias; Question wording; Longitudinal survey.

1. INTRODUCTION

1.1 Background

Surveys on smoking conducted over the past 10 years have consistently reported smoking rates between about 29% and 34%. These surveys have used a variety of methodologies ranging from Labour Force Survey supplements to Random Digit Dialling; proxy and non-proxy responses; sample sizes varying from 10,000 to 20,000; collection in a single month, quarterly or monthly. The state of question development had evolved since the earliest Statistics Canada surveys in the 1970's to a point where most health experts felt that the survey estimates on smoking prevalence were quite reliable. Nevertheless, in the early 1990's people began noticing a large decrease in the number of cigarettes manufactured for domestic sales while at the same time there was no corresponding decrease in the number of people smoking or the number of cigarettes they smoked. At this time there was a growing disparity between the cost of purchasing cigarettes in Canada as compared to prices in the United States, primarily because of Canada's heavier rate of taxation. The conclusion which was eventually reached was that the majority of cigarettes being manufactured in Canada for sale in the U.S. (which had undergone a huge increase over the same time period) was being smuggled back into Canada and sold on the black market, at prices substantially lower than those consumers would pay at the corner store.

In February 1994, the federal government made the decision to cut taxes on Canadian cigarettes in an attempt to combat cigarette smuggling. Many provinces also decided to cut provincial taxes on cigarettes, resulting in a net reduction of up to 50% on the retail purchase price of cigarettes. Health Canada and other health agencies were concerned that the lower prices would lead to an increase in smoking prevalence and an increase in amount smoked. As a result, Health Canada wanted to conduct a survey which would measure and monitor changes in the smoking rate over the 12 months following the price change.

Every percentage point in the prevalence estimate was expected to be contentious to people on each side of this highly polarized issue. Knowing this, the questions selected for the survey were those which had been used successfully in past surveys and were expected to measure smoking rates with the highest level of accuracy. Nevertheless, it was recognized by both Health Canada and Statistics Canada that certain types of behaviour cannot be measured as accurately as one might like because of factors such as rapidity of change, transition states (impacting on respondents' ability to report accurately), and perceptions of social acceptability (impacting on willingness of the respondents to report accurately).

---

[1]Lecily Hunter, Special Surveys Division, Statistics Canada, 5D5 Jean Talon Building, Tunney's Pasture, Canada K1A 0T6.

Statistics Canada conducted the Survey on Smoking in Canada (SoSiC) as a longitudinal survey with four interviews, three months apart. The initial results showed a prevalence rate of about 30%, the same level as in 1991. The follow-up interviews showed decreasing rates, down to 27% by the fourth quarter. This went against all the expectations of the health experts, who called into question the validity of the survey results. This paper reports the various methods which were used to examine the information available to assess the possible impact of non-sampling errors on the survey.

1.2 Survey Design

The initial sample was selected using Random Digit Dialling (RDD) methodology. One respondent age 15 or over was selected per household. This selection was done with unequal probability to satisfy Health Canada's requirement of reliability within each of four age groupings: 15-19, 20-24, 25-64 and 65+. Furthermore, a portion of the households with only 25-64 year-olds were screened out, to reduce costs.

The respondents from this initial interview were re-contacted three, six and nine months later. This satisfied Health Canada's second requirement of being able to detect relatively small changes in the rate of smoking prevalence.

The interviewing was done using Computer-Assisted Telephone Interviewing (CATI), which allowed us to satisfy Health Canada's third requirement of getting into the field and turning the data around quickly. A base set of questions was used in each quarter to determine the current smoking status of the respondents, but the remainder of the questions varied each quarter to allow a focus on different aspects of the smoking issue. Data collection took place in April/May, August and November 1994 and in February 1995.

1.3 The Expectations

Many health experts expected to see a dramatic rise in both the proportion of smokers and in the number of cigarettes smoked. This expectation was based on (1) looking at the cigarette manufacturing data and (2) calculations based on "price elasticity" (lower prices lead to increased demand).

1.4 The Results

Table 1. Estimates of smoking rate and amount smoked, by collection cycle of SoSiC.

Collection Cycle	# respondents	proportion of smokers	average # cigarettes smoked
1 (April/May 1994)	15,804	Current- 30.4% Daily - 24.8% Occasional - 5.7%	Current - 15.8 Daily - 18.7 Occasional - 3.2
2 (August 1994)	13,398 (85% of Q1)	Current - 29.9% Daily - 24.0% Occasional - 6.0%	Current - 15.6 Daily - 18.7 Occasional - 3.2
3 (November 1994)	12,808 (81% of Q1)	Current - 29.5% Daily - 24.6% Occasional - 4.9%	Current - 15.6 Daily - 18.3 Occasional - 1.8
4 (February 1995)	12,424 (79% of Q1)	Current - 27.4% Daily - 23.2% Occasional - 4.2%	Current - 15.3 Daily - 17.7 Occasional - 2.0

11,119 of these people responded to the survey in all four quarters.

The group which was felt to be most at risk was young people, under the age of 20. The experts thought that this group was more price-sensitive than other ages, and so would take up smoking in higher number, stick with it longer, and increase their number of cigarettes smoked per day. This would countermand any progress which had been made in the area of educating teens on the dangers of smoking in an effort to keep kids from starting smoking in the first place.

1.5 Reaction

When the results from the first quarter were published, the immediate reaction from both the media and the non-smoking interest groups was that the survey must be flawed because the result of 30.4% (current smokers) was not "high enough". In fact, there was no apparent difference from the previous Statistics Canada survey which measured smoking rates: the 1991 General Social Survey (GSS) estimated current smokers at 31% (26% daily; 5% occasional). [NOTE: GSS91 did not calculate an average number of cigarettes smoked, although they did ask number of cigarettes smoked, using ranges.]

With the release of each subsequent quarter's results showing a downward trend in smoking prevalence, particularly in the fourth quarter, criticism about the reliability of the results grew stronger. To address these issues, we essentially had to answer the question: Is this survey describing "the truth" or were the results contaminated because of non-sampling error?

2. EVALUATION OF THE RESULTS

2.1 Survey Non-Response to the First Interview

Of the 29,149 sample cases which were identified as households, 25,598 responded at the household level. This means that we were able to collect the household roster and either select one person to be the survey respondent or screen out the household.

To determine if our initial sample of 25,598 responding households was representative of the population, we compared the household composition with data from the May 1994 Labour Force Survey (LFS), and with the 1994 GSS data (Table 2).

Table 2. Proportion of households by household composition: SoSiC, LFS and GSS

	SoSiC - Apr/May 1994 (unweighted)	LFS - May 1994 (weighted)	GSS 1994 (unweighted)
One person household	20.3	24.3	24.2
One person 15+ plus at least one person <15	3.5	3.9	3.2
Two people 15+ plus at least one person <15	20.1	19.3	16.9
Two people 15+, no children <15	29.7	30.6	32.0
More than two people 15+, with or without children <15	26.3	22.0	23.7

From the numbers shown in Table 2, it appears the SoSiC was somewhat under-representing smaller (one-person) households and over-representing larger households. This is not an unusual pattern in an RDD survey where it

is typically easier to reach someone at home the larger the size of the household. Two factors probably played a part in these results: the duration of the collection period and the call scheduling system.

Even though it is more difficult to find someone at home in smaller households, extending the collection period by even one or two weeks can have a dramatic impact on the reduction of non-contact cases (the major source of non-response in SoSiC). The collection period for this survey was six weeks, which appears to have been too short for the size of our sample.

The automated call scheduling system also had an impact on the collection. A large portion of the sample was not called for the first time until at least halfway through the collection cycle, even though some cases had been called numerous times with Ring-No-Answer outcomes. More sophisticated call scheduling systems allow users to rank cases according to the evaluation of a number of criteria; the criteria to be used as well as the weight assigned to these criteria are flexible and can be set by the survey design team. Because of the more "primitive" nature of the call scheduler used with the SoSiC survey, many cases had a shortened window in which to make contact.

2.2 Survey Non-Response to the Follow-up Interviews

There were 15,804 respondents to the first interview of SoSiC. Even assuming that the initial sample was representative, it is possible that non-response to the follow-up interviews could have biased the results from the subsequent quarters. To test this hypothesis, we first assumed that those people who did respond were truthful and accurate; for those who did not respond, we imputed a smoking status for the quarter(s) which were missing. Different imputation strategies were used and the resulting estimates were compared (Table 3).

Table 3. Smoking rates, after imputation for non-responses to subsequent interviews, using assorted imputation strategies.

	Interview 1	Interview 2	Interview 3	Interview 4
No imputation	30.4	29.9	29.5	27.4
1. Non-respondents were assigned their last known smoking status	30.4	30.3	30.1	28.4
2. Non-respondents were assigned their smoking status from 1st interview	30.4	30.3	30.0	28.5
3. Same as 1., but if person stopped smoking between Jan and May 1994, status was imputed as smoker	30.4	30.4	30.3	28.8
4. Same as 1., but if person had change in smoking status during survey period, status imputed as smoker	30.4	30.4	30.6	29.0
5. Same as 3., but all non-respondents under age 20 had status imputed as smoker	30.4	31.1	31.3	29.8

Even with fairly radical guesses as to the smoking status of non-respondents, in all cases the estimates point to a decrease in smoking rates between quarters 3 and 4. So it seems that the effect we see was not caused by non-response errors - it reflects what people actually reported as changes in their behaviour. This raises the next issues for non-sampling errors: was the survey measuring real changes or were respondents reporting their behaviour incorrectly.

2.3 Learning Curve (familiarity with the questions in subsequent interviews)

One of the possible negative effects of longitudinal surveys is that respondents become familiar with the questions and figure out which answers will get them through the survey fastest. Over time you would expect more and more people going down the "shortest path" if this were the case.

For SoSiC, each interview lasted about 10 minutes. Both smokers and non-smokers were asked approximately the same number of questions; the interview was no more than one minute longer for smokers than for non-smokers. Respondents reporting a change in smoking status since the previous interview were asked two additional questions to determine when the change took place and why.

Altogether, 89.2% of respondents reported no change in smoking status over the 4 interviews. Of the remaining 11.8%, a little over half reported one change while the remainder reported two or more changes. Table 4 below shows the proportion of respondents (who responded in all 4 quarters) who reported a change in smoking status (in either direction). There does not appear to be a trend supporting the hypothesis of learning. However, one important point can be made from the data on change in smoking status. The estimate of change collected retrospectively is about one-half the size of the estimates measured directly; this is likely because transitions such as quitting tend to be frequent and short-lived, and therefore forgotten or ignored when asked retrospectively.

Table 4. Proportion of longitudinal respondents reporting change in smoking status over reference period.

	total % changed	% changed to smoking	% changed to non-smoking
Jan 94 vs. Apr/May 94 (asked retrospectively)	1.6	0.6	1.0
Apr/May 94 vs. Aug 94	4.9	2.1	2.8
Aug 94 vs. Nov 94	4.3	2.1	2.3
Nov 94 vs. Feb 95	4.1	1.1	3.0

2.4 Social Acceptability/Sensitivity

It has been suggested that the anti-smoking campaigns of the past 10 years have created a social stigma about cigarette smoking. If this is the case, respondents may be unwilling to admit that they smoke. There is little hard evidence to support this proposition. Nearly a quarter of the SoSiC respondents reported themselves as smokers in all four interviews. Over the past ten years, there has been relatively little change in the smoking rates overall, but there has been a major shift between daily and occasional smokers. If social acceptablility plays a role in how people answer smoking questions, it most likely has an impact on how they report the frequency of their smoking or how much they smoke, rather than completely denying the smoking behaviour.

2.5 Question Wording

The wording of questions and the context in which they are asked is known to have an effect on the quality of the responses. Since several surveys collected smoking information at about the same time period, we were able to compare results and how they relate to wording (Table 5).

Asking first about current behaviour (GSS91, SoSiC and NPHS) appears to produce dramatically different estimates of current smokers compared to asking first about past behaviour (CADS and GSS95). The differences are mostly in the sub-group of "occasional" smokers. This group has grown in size compared to estimates from about 10 years ago, both because of increases in people trying to quit or starting to smoke, and because of increases in restrictions on where people can smoke.

Table 5. Question wording and resulting estimates of smoking rates from recent surveys asking about smoking.

Survey	Question Text	Smoking Rates
GSS91[2] (monthly collection - annual estimate for 1991)	The next questions are about cigarette smoking. Do you smoke cigarettes daily, occasionally or not at all?	31% current 26% daily 5% occasional
SoSiC[3] (first interview collection in April/May 1994)	(Survey intro) At the present time do you smoke cigarettes every day, occasionally or not at all?	30% current 25% daily 6% occasional
NPHS[4] (quarterly collection - annual estimate for April 1994-June 1995)	The next few questions are about smoking. 1. Does anyone in this household smoke regularly inside the house? 2. At the present time do/does ... smoke cigarettes daily, occasionally or not at all?	31% current 26% daily 5% occasional
CADS[5] (collection in October 1994)	Now I'd like to ask you some questions about smoking. 1. Have you ever been a cigarette smoker? 2. How old were you when you started smoking? 3. At the present time do you smoke cigarettes? 4. Thinking back over the last 7 days, did you smoke any cigarettes? 5. Did you smoke the same number of cigarettes each day?	27% current (no breakdown of daily and occasional was derived)
GSS95[6] (monthly collection - annual estimate for 1995)	Now I am going to ask you a few questions about your exposure to smoke from cigarettes. 1. Have you ever smoked cigarettes? 2. At the present time, do you smoke cigarettes? 3. Do you usually smoke cigarettes every day?	26% current 25% daily 2% occasional

2 General Social Survey (January - December 1991)

3 Survey on Smoking in Canada, Cycle 1 (April/May 1994)

4 National Population Health Survey (April 1994 - June 1995)

5 Canadian Alcohol and Other Drugs Survey (October 1994)

6 General Social Survey (January - December 1995)

These surveys all take the approach of asking the respondent to classify himself as a smoker or a non-smoker, and then all the subsequent questions are determined by that self-classification. For most people, this is not a real problem, but for 5-10% of the population who changes smoking status one or more times during a 12-month period, this classification may not be clear. For CADS and GSS95, the question "have you ever been a cigarette smoker / have you ever smoked cigarettes" may be mis-interpreted for those people for whom smoking is not a regular behaviour. If the respondent smokes occasionally or is just beginning to experiment with smoking, there may be a tendency to answer in the negative, thereby screening themselves out of all subsequent smoking questions.

In future surveys, a better approach may be to have a series of questions which measure specific behaviours within specific time periods. At the analytical stage we then base our definitions on those behaviours rather than asking respondents to classify themselves and having that classification drive the flow of subsequent questions.

2.6 External Effects

Smoking is a volatile characteristic, which makes it difficult to measure. Although the overall smoking rate may appear stable over a period of time, the specific individuals who are smoking or not smoking changes rapidly. At the time of the 4th quarter collection of SoSiC (Feb. 1995), the federal government had just begun running a series of television ads informing the public about the hazards of second-hand smoke on babies and young children. Many provincial governments had recently had other anti-smoking campaigns; Ontario had just implemented laws banning the sale of cigarettes in drug stores and making stores display signs to warn under-age smokers that they would not be sold cigarettes without proper age identification; and the federal government implement a small increase in cigarette taxes. It is possible that some of these factors could have contributed to people (at least temporarily) quitting smoking.

Seasonality may also play a role in smoking. Some people give up smoking when they are active in sports; others may stop smoking as a new year's resolution; young people may start smoking near the beginning of the school year. We do not have any data which can measure such factors reliably, but some health workers believe that such effects may exist.

In the SoSiC data, the largest changes in smoking status occurred between Nov. 94 and Feb. 95 when about 1.1% of the population shifted from non-smoker to smoker, and about 3% of the population shifted from smoker to non-smoker, for a net decrease of 2% in the smoking rate. About 40% of the people who "quit" smoking had been non-smokers at one of the previous interviews, while the other 60% had been smokers in each of the three previous interviews. In looking at the reasons people gave for their change in behaviour, there were no apparent differences between the fourth quarter and the previous three quarters - most people cited health concerns or cost of smoking as their main reasons for stopping smoking.

2.7 Measuring a Moving Target

One of the greatest hurdles to measuring smoking prevalence is that it is always changing. There is always a portion of the population which is in some stage of transition (starting or stopping smoking) at any given time. It is difficult for these people to answer survey questions in a meaningful way since surveys are usually designed with the idea of classifying people into distinct categories which are meaningful for analysis (such as smoker or non-smoker). This problem can be addressed in part by question designers recognizing that these transition groups exist and adjusting their questions and/or definitions to accomodate them.

Table 6. Proportion of population classified into groups related to risk for changing smoking behaviour.

Age 25+, never smoked	low risk of change	30%
Age 25+, smoker in all 4 cycles	low risk of quitting (at least in near future)	20%
Age 25+, former smoker for 5 or more years	low risk of change	20%
Age 25+, recently stopped smoking (<5 years ago) or showed pattern of starting and stopping over 4 cycles	high risk of change (both starting and stopping)	13%
Age 15-24, smoker in all 4 cycles	low risk of stopping smoking (at least in the near future)	4%
Age 15-24, never smoked	high risk of starting smoking	10%
Age 15-24, pattern of starting and stopping smoker, or tried smoking in past	high risk of change (both starting and stopping)	4%

In Table 6 above, the stable populations make up about 74% of the population, these being the people aged 25 or more who have never smoked; or are regular smokers with no indication of attempting to stop; or are former smokers age 25+ who have not smoked in 5 years or longer. The remaining 26% of the population is the group which may be described as "at risk" for changes in smoking behaviour: (4) the people aged 25 or more who have demonstrated a pattern of stopping and re-starting smoking; (5, 6, 7) all people aged 15-25, who are most at risk for starting smoking, but many of whom will not stick with it for very long. Even within the high risk groups, only a portion of these people will actually be in transition at any given moment, so getting a precise estimate of the smoking prevalence rate really amounts to our ability to measure a relatively small, but influential, segment of the population. Depending on what these people are doing at the time we survey them, we could legitimately get estimates ranging from 24% to 36%.

3. CONCLUSIONS

3.1 What really happened to the smoking rates?

There are really two questions to answer: (1) Why didn't the smoking rates show an increase after the taxes (and thus the market price) for cigarettes were decreased?; and (2) Why did SoSiC show a decrease in smoking rates over the 4 interview periods? Much of what was observed in those months following the February 1994 reduction in cigarette costs could be explained (rightly or wrongly) by factors totally unrelated to survey sampling or non-sampling issues. Some examples to consider:

- The majority of people for whom price was an important determinant in their decision to smoke or the amount to smoke had already adjusted their behaviour to the lower prices available through smuggling long before the "official" prices were lowered.
- Lower prices of cigarettes freed up a certain amount of money per week for regular smokers, but that doesn't mean that everyone is going to spend that money on more cigarettes.
- The population is ageing and that large block of people in the baby boomer group is now in the "prime time" for quitting smoking (age 40+). As more and more of these people quit smoking, demographics will be a driving factor in decreases in smoking rates.

3.2 What was the effect of non-sampling error and what have we learned?

Non-sampling errors certainly played a part in influencing the estimates of smoking prevalence, although the extent of this effect is difficult, if not impossible, to measure. The major influences - ones which were preventable - were:

(1) the effect of the call scheduler on reaching households, which had an impact on the types of households contacted;
(2) the effect of question wording (related to survey definitions), which caused a portion of the respondents some difficulty in correctly classifying themselves;
(3) lack of awareness of the types and amounts of transitions going on within the population being measured (also related to survey definitions).

The impact of call scheduling on the estimates was probably minimal because post-stratification in the weighting likely corrected much of the imbalance. Nevertheless, we are currently working on creating a new system for call scheduling which will allow survey designers and managers better control over case management.

The other two effects are best solved by people involved in the measurement of smoking-related behaviour assessing what this survey (and other recent surveys) have demonstrated about the relationship between the analytical need to classify people in a meaningful way and the tools required to acheive that. This means we have to recognize and understand the underlying behaviour patterns in the populations we are trying to measure, then develop definitions based on those behaviours, which can subsequently be transformed into the classifications we want to use. This is almost the opposite of our current approach in which we develop classifications and then design survey questions which attempt to ask people which category they fit in.

Despite the best efforts of survey designers, estimates of populations which are constantly in flux will never be easy to measure. When the issue is as polarized as this one, with health experts, non-smoking activists and tobacco companies fighting over the meaning and reliability of every percentage point, no survey estimate will be accepted without question. Conducting a survey in that kind of environment is a challenge to all involved, and it is absolutely essential that everyone understands the limits of our collection tools (i.e. survey questions) when confronted with measuring certain types of behaviours.

SESSION 7

MEASUREMENT ERRORS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

EVALUATOR ERROR IN THE ASSESSMENT OF INTERVIEWER PERFORMANCE

Paul P. Biemer[1]

ABSTRACT

Because of increasing awareness and concern for interviewer errors in survey work, the evaluation of interviewer performance has become an important part of survey quality improvement. Evaluation of interviewer performance can take many forms including: reinterview programs; telephone interviewer monitoring; in-person interview observation by a supervisor or other senior interviewer; and behaviour coding of interviewers from audio or video tape recordings of the interactions of the interviewer and the respondent. If the purpose of the evaluation is to improve interviewer performance, then it is customary to compare the performance of the interviewer with either another interviewer, the group average, or some other standard for performance.

The problem we deal with in this paper arises when the measures of interviewer performance are biased and do not accurately measure real performance primarily as a result of evaluator error. When the interviewer performance criteria is the same for all interviewers, the primary reason for inconsistent evaluations is the evaluator - that is, the reinterviewer, monitor, behaviour coder, observer, or other rater. First, we review the literature on inter-rater reliability with emphasis on the statistical models that have been used to estimate evaluator variance and the conditions required for estimability conditions and illustrate its use in two examples: one from a telephone monitoring operation conducted at the Research Triangle Institute (RTI) and another from the Bureau of the Census Current Population Survey Reinterview Program.

[1]Paul P. Biemer, Chief Scientist, Research Triangle Institute, 3040 Cornwallis Road, P.O. Box 12194, Research Triangle Park, North Carolina, U.S.A.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

TIME-, RESPONDENT- AND INTERVIEWER-RELATED CAUSES OF ITEM-NONRESPONSE ON THE CES-D DEPRESSION SCALE: A MULTILEVEL MODEL[1]

Pieter van den Eeden[2], Johannes H. Smit[3], and Aart-Jan Beekman[4]

ABSTRACT

Although much effort has been expended on mitigating the effects of item-nonresponse bias, the study of its causes has been less energetically pursued. In this paper we devote attention to time-, respondent-, and interviewer-related causes and we aim both to provide a general model for its analysis and to illustrate this model. Since occasions can be viewed as being nested under respondents, and respondents under interviewers, the multilevel model of analysis is appropriate. We define item-nonresponse as a dependent variable of a dichotomous nature, so the logit regression type of multilevel analysis was used. The illustrative data are adopted from the LASA (Longitudinal Aging Study Amsterdam) project, consisting of 655 55-85 year old people from a nation-wide representative sample in the Netherlands, where the data were collected in seven occasions. The item-nonresponse analysis concerns 20 items of the Center for Epidemiologic Depression Scale (CES-D). Since the study incorporates both interviews and self-administered mail questionnaires it enables us to assess the net effect of the interviewing mode as well as the effects of item-nonresponse in a longitudinal design. It turned out that item-nonresponse is lower in self-administered mailed questionnaires than in interviews, that it steadily decreases over time, that not respondent's gender and age, but their physical difficulties in daily life, depression, and pain experience are the main determinants, and that it is higher for negatively worded items than for positvely worded items. No interviewer-effects were found.

KEY WORDS: Mixed linear models; Variance components; Longitudinal research, Item-nonresponse

1. INTRODUCTION

Item-nonresponse can be defined as missing data occurring in individual items of a survey after the sampled person has agreed to participate in the study. Item-nonresponse can occur for various reasons: the interviewer fails to ask a question, the respondent refuses to provide an answer to a question, the respondent is unable to provide an adequate (scorable) answer to a question, or the interviewer fails to record the answer provided (Groves, 1989: p. 156).

In the present study we will focus on the occurrences and causes of longitudinal item-nonresponse in the Center for Epidemiological Depression Scale (CES-D, Radloff 1977). Data were adopted from the Longitudinal Aging Study Amsterdam (LASA), which covers 3107 respondents, from which a selection of 662 respondents were incorporated in a follow-up study on the course of depression. The follow-up consisted of six successive self-administered mail questionnaires sent at intervals of approximately five

---

[1] This study is based on data collected in the context of the Longitudinal Aging study Amsterdam (LASA), conducted at the department of Psychiatry of the Faculty of Medicine and the department of Sociology and Social Gerontology of the Faculty of Social and Cultural Sciences of the *Vrije Universiteit* of Amsterdam. The study was funded by the Dutch State Ministry of Health, Welfare, and Sports.
[2] Department of Social Research Methodology, *Vrije Universiteit*, De Boelelaan 1081c, 1081 HV Amsterdam, The Netherlands. E-mail: PJWM.vd_Eeden@scw.vu.nl [3] Department of Sociology and Social Gerontology, *Vrije Universiteit*, [4] Department of Psychiatry, *Vrije Universiteit*.

months. Compared with past studies, this study is unique with respect to sample setting and response mode. This is, to our knowledge, the first investigation to examine item-nonresponse in an 'one interview followed by multiple mailed questionnaires' type of longitudinal design.

In survey designs a number of respondents are usually questioned by the same interviewer; so the respondents are typically nested under interviewers. In longitudinal research, time occasions are nested under respondents. As a result, the multilevel model of analysis, using three levels, is appropriate for detecting the effects of time, and respondent- and interviewer characteristics.

In section 2 some theoretical notions and some hypotheses on item-nonresponse will be presented. Section 3 describes the data and design. In section 4 the results are presented; in this section some attention will briefly be paid to unit-nonresponse, since it biases the outcome of analyses of item-nonresponse. Section 5 concludes the paper with a summary and a discussion.

2. THEORETICAL PERSPECTIVE

The emphasis in this paper is on a number of factors related to item-nonresponse studied within the framework of a longitudinal design. These factors are:

Time. Little research on item-nonresponse in longitudinal research has been done. It is hard to formulate a prediction, although it could be supposed that, if item-nonresponse expresses a lack of respondents' motivation and this motivation decreases in time, item-nonresponse will also increase.

Data collection mode. It is widely assumed that face-to-face interviews produce less item-nonresponse because of the opportunities the interviewer has to repeat, to explain questions or to probe to get an answer. Groves and Fultz (1985) note that interviews conduced by telephone show more missing data than to face-to-face interviews. In a meta-analysis of mode-comparison studies on item-nonresponse, De Leeuw (1992) did indeed find that face-to-face interviews had a slightly lower overall item-nonresponse, but she also found that self-administered mail questionnaires perform better when sensitive questions are asked. These findings were replicated in an experiment in which three different modes of survey research were compared (De Leeuw, 1992).

Respondent characteristics. The characteristics most often mentioned in relation to higher item-nonresponse are age, gender, education, and (cognitive) health (Hox, De Leeuw & Kreft, 1991). Age effects have been identified by, among others, Craig and McCann (1978), Ferber (1966), and Francis and Bush (1995). Colsher and Wallace (1989) report in a study among the elderly that item-nonresponse increases with age. They hypothesize that the elderly are unfamiliar with survey research methods and reluctant to share their personal experiences with strangers. Ying (1989), however, found higher item-nonresponse among depressed and unhealthy respondents.

Interviewer characteristics. The systematic effects of specific interviewer characteristics on item-nonresponse are not reported in the literature. For example, Berk and Bernstein (1988) found a negative effect of interviewer age on item-nonresponse, while Collins (1980) did not find a clear effect of interviewer age. Hox, De Leeuw and Kreft (1991) also found no effect of interviewer characteristics on item-nonresponse, although between-interviewer differences were assessed. However, the fact that interviewers can influence item-nonresponse is undisputed (they can probe, they can fail to record, and so on). The emphasis in the case of interviewers should probably be on their actual interviewing behaviour rather than on their background characteristics. In line with this view, personal interviewer style (Rogers,

1976) and suggestions by interviewers that certain questions are hard to answer (Sudman *et al.* 1977) contribute to a higher item-nonresponse rate.

Question wording. There is some evidence that question difficulty is related to item-nonresponse. In a telephone surve, Leigh and Martin (1987) found that the more difficult a question was, the higher was the item-nonresponse. In a study on well-being, Ying (1989) reports that positively formulated items have higher item-nonresponse percentages. In a meta-analysis, Viswesvaran, Barrrick, and Onnes (1993) found consistent effects of item difficulty on the percentage of item-nonresponse. On the other hand, Molenaar (1988) states that although item difficulty seems related to item-nonresponse there is no evidence that this relation is due to the linguistic difficulty of the items (measured with, for example, a Reading Ease Formula).

In the present study a number of questions are formulated. We restrict ourselves to time, the respondent characteristics of gender, age, feelings of pain, physical difficulties in daily life, the CES-D score, and a number of variables related to CES-D (for example, education and the urban environment of the dwelling place), the interviewer characteristics of age and education, and question wording using a positive/negative dichotomy.

The question can be summarized as follows: What relation exists between age and item-nonresponse over time in the case of CES-D, having controlled for data collection mode, and a number of relevant respondent- and interviewer characteristics, and item wording, using the multilevel model of analysis?

3. DATA AND FIELDWORK

The data for illustrating the model are adopted from the Longitudinal Aging Study Amsterdam (LASA). LASA is a longitudinal, multi-disciplinary study of changes in the autonomy and quality of life of older persons, and of predictors and consequences of such changes (Deeg, Knipscheer & Van Tilburg, 1993). The LASA cohort is based on a random sample, stratified for age, gender, and expected mortality over five years. The registries of eleven municipalities in areas in the west, east, and south of the Netherlands provided the sampling frame. Within strata, the sample is representative of the older population (aged 55-85) of the Netherlands. From September 1992 to September 1993, participants in this study were interviewed in the first LASA wave (N = 3107).

To standardize interviewer behaviour an interviewer training course was held (five sessions, each lasting six hours). Video examples illustrated basic interview rules and role playing was used to practice interviewer skills. Additional training sessions were held during fieldwork. The interviewers filled out a questionnaire in order to assess their characteristics. Within regions, respondents were randomly assigned to interviewers. All interviews were tape-recorded in order to control the quality of the data. All interviews were held in the respondents' own residence.

In the baseline interview, minor depression was measured using the Center for Epidemiologic Depression Scale. This is a twenty-items self-reporting scale developed to measure depressive symptoms in the community. It has been widely used in older community samples and has shown good psychometric properties in elderly samples (Hertzog 1990, Radloff 1986). The Dutch translation of this scale had similar psychometric properties in three samples of elderly in the Netherlands (Beekman *et al.* 1994). The absence of an overlap with symptoms of physical illness is crucial in studies of the elderly, and it have shown this to be minimal in a number of studies (Berkman *et al.* 1986). The CES-D scale generates a total score which can range from 0 to 60. In order to identify respondents who were clinically relevant we applied the generally used cut-off score >15.

After the interviews 662 respondents were incorporated into a three-year follow-up study (six waves with an approximately 150-day interval). The database consisted of all respondents above the CES-D cut-off score >15 and a random sample of respondents with a score between 0 and 15 on the CES-D. The follow-up was done by means of a mailshot in which the CES-D scale was determined asking a self-administered questionnaire. Before the start of the follow-up study, seven respondents withdrew from the study (for illness, death, and other reasons) resulting in 655 eligible respondents. In the present paper, data from both the baseline study and the follow-up study are presented.

4. RESULTS

Unit-nonresponse. In a survey study nonresponse is the proportion of people in a sample arranged aimed by an investigator that does not provide the data required. The total nonresponse of a given study can be decomposed in a part that consists of people who did not cooperate with the study ('unit-nonresponse') and another part that consists of people who cooperated but who did not give a useful answer to a posed question ('item-nonresponse'). If nonresponse is random, it does not affect the generalizability of the outcomes to the population. If, however, nonresponse is not random, it is worth to detecting the sources of the bias. The analysis of nonresponse is therefore directed towards the detection of the sources and factors which significantly contribute to bias. If we know what these are, we can try to correct the outcomes, for example by adopting imputation techniques (as described by Little and Rubin, 1987). However, it would seem wiser to obviate the application of such techniques by determining the causes of item-nonresponse.

The analysis of item-nonresponse has to be preceded by an analysis of unit-nonresponse, especially in a longitudinal study. It turned out that the unit-nonresponse rates are 20.5 %, 17.8 %, 31.1 %, 34.4 %, 40.7 % and 31.9 % in occasions 2 to 7, respectively. There is a general steady decline of the unit-nonresponse rate, with occasion 6 as an exception. An explorative analysis of our study showed that unit-nonresponse increases over time, with increasing age and depression, and decreases with educational level of the respondents. Men have lower nonresponse rates than women. Table 1 shows the unit-nonresponse rates per occasion.

Item-nonresponse. First, Table 1 shows the percentages of item-nonresponse per item. The figures of the summarized item-nonresponse are rather low (.4, 3.0, 2.4, 1.4, 1.0, 1.0 and .6 in occasions 1 to 7, respectively). The lowest percentages are at occasion 1, where interviews were held. In the mailed questionnaire mode they show a steady decrease of item-nonresponse over time.

Methodology. Most variables in the study had a skewed distribution. In order to correct this, z-scores for interval variables were used in the analysis. These z-scores were not centred with respect to the levels.

Since the dependent variable is dichotomous and the independent variables are of an interval level, we used logistic regression models for a binary response (Goldstein, 1991). In such a logistic regression model, the logit has to be predicted. In this study the logit is the natural logarithm of the ratio of the probability of item-nonresponse to the probability of item-response. The models were analyzed using the programme MLn (Woodhouse *et al.* 1995).

The research questions were translated into the random coefficient model of multilevel analysis, which, in its simplest form, consists of two equations. The first equation (level 1: occasions) concerns within-occasion regression where the logit of the nonresponse rate per item (INR) as a dependent variable is explained by time. With respect to occasion i of respondent j the equation is as follows:

$$\ln(INR_{ij}/(1\text{-}INR_{ij})) = \beta_{0j} + \beta_{1j}\, t_{ij} + e_{ij} \qquad (1)$$

Table 1. Percentage of item-nonresponse per item and per occasion

occasion	1	2	3	4	5	6	7	
no respondents	655	521	540	517	495	388	446	mean
item								
1	.3	2.7	2.0	1.4	.8	.8	.0	1.14
2	.0	2.3	2.0	1.2	.8	1.0	.0	1.04
3	.0	1.9	2.2	1.4	.6	1.0	.0	1.01
4	1.2	3.5	2.4	1.9	1.4	.8	.4	1.60
5	1.2	1.9	2.4	1.7	1.2	1.3	.2	1.41
6	.3	2.5	2.4	1.4	.6	1.3	.2	1.24
7	.2	2.5	2.6	1.7	.6	1.3	.2	1.30
8	2.5	4.2	3.2	2.3	2.0	1.0	.1	2.19
9	.2	3.5	3.2	2.1	1.1	1.3	.7	1.73
10	.2	3.5	2.6	2.1	1.0	1.0	.3	1.53
11	.2	2.5	2.0	1.4	.8	.8	1.0	1.24
12	1.0	4.0	2.8	2.1	1.6	1.0	1.1	1.94
13	.3	4.2	3.0	1.7	1.1	1.3	1.6	1.89
14	.0	1.7	1.3	1.2	1.2	1.3	1.1	1.11
15	.0	3.5	3.2	1.2	1.2	1.3	1.1	1.64
16	.5	3.8	2.2	1.0	.8	.5	.9	1.39
17	.2	3.1	2.4	1.0	1.0	.8	1.1	1.37
18	.3	2.3	2.6	1.0	1.4	.8	1.3	1.39
19	.2	4.4	2.4	.4	1.2	1.0	.8	1.37
20	.0	3.3	1.3	.8	.6	1.0	.8	1.11

where i refers to the occasion (i = 1,...,7), j indicates the respondent (j = 1,...,J) and ln (INR_{ij}/ (1-INR_{ij})) the dependent variable, β_{0j} is the mean for occasion i, and β_{1j} is the regression coefficient of the dependent variable $\ln(INR_{ij}/(1\text{-}INR_{ij}))$ on t_{ij}. Erroneously neglected dependent variables and measurement errors (among those the effects of mode, respondent and interviewer characteristics, and item wording) are expressed in the disturbance term (e_{ij}). $e_{ij} \sim N(0,\sigma_e^2)$. The common assumptions of regression analysis apply.

At the respondent level the equation runs as follows:

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}\, Z_{1j} + \ldots + \gamma_{kN}\, Z_{Nj} + u_{kj} \qquad (2)$$

where Z_{nj} are independent variables (n=1,...,N) of respondent j, γ_{k0} (k=0,1) indicates the intercept of the regression, γ_{kn} is the regression coefficient of the dependent variable β_{kj} on the independent respondent variable Z_{nj}. u_{kj} ($u_{kj} \sim 0,\sigma_u^2$) refers to the disturbance term belonging to β_{kj}. $E(e_{ij}u_{kj})=0$. Substituting eq. (2) into eq. (1) creates an equation where the terms that contain γ_{kn} refer to the fixed part of the model, and where the terms containing e_{ij} or u_{kj} refer to the random part of the model. This approach has some similarity with that of Singh *et al.* (1996); they did not control the intermediate level of the respondents, however.

In the analysis, interviewers and items are added as a third level; the corresponding equations run analogously.

Table 2 shows the hierarchical nesting of the data as a basis for multilevel analysis. Since the occasions are nested under respondents, the cases with missing values on the dependent variable due to unit-

nonresponse can be omitted. All possible occasions are present in the case of respondent 1, but not in the case of respondent 655.

Table 2. Hierarchical data structure (fictitious example). Item-nonresponse = 1; item-response = 0. Cases with missing values on the dependent variable, due to unit-nonresponse, are omitted from the data file.

occasion number	respondent number	interviewer number	item number	item-k response
1	1	1	1	0
2	1	1	1	1
3	1	1	1	0
4	1	1	1	0
5	1	1	1	0
6	1	1	1	0
7	1	1	1	0
...	...	...	...	...
1	655	24	20	0
3	655	24	20	1
4	655	24	20	0
5	655	24	20	0
7	655	24	20	0

Time effects and mode effects. In addition to the analysis of the decrease of the summed nonresponse in time, we performed the analysis per item. In this analysis, the hypothesis that item-nonresponse is higher in interview situations than in mail questionnaires (the mode effect) is tested. The outcome is shown in Table 3.

Table 3. Item-nonresponse logit of mean and regressions of TIME and MODE. Standard errors in brackets.

item	MEAN		TIME		MODE
1	-3.91	(.44)	-.45	(.12)	1.32	(.39)
2	-7.76	(10.98)	-.35	(.11)	4.77	(10.98)
3	-7.70	(9.00)	-.37	(.12)	4.75	(9.01)
4	-3.32	(.29)	-.37	(.10)	.75	(.52)
5	-4.64	(.56)	-.25	(.10)	1.49	(.23)
6	-4.09	(.43)	-.36	(.11)	1.25	(.38)
7	-4.12	(.56)	-.35	(.10)	1.62	(.52)
8	-2.95	(.23)	-.29	(.08)	.41	(.18)
9	-4.33	(.54)	-.32	(.09)	1.73	(.52)
10	-4.49	(.54)	-.26	(.09)	1.64	(.51)
11	-4.67	(.56)	-.25	(.10)	1.47	(.52)
12	-3.46	(.29)	-.29	(.08)	.83	(.24)
13	-4.06	(.41)	-.27	(.08)	1.36	(.37)
14	-7.63	(16.24)	-.08	(.98)	4.62	(16.24)
15	-7.22	(7.91)	-.29	(.09)	4.44	(7.91)
16	-3.68	(.37)	-.42	(.10)	1.18	(.32)
17	-4.52	(.55)	-.29	(.10)	1.57	(.52)
18	-4.42	(.42)	-.19	(.09)	1.08	(.38)
19	-4.23	(.55)	-.40	(.09)	1.76	(.52)
20	-7.61	(6.44)	-.32	(.11)	4.55	(6.44)

The table shows that all time effects are negative; so item-nonresponse generally tends to decrease for all items. The table supports the mode-effect hypothesis: interviewers have lower item-nonresponse rates than do mail questionnaires.

The interpretation of the figures is as follows. Let us take item 1 as an example. The mean logit is -3.91, which means that on average 4.9 % [$=100*\exp^{-.391}/(1+\exp^{-.391})$] of the respondents tends to refuse. In the

interviews this percentage is 2.7 % [$=100*\exp^{-.391+1.32}/(1+\exp^{-.391+1.32})$], and at the fictitious middle of the mail questionnaire series 5.9 % [$=100*\exp^{-3.91-1.32}/(1+\exp^{-3.91-1.32})$]. At occasion 3, being the middle of the time occasions, this percentage is 5.5 % [$= 100*\exp^{-3.91-.45+1.32}/(1-\exp^{-3.91-.45+1.32})$].

Respondent and interviewer effects. The next step in the analysis is to assess the differences in the time-function and try to explain these by means of respondent variables. The model of analysis can be characterized as follows. Occasions, respondents, and interviewers were incorporated in the analysis as separate levels. The respondent variables mentioned were also incorporated into the model. However, in none of the cases was any significant interviewer difference found. If we restrict ourselves to the first occasion, that of the interviews, the differences are very slight; η^2 generally equals .03.

Table 4 offers the outcome for each item.

Table 4. Significant fixed coefficients of the effects of MEAN, TIME, GENDER and AGE on the logit of item-nonresponse (standard errors are not mentioned).

item	MEAN	TIME	GENDER	AGE
1	-5.22	-.47	1.48	.81
2	-5.84	-.36	1.34	.72
3	-5.60	-.38	1.34	.77
4	-4.27	-.37	.69	.62
5	-4.89	-.26	.77	.58
6	-4.27	-.37	.46	.78
7	-4.91	-.36	.92	.67
8	-3.96	-.29	.57	.58
9	-4.72	-.33	.89	.81
10	-5.31	-.28	1.15	.79
11	-5.30	-.36	.97	.77
12	-5.24	-.39	1.24	.73
13	-5.63	-.27	1.32	.89
14	-5.64	-.00	.83	.58
15	-5.13	-.29	1.03	.82
16	-5.60	-.42	1.37	.80
17	-5.53	-.30	1.13	.84
18	-5.16	-.18	.86	.65
19	-4.61	-.41	.76	-1.01
20	-5.01	-.32	.00	.70

Since time and mode are intercorrelated, however slightly, and the outcome regarding the mode effect hypothesis is so clear, we decided to omit the data obtained by interviews from the dataset. The number of occasions is then restricted to six.

On the basis of this modified database a multilevel analysis was carried out on a sample of all items. Table 5 shows the outcomes. From the significant figures in the random part of Model 1 it can be concluded that the item-nonresponse trajectories do differ among the respondents. As expected, the time-related coefficient is negative. In Model 2 all respondent variables mentioned above were incorporated into the model. It turned out in the fixed part that gender and age effects disappeared, and that difficulty with daily life (DIFF), depression (CES-D), and pain experience (PAIN) turned out to have significant effects. Education, church attendance, urban environment, and absolute and relative subjective health did not show any effect. The respondent-specific trajectories remain.

Table 5. Multilevel analysis of three models over all items on six occasions. The dependent variable is the item response-related logit. In model 1 there is no control of respondent variables; in model 2 all respondent variables are introduced. In model 3, item is added as a third level. Standard errors in brackets; non-significant coefficients omitted.

	model 1		model 2		model 3	
fixed part						
CONS	-.98	(.17)	-2.01	(.13)	-1.40	(.11)
TIME	-.94	(.10)	-1.31	(.08)	-.70	(.04)
DIFF			5.19	(.16)	-.16	(.06)
CES-D			.37	(.06)	.54	(.05)
PAIN			3.92	(.13)	-.43	(.05)
NEG/POS					-1.30	(.15)
random part						
occasion level						
CONS/CONS					.00	(.00)
respondent level						
CONS/CONS	63.86	(2.82)	15.34	(1.59)	6.24	(.86)
CONS/TIME	-42.73	(1.70)	-14.62	(1.08)	-2.44	(.38)
TIME/TIME	27.62	(1.04)	11.68	(.74)	.16	(.16)
time level						
CONS/CONS	1	(0)	1	(0)	1	(0)

Item wording effects. One of our questions is whether or not there are differences between items according to their positive or negative formulation. In order to assess a possible effect, we added items as a third level and the negative *vs.* positive formulation as a variable. Model 3 in Table 5 reveals that, in addition to the effect of the respondent variables, negatively worded items have a lower probability on item-nonresponse than the positively worded items.

6. CONCLUSIONS

This paper has mainly been concerned with how to analyse the item-nonresponse trajectory, taking into account five possible sources of bias. In order to obtain some insight into this problem, the following questions were considered: (a) What relation exists between item-nonresponse and time? In our study item-nonresponse turned out to decrease. Presumably, capable and motivated respondents remain in the dataset. (b) What differences between item-nonresponses were produced by the mode of data collection? According to our expectations, we found that face-to-face interviews have lower item-nonresponses than do mailed questionnaires. (c) Does item-nonresponse depend on respondent characteristics? We expected that gender, age, education and health of respondents would influence item-nonresponse rates. In line with the literature, we found that item-nonresponse increases with age, and that women have higher item-nonresponse rates. Our analysis revealed that indeed gender and age do indeed appear to contribute to item-nonresponse, but that the introduction of physical health and depression variables shows that item-nonresponse can be explained by these variables; education, urban environment, and subjective health have no effect. This outcome means that the survey unfamiliarity hypothesis of Colsher and Wallace (1989) has not been supported. It would seem to have been replaced by the physical health and depression hypothesis as a cause of item-nonresponse. Moreover, considerable differences in item-nonresponse trajectories in time have been assessed. (d) Does item-nonresponse depend on the interviewer? We did not assess differences between them. (e) Are there differences according to item wording? The outcome shows that the item-nonresponse rates are lower for the positively worded items than that they are for the negatively worded items.

In this paper we proposed a procedure for the assessment of the effects of time and of respondent and interviewer characteristics on item-nonresponse in a longitudinal design. The paper showed that the

nesting of occasions under respondents, and the nesting of respondents under interviewers necessitates a multilevel model. Since item-nonresponse is a dichotomous variable, a multilevel logistic regression analysis has to be used. Hence, this paper offers a general description of an adequate model for studying these effects as they originate from variously leveled sources. This paper's illustration also showed the existence of time effects of item-nonresponse in a longitudinal study. It showed the existence of mode effects, in so far as an interview produces lower item-nonresponse rates than does a mail questionnaire. The respondents' age effects are not replicated, but it turned out that their physical difficulties in daily life, depression, and pain experience are the main determinants.

Although the percentages of item-nonresponse are generally low, it is worth to explaining them, although the process is laborious. Paying attention to the details of data collection procedures is the first step towards increasing the quality of survey data.

7. REFERENCES

Beekman, A.T.F, Van Limbeek, J., Deeg, D.J.H., Wouters, L. and Van Tilburg, W. (1994). Een screeninginstrument voor depressie bij ouderen in de algemene bevolking: de bruikbaarheid van de Center For Epidemiologic Studies Depression Scale (Ces-D). (A screening instrument for depression among elderly people in the general population: the utility of the CES-D scale). *Tijdschrift voor Gerontologie en Geriatrie.*

Berk, M.L and Bernstein, A.B. (1988). Interviewer characteristics and performance on a complex health survey. *Social Science Research. 17.* 239-251.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (Eds.). (1991). *Measurement Errors in Surveys.* New York: John Wiley and Sons.

Collins, M. (1980). Interviewer variability: A review of the problem. *Journal of the Market Research Society,* 77-95.

Colsher, P.L. and Wallace, R.B. (1989). Data quality and age: health and psychobehavioral correlates of item nonresponse and inconsistent responses. *Journal of Gerontology, 44,* 45-52.

Craig, C.S. and McCann, J.M. (1978). Item nonresponse in mail surveys: extent and correlates. *Journal of Marketing Research, 15,* 285-289.

Deeg, D.J.H, Knipscheer, C.P.M. and Van Tilburg, W. (1993). *Autonomy and Well-being in the Aging Population: Concepts and Design of the Longitudinal Aging Study Amsterdam.* Bunnik: NIG.

De Leeuw, E. (1992). *Data Quality in Mail, Telephone, and Face to Face Surveys.* Amsterdam, T-Publications.

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

Francis, J.D. and Bush, L. (1975). What we know about "I do'nt knows". *Public Opinion Quarterly, 34,* 207-218.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika, 78,* 45-51.

Groves, R.M. (1989). *Survey Error and Survey Cost*. New York etc: John Wiley and Sons.

Groves, R.M. and Fultz, R.L. (1979). Gender effects among telephone interviewers in a survey of economic attitudes. *Sociological Methods and Research, 14*, 31-52.

Hertzog, C, Van Alstine, J., Usala, P.D., Hultzch, D.F. and Dixon, R. (1990). Measurement proportions of the CES-D scale in older populations. *Psychological Assessment, 2*, 64-72.

Hox, J.J. De Leeuw, E.D. and Kreft, I.G.G (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In: P.P. Biemer, R.M Groves, L. E. Lyburg, N.A. Mathiowetz, and S. Sudman (eds). *Measurement Errors in Surveys*. New York: Wiley, 439-462.

Leigh, J.H. and Martin, Cl.R. (1987). "Don't know" item nonresponse in a telephone survey: Effects of question form and respondent characteristics. *Journal of Marketing Research. 24*, 418-424.

Little, D.B. and Rubin, R.J.A. (1987). *Statistical Analysis with Missing Data*. New York, etc. Wiley.
Madow, W.G., Nisselson, H. and Olkin, I. (1983). *Incomplete Data in Sample Surveys, Vol 1: Report and Case Studies*. New York: Academic Press.

Molenaar, N.J. (1988). *Formuleringseffecten in Survey-interviews: een Non-experimenteel Onderzoek*. (Question Wording in Survey-Interviews). Amsterdam: Vrije Universiteit.

Radloff, L.S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 3*, 385-401.

Rogers, Th.F. (1976). Interviews by telephone and in person: Quality of responses and field performance. *Public Opinion Quarterly, 42*, 407-410.

Singh, A.C. and Merkouris, P. (1996). A diversity component analysis of interviewer effects in categorical survey data. In: SMPQ (eds.) *Proceedings of the International Conference on Suvey Measurement and Data Quality*. American Statistical Association, Alberta, Vi. 332-337.

Sudman, S., Bradburn, N.M., Blair, E. and Stocking, C. (1977). Model expectation: The effect of interviewer's prior expectations on responses. *Sociological Methods and Research, 6*, 171-182.

Viswesvaran, Ch., Barrick, M.R. and Ones, D.S. (1993). How definitive are conclusions based on survey data: Estimating robustness to nonresponse. *Personel Psychology, 46*, 551-567.

Woodhouse, G. (eds). (1995). *A Guide to Mln for New Users*. London, Multilevels Projects, Institute of Education, University of London.

Ying, Y.W. (1989). Nonresponse to the Center for Epidemological Studies Depression scale in Chinese Americans. *International Journal of Social Psychiatry, 35*, 156-163.

APPENDIX 1: ITEMS OF THE CES-D DEPRESSION SCALE

Please indicate how often you have felt this way during the past week (0. rarely or never 1. some of the time/2. occasionally/3. mostly or always)

1. During the past week I was bothered by things that usually don't bother me.
2. During the past week I did not feel like eating, my appetite was poor.
3. During the past week I felt that I could not shake off the blues even with help from my family and friends.
4. During the past week I felt that I was just as good as other people.
5. During the past week I had trouble keeping my mind on what I was doing.
6. During the past week I felt depressed.
7. During the past week I felt that everything I did was an effort.
8. During the past week I felt hopeful about the future.
9. During the past week I thought my life had been a failure.
10. During the past week I felt fearful.
11. During the past week my sleep was restless.
12. During the past week I was happy.
13. During the past week I talked less than usual.
14. During the past week I felt lonely.
15. During the past week people were unfriendly.
16. During the past week I enjoyed life.
17. During the past week I had crying spells.
18. During the past week I felt sad.
19. During the past week I felt that people dislike me.
20. During the past week I could not get "going".

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

ASSESSING NONSAMPLING ERRORS IN SURVEY DATA THROUGH RANDOM INTERCEPT MODELS

Dale Atkinson [1]

ABSTRACT

Variability in reporting from survey to survey by individual survey respondents can result for many different reasons. Indicated changes in reported inventories reflect real change confounded with nonsampling error. The objectives of this study were 1) to identify systematic components in survey to survey variability that could indicate nonsampling error, 2) to quantify the potential nonsampling error underlying these components, and 3) to target areas where corrective measures may be taken to reduce the problem.

Random intercept modeling was used to quantify the percentage of variability explained by differential effects of enumerator assignment and respondent change from one survey to the next. Under certain assumptions, the modeling provides estimates of standard survey quality measures such as reliability. When enumerator assignment is incorporated in the model, the approach can be used to identify individual assignments which contribute inordinately to this variability. Therefore, the approach provides a tool to target potential enumeration problems where additional concept training might be needed. Considerable information about the data collection process can be obtained directly from the survey data without additional data collection requirements.

This paper discusses the random intercept models used, results of applying the approach to recent data collected by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture, and the potential for incorporating this survey quality monitoring tool as an integral part of the NASS survey management process.

KEY WORDS: Sample survey; Nonsampling error; Reporting variability; Random intercept models.

1. INTRODUCTION

1.1 Description of the Problem

Despite our best efforts to avoid them, nonsampling errors occur in survey collected data. These are of many types and affect the data at various stages of the survey process. They result from actions before, during and after the interview. This paper will address errors during the interview, and focus on whether a systematic (and hopefully correctable) component of reporting variability can be identified.

Errors at interview time can be attributable to the enumerator, the respondent or an interaction of the two. One potential source of quarter-to-quarter reporting differences is change in respondent. Another is differential interviewer effect on the response. This study examined both of these possibilities to determine whether either had a significant impact on the collected data.

2. BACKGROUND

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture (USDA) conducts an ongoing series of quarterly sample surveys from which it estimates inventory and production of various agricultural commodities. The quarterly surveys conducted in June, September, December and March have a multiple frame design consisting of both list and area frame samples. In June, sampled area segments averaging about one square

[1]Dale Atkinson, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, 3251 Old Lee Hwy, Room 305, Fairfax, Va., U.S.A., 22030.

mile in size are completely enumerated to record all agricultural activity within their boundaries. The individual land operating arrangements (tracts) within the sampled segments are the units for which survey data are collected. The data from these tracts provide full area frame estimates and also the area component of multiple frame indications. The multiple frame indications include area data only for area tracts with no chance of list selection. These are referred to as non-overlap (NOL) tracts.

The NOL tracts identified in June are subsampled to account for list incompleteness in multiple frame indications from the follow-on quarters of September, December and March, when no full area enumeration occurs. Since for many commodities NASS' list frames are fairly complete, there are relatively few NOL tracts. For example, the numbers of NOL tracts in June 1995 varied by State from about 10 to about 300. However, in spite of their small numbers, the NOL tracts account for a large amount of the variability in multiple frame indications. Consequently, the NOL tracts are sampled at very high rates in the follow-on quarters, resulting in repeated measures on some survey items for most NOL tracts.

Various analyses of the NASS quarterly survey data over the past few years have indicated a substantial amount of quarter-to-quarter reporting variability for individual sample units in survey items that should be fairly stable within a survey year. "True value" reinterview surveys have been conducted by NASS to address similar reporting concerns in the past. However, these create additional respondent burden, which is especially troublesome with the already heavily burdened NOL operations. A preferable way (if it works) to assess the data quality and to identify areas needing improvement is to glean as much survey quality information as possible from the regular survey contact. This type of approach was explored in this study.

3. METHOD

3.1 General

In many studies of nonsampling errors, either an administrative source or a reinterview survey provides the "true values" by which the quality of survey data is evaluated. In this study neither is available. However, NASS invests very heavily in the quality of data collected in its June Agricultural Survey, and a case can be made that these data may be of better quality than data collected in subsequent quarters. In essence, the June data possess many of the attributes associated with "true value" reinterview data. They are collected through personal enumeration (thought to be the best mode of data collection) by the best trained and most agriculturally experienced enumerators available to NASS. By comparison, most data collected in the follow-on quarters are by telephone from one of NASS' State Statistical Offices (SSOs). The telephone enumerators used are generally hired locally and often have less experience in agricultural surveys than the field enumerators used in June.

Obviously both the June and follow-on quarter data contain errors. However, if the June data contain fewer errors and can be viewed as a reasonable proxy to the truth, then data quality measures such as reliability and indices of inconsistency can be estimated directly from the quarterly data. The random intercept regression approach used in this study also yields estimates of intra-group correlation and the effects of various levels of a grouping variable (e.g., follow-on quarter enumerator assignment or whether or not a respondent change has occurred between quarters). Percentages of model variability (in predicting a follow-on quarter's response with the June response) that were attributable to differential group effects were calculated.

3.2 The Model

Biemer and Atkinson (1995 (1 & 2)) discussed an approach by which measures of data quality and group effects for arbitrary grouping variables could be obtained from two-phase samples, where the second phase sample was selected for reinterview with reconciliation to obtain true values. This paper is an attempt to apply the approach to the situation where reinterview data are not available, but where independent repeated measures on sample units are available through on-going quarterly surveys. In the present case, one response (June's) is expected to be generally superior to the other and to represent a reasonable "proxy to the truth." The underlying model development is described in great detail in the previous references and will be discussed much less rigorously here. In general we fit a model of the form:

$$y_i = \gamma_0 + \gamma\mu_i + z_{gi} \qquad (3.2.1)$$

where μ_i is the June value of the item, γ_0 and γ are constants, and z_{g_i} is a random error term. Insofar as μ_i is the "true" value for the item of interest, the parameter γ_0 may be interpreted as a constant or absolute bias that is added to all observations, while γ is a "proportional" bias. As an example, suppose μ_i is some measure of farm size (e.g., all land in farm or total acres of cropland). The magnitude of the error in y_i is often proportional to size and is therefore appropriately modeled by $\gamma\mu_i$. The term z_i is the sum of two random components, d_g and δ_i where d_g is the "bias" or "group effect" associated with group g, and δ_i is an independent unit-level error. We assume that $d_g \sim (0, \sigma_d^2)$ and $\delta_i \sim (0, \sigma_\delta^2 \mu_i^\lambda)$ where λ is a known constant. In this study a value of 0 was used for λ; however, it is possible to estimate λ from the data (see, for example, Wright, 1983).

With the above model, further assume the conditional covariance of the errors for $i \in G_g$ is given by

$$\begin{aligned} \text{Cov}\,(z_{gi}, z_{g'i'} \mid i) &= \sigma_d^2 + \sigma_\delta^2 \mu_i^\lambda \quad \text{for } i = i' \\ &= \sigma_d^2 \quad \text{for } i \neq i';\ g' = g \\ &= 0 \quad \text{for } g' \neq g \end{aligned}$$

Let $E(\cdot)$ denote the conditional expectation given the unit i over the measurement error distribution and $\text{Var}(\cdot)$ denote the unconditional variance with respect to the sampling distribution. If we assume that all the G_g, $g = 1,\ldots,J$ are of equal size (say m) and that the finite population correction is ignorable, then Biemer and Stokes (1991) show that $n\text{Var}(\bar{y}) = R^{-1}\text{Var}(\gamma\mu_i)[1 + (m-1)\rho_y]$ where R, referred to as the reliability ratio, is

$$R = \frac{\text{VarE}(y_i \mid i)}{\text{Var}(y_i)} = \frac{\gamma^2 \sigma_\mu^2}{\sigma_y^2}, \qquad (3.2.2)$$

and (when a grouping variable of enumerator assignment is used) ρ_y, referred to as the *intra-enumerator correlation coefficient*, is the correlation between pairs of units within an enumerator's assignment.

The reliability ratio, R, is the ratio of the variance of the "true" value for the data item -- *viz.*, $\text{Var}(\gamma_0 + \gamma\mu_i)$ -- to the variance of the observation y_i. Estimation of R usually requires repeated measurements obtained under identical survey conditions and such that the measurement errors associated with each measurement are independent (between measurements) and identically distributed (see Biemer and Stokes, 1991). These assumptions are perhaps best satisfied with a well-designed and executed reinterview survey, but with the present approach R is estimated directly from the quarterly survey data.

Under model (3.2.1) the intra-enumerator correlation coefficient, ρ_y, is given by

$$\rho_y = \frac{\sigma_d^2}{\sigma_y^2}. \qquad (3.2.3)$$

It is widely used in measurement error studies to describe the degree to which the quality of interviewing varies by enumerator (see for example, Groves, 1989).

In this study, ρ_y was estimated by defining the grouping variable to be the follow-on survey enumerator assignment. A large estimate of ρ_y indicates that large enumerator effects (d_i) are present in the data, and an analysis of the large absolute values of d_i can help identify which enumerator assignments are contributing the most to the enumerator variance. This paper presents estimates of R and ρ_y, as well as a distribution of the standardized d

associated with the enumerators for the 1995-96 follow-on surveys.

3.3 The Data Analyzed

To explore the usefulness of this approach in studying quarter-to-quarter reporting variability, a four-quarter survey data set covering June 1995, September 1995, December 1995 and March 1996 was constructed. For the purposes of this study it was necessary to eliminate as thoroughly as possible "real" quarter-to-quarter inventory changes, since the success of the approach for monitoring data quality is predicated on the assumption that data differences between June and the follow-on quarters are indicative of measurement error. Insofar as this is true, the approach can be used to help identify and quantify sources of the measurement error.

Several steps were taken to minimize the confounding effects of real change. First, the items analyzed were limited to all land in farm and cropland, items which are less likely than others to legitimately change during the course of the survey year. Also, records with a change in reporting unit, indicative of an operation change, were eliminated. Finally, since all four quarters of data were analyzed together (to increase usable sample size, thus improving the models), a further step was taken to limit the influence of tracts for which consistent results were obtained across all follow-on quarters after an initial inconsistency with the June report. Consistent follow-on quarter results that differ from the June report could suggest a problem in June. To address this concern, only the earlier (earliest) follow-on quarter's report was retained in the modeling data set in cases where identical acreages were obtained in consecutive follow-on quarters.

Separate analysis data sets were created for all land in farm and cropland, based on usability for these items. Records in the data sets represented a usable follow-on quarter's response with an associated usable June response. For a record to be included, the analysis item had to have been reported (not estimated or imputed) in both June and the follow-on quarter. The resulting four-quarter data sets with all States' data contained about 7,100 records for all land in farm and about 7,000 records for cropland.

3.4 Outlier and Leverage Point Elimination

Since the current approach utilizes least squares regression modeling, the handling of influential data is very important. Outliers create a special quandary in this application. Since we're trying to identify problematic situations (and the extent to which these can be attributed to differential effects of the various levels of the grouping variable) we're reluctant to throw them out. However, if we don't eliminate some outliers, the basic regression parameters (and as a result the group effects) will be poorly estimated. Identifying a procedure for eliminating the proper level of influential data points is a challenging endeavor -- one that considerable time was spent on in this study, and one for which more work should probably be done.

Since little literature (and even less supporting software) was found on approaches for handling influential data in mixed models, the elimination of leverage points for this application was approached by temporarily ignoring the random group effects and using linear regression methodology. After the leverage points identified through linear regression techniques were eliminated, the resulting data set was used in the random intercept (mixed) modeling.

Specifically, records whose "hat value" (from the linear regression of the follow-on quarter value on the June value) exceeded 3*p/n=6/n (where p is the number of parameters in the model and n is the usable sample size) were eliminated as leverage points. While 2*p/n is often used as a rule of thumb for reviewing observations, the 3*p/n threshold seemed to work better for an automated screening. The automated approach is necessary since one of the goals of this study is to implement a batch-run quality monitoring tool.

After the leverage points were removed, the following two candidate procedures were tested for use in eliminating outlier data:

1.) Iteratively refit the model and on each of five passes eliminate observations that account for more than a constant, c, times the average observation contribution to the residual sum of squares (RSS). That is, observations whose squared residual $> c * RSS / n$.

2.) Noniteratively remove observations meeting one or more of the following Belsey, Kuh, and Welsh (1980)

linear regression criteria:

a.) Externally Studentized residual > 2
b.) |Covariance ratio| $\geq$ 6/n
c.) DFFITS > 2*sqrt(2/n).

Procedure 1 seemed to perform somewhat better than procedure 2, though both retained some observations that subjectively might have been eliminated and eliminated some observations that subjectively might have been retained. Procedure 2 tended to eliminate more observations (overall about 10 percent of them), even run noniteratively. This may be due to a "swamping" effect. Procedure 1 seemed to make relatively good decisions (based on a subjective evaluation), especially with larger values of the constant c. The iterative nature of procedure 1 also seemed to help overcome outlier masking, as some observations that were not influential in the presence of other outliers, showed up as overly influential in a later iteration.

There may be some opportunity for meaningful refinement in these procedures by using outlier identification techniques specifically suited for identifying outliers in multiple outlier situations, such as those discussed in Swallow and Kianifard (1996).

For the present, all results presented in this paper are based on outlier elimination procedure 1, with a value of c=20. Subjectively, this value of the scale parameter seemed to give the best results of several that were tried. About 8 percent of the all land in farm records and about 7 percent of the cropland records were discarded as outliers or leverage points in the final analysis.

3.5 The Model Selection Process

Using the model specification in 3.2.1 separate models were generated using two basic grouping variables -- follow-on quarter enumerator assignment and an indicator variable for whether or not there was a respondent change between June and the follow-on survey. A refinement to the basic model to adjust for the differential size of operations in the various levels of the grouping variable (i.e., average size of operation in an enumerator's assignment) was considered. However, this refinement didn't improve the modeling since the variable calculated to capture this information -- average June acreage (all land in farm or cropland) for a group -- was both statistically and practically insignificant in virtually all the models. The overall proportional adjustment, γ, appeared to be sufficient to account for differences in the average size of operations in each group. This was the case, regardless of whether the grouping variable used was enumerator assignment or respondent change indicator.

Another modeling possibility, including random effects for both grouping variables (and their interaction) in the same model, was explored. This approach did not work adequately, however, since crossing enumerator assignment with respondent change indicator resulted in too many empty or sparse cells. In general, there are relatively few respondent changes from June to a follow-on quarter. In the analyzed data sets this occurred only about 25 percent of the time. As a result, crossing this variable with another one was not a viable modeling alternative.

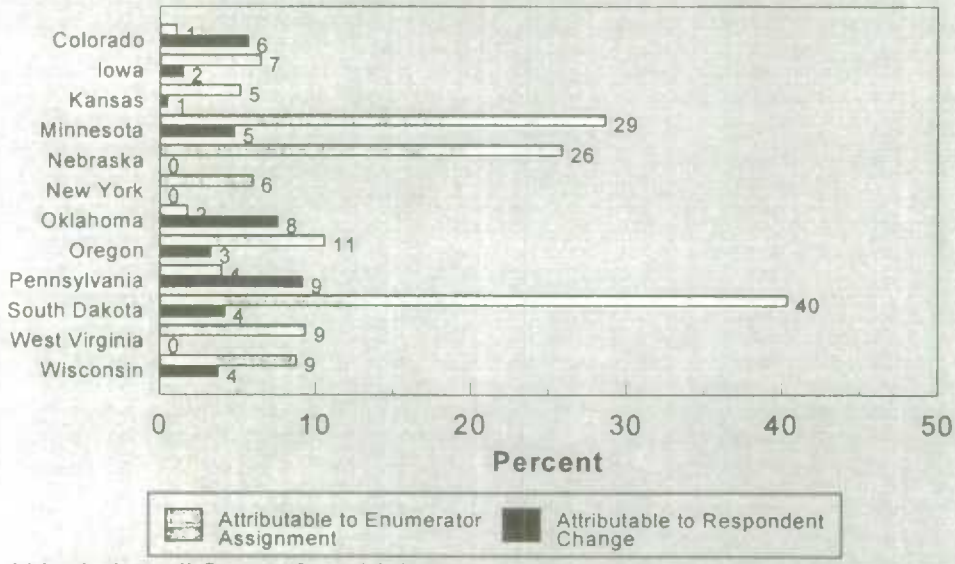
Therefore, our selected models were marginal ones with the two candidate grouping variables incorporated separately. To eliminate any confounding effect on the models, all records with a respondent change were excluded when models using enumerator assignment as the grouping variable were fit.

Finally, to better reflect State-to-State differences in the modeling, the models were created at the State level.

3.6 Model Results

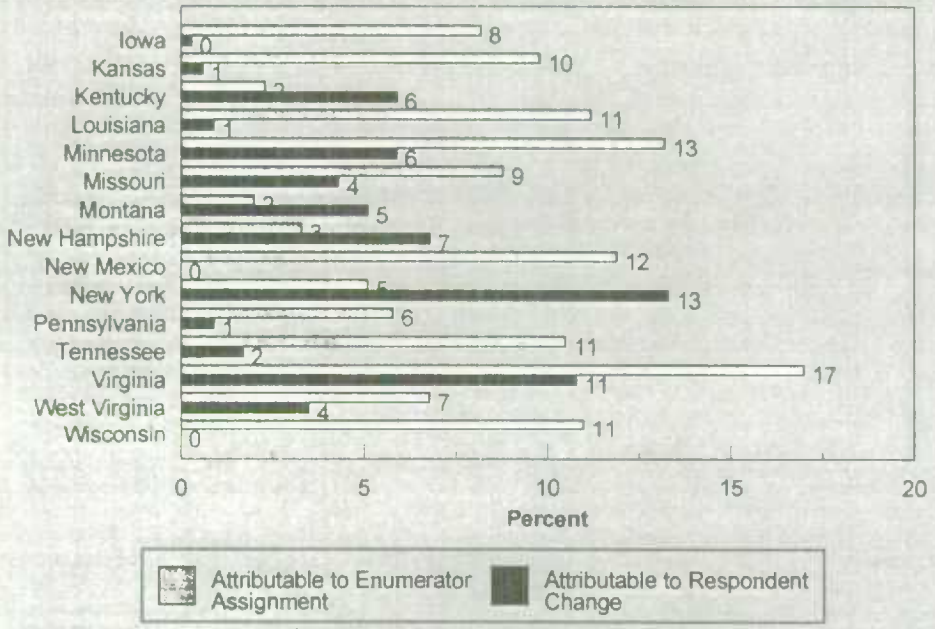
Somewhat surprisingly, although again resulting largely from the rarity of the event, respondent change generally had a small effect. In most States enumerator assignment effect was larger, and in some States it was substantial. Figures 1 and 2 compare the percentage of model variability attributable to enumerator assignment vs. respondent change for all land in farm and cropland, respectively. Some caution is needed in interpreting the percentages in Figures 1 and 2, since in a few cases (most notably Minnesota and Nebraska for all land in farm) the large percentages of variability accounted for were on very small bases. Some States showed substantial overall model variability while other States showed very little.

Figure 1: Estimated Impact of Enumerator Assignment and Respondent Change on Quarter to Quarter Variability for the 1995 Crop Year -- All Land 1/



1/ Includes all States for which enumerator assignment or respondent change accounted for at least 5 percent of the indicated variability.

Figure 2: Estimated Impact of Enumerator Assignment and Respondent Change on Quarter to Quarter Variability for the 1995 Crop Year -- Cropland 1/



1/ Includes all States for which enumerator assignment or respondent change accounted for at least 5 percent of the indicated variability.

Figure 3 shows the distribution of estimated enumerator effects for cropland, standardized to reflect within-State variability. The tails of distributions like this indicate enumerator assignment effects that are abnormally large and can help identify where problems may exist.

Figure 3: The Distribution of Standardized Enumerator Effects for Cropland

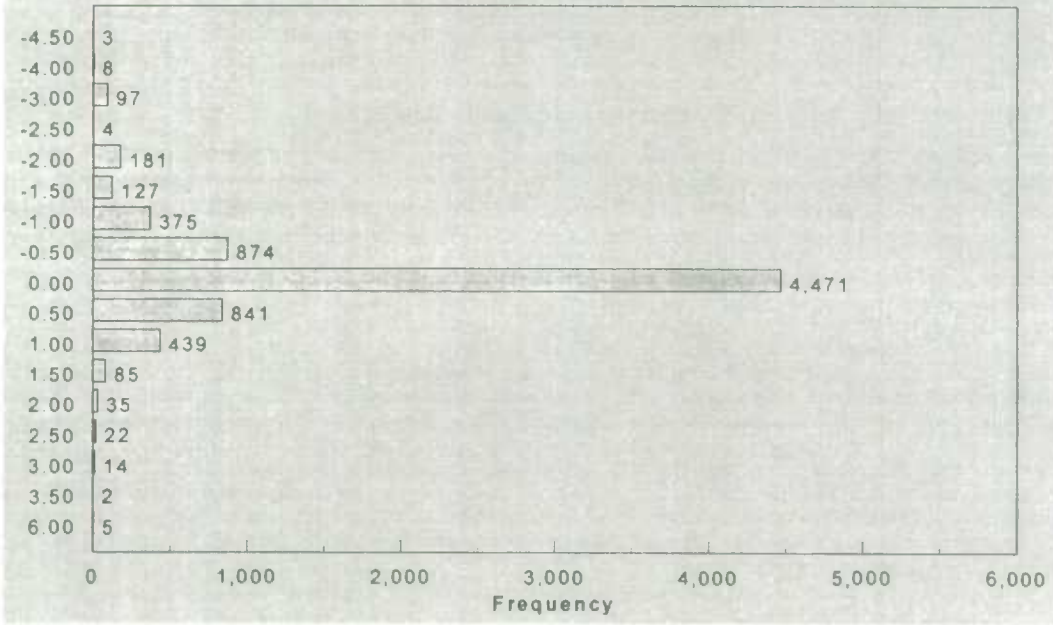


Table 1 shows our estimates of data quality at the U.S. level for all land in farm and cropland , both including and excluding outliers. From the "excluding outlier" analysis, cropland appears to be slightly less reliable, with higher intra-enumerator correlation than all land in farm. Based on the average enumerator assignment size in the quarterly surveys analyzed in this study (m=17), an intra-enumerator correlation of .008 could be expected to cause an increase in variance in an estimate of about (17-1)(.008)=.13, or 13% (using equation 3.2.2). Estimates of the intra-enumerator correlation varied widely by State, so in some States a larger variance inflation could be expected.

Finally, notice the large impact of outliers on the estimates of data quality. Clearly the 7-8 percent of data screened in this application strongly affected the estimates of both intra-enumerator correlation and reliability. Standard errors shown in Table 1 were calculated through bootstrapping.

Table 1. Estimates of the Intra-Enumerator Correlation Coefficient and Reliability

Item	Screening Scenario	$\hat{\rho}_y$ (*s.e.*)	$\hat{R}$ (*s.e.*)
All Land in Farm	Excluding Outliers	0.004 (.0017)	.91 (.005)
	Including Outliers	0.009 (.0043)	.76 (.018)
Cropland Acres	Excluding Outliers	0.008 (.0027)	.88 (.007)
	Including Outliers	0.010 (.0042)	.77 (.017)

3.7 Concluding Remarks

This paper documents an attempt to mine existing data to obtain an indication of the relative quality of items collected in an on-going survey program, to assess the relative impact of sources of nonsampling error and to provide a tool to help target areas where additional training may be needed.

Like most situations where data are put to use in a way for which they're not specifically designed, the validity of some of the underlying assumptions may be questionable. In particular, the assumption of the June value representing truth, on which the interpretation of our estimates of reliability and intra-enumerator correlation are largely predicated, can be debated. Also, the assumption of equal group sizes does not strictly hold in our normal survey enumeration.

In spite of these weaknesses, however, the process does provide useful information. Estimates of survey quality comparable to those previously produced at much higher cost through reinterview surveys were obtained, both with and without outlier removal. Which are the better estimates and how outlier screening should be performed (if indeed it should be performed at all) are open issues. However, the fact that 7-8 percent of the data had such a significant impact on the calculated estimates indicates the presence of very influential observations that should be checked for validity. The large impact of these observations also raises the issue of whether a more robust model with no outlier screening would provide even better results. This option will be explored in a later study.

Because of unequal size groups and a confounding of errors, the enumerator assignments identified as problematic were not always indicative of poor enumeration at all, but sometimes a combination of a small assignment size and a serious key-entry error. Whatever the reason for their identification, however, the groups with large, absolute group effects were generally ones that should be examined.

Finally, the computer program used in this analysis was designed with the idea of its potential implementation as an on-going tool to help statisticians in our SSOs manage the data collection process. It produces review listings of all outlier and leverage point samples eliminated in the modeling, and of all samples in enumerator assignments whose effects were statistically significant. The results of this study indicate that it may have potential for operational use.

4. REFERENCES

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). "Regression Diagnostics", New York: John Wiley & Sons, Inc.

Biemer, P. and Atkinson, D. (1995). "An Integrated Approach for Estimating Measurement Error Bias and Variance in Two-Phase Samples," *Proceedings of the 1995 Annual Research Conference*, 355-357.

Biemer, P. and Atkinson, D. (1995). "Estimating Measurement Error Bias and Variance in Two-Phase Samples," American Statistical Association - *Proceedings of Section on Survey Research Methods*, Volume II, 775-780.

Biemer, P. and S.L. Stokes (1991). "Approaches to the Modeling of Measurement Errors." In P.P. Biemer, R.M. Groves, L.E Lyberg, N.A. Mathiowetz, S. Sudman (Eds.) *Measurement Errors in Surveys*. New York: John Wiley & Sons, 487-516.

Groves, R.M. (1989). "Survey Errors and Survey Costs", John Wiley & Sons, N.Y.

Swallow, W. and Kianifard, F. (1996). "Using Robust Scale Estimates in Detecting Multiple Outliers in Linear Regression," *Biometrics*, **52**, 545-555.

Wright, R.L. (1983). "Finite Population Sampling with Multivariate Auxiliary Information," *Journal of the American Statistical Association*, **78**, 879-884.

SESSION 8

ADMINISTRATIVE DATA

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

SOME DATA QUALITY IMPACTS WHEN MERGING SURVEY DATA ON INCOME WITH TAX DATA

Sylvie Michaud and Michel Latouche[1]

ABSTRACT

Generally, measurement error causes some problems. A lot of work has been done in attempting to measure it and compensate for it. It has been shown that measurement error can create more problems in a longitudinal survey, especially if the data are used in regressions. The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey that attempts to measure the impact of changes in labour market activities and family characteristics on income. To try to reduce response burden and improve data quality, the survey has offered a choice to respondents: either respond to the income survey or give permission to SLID to use their administrative records. This paper aims to quantify the impacts of this mixed approach on the response error, especially on the measures of change.

KEY WORDS: Longitudinal data; Response error; Underreporting; Response rate; Trends.

1. INTRODUCTION

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey that measures the impact of changes in labour market activities and/or family circumstances on income. People in a given panel stay in the sample for six years and are interviewed twice a year. In January, labour information is obtained while the May interview collects income data. The income interview is done in May because Canadians file their income tax return by the end of April and it is generally felt that people are in a better position to provide accurate income data around that time. In May 1995, in an effort to reduce response burden, respondents in SLID were offered the choice between responding to the income survey and allowing SLID to get the income information from their tax return. This question is re-asked every year and after three years of collection, more than 75% of the respondents use the second option. However, the integration of survey data and tax data is not without problems. Definitions are not always compatible and there are linkage problems. This is balanced against quality issues usually found with income surveys (such as under-reporting of certain income sources) and the need for imputation. This paper gives an overview of the different sources of errors that occur with this methodology and presents some results on the impact of this mixed approach. The research has focused on micro-comparisons and has attempted to quantify the impact on measures of change.

2. SLID'S SAMPLE DESIGN

The SLID sample is selected using a multi-stage sample design. Respondents selected in the sample have already been part of the Canadian Labour Force Survey (LFS) for six months before being selected to participate in SLID. They are then interviewed twice a year for six years. A first interview in January asks detailed labour information. It also records changes in family composition and dates of the changes. The second interview in May collects detailed income information, according to 24 categories. The survey also collects income tax paid to determine after-tax income. Income is collected for each individual in the household aged 16 years and older. It is aggregated at the family level to determine low-income measures.

Collection of income information is not a new process within Statistics Canada. The Survey of Consumer

[1]S. Michaud and M. Latouche, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

Finances (SCF) has been collecting annual income data for the last thirty years, and SLID's income questions are identical to those used by SCF. This experience will greatly aid SLID.

In general, income surveys suffer from lower response rates than many other surveys. While a non-income survey such as the Canadian Labour Force Survey has a usual response rate of 95%, the Survey of Consumer Finances has a response rate of 80% . The SLID response rate for the income interview is 76%. The data have also been linked to other sources for data quality evaluations. Based on these comparisons, there is under-reporting of certain income sources such as unemployment insurance benefits, social assistance and interest and dividends[5].

Tax data have also been used more recently as a source of income information. In particular the Longitudinal Administrative Data (LAD) is a longitudinal file based on tax data[7]. A 10% sample of survey respondents has been randomly selected and families are reconstructed, based on the information provided on the tax return (spouses and children are created based on other fields from the tax form). LAD does not agree perfectly with the other administrative sources. Only census families (father-mother-children) can be constructed and there is a tendency to overestimate families of one person. There is also an under representation of certain age groups (particularly older persons) and of low income families. Recently however, with the implementation of tax credits, the population coverage of the universe by the tax system has improved. The problem of constructing families still remains. On the other hand, the quality of income data from tax sources is felt to be superior to that from a survey.

SLID uses the mixed approach to try to maximize response rates and data quality. There are however issues with such an approach.

3. THEORETICAL SOURCES OF ERRORS USING SURVEY DATA AND USING TAX DATA

There are a number of issues with the principle of using administrative data for income sources. For example timeliness of the data may have an impact on the survey's target dates. The link to the administrative files with or without unique identifier may also raise some problems, and if one tries to assess the overall impact of a mixed strategy, this should be included. However, in this case the discussion will be restricted to the issues of combining both of the sources from a data quality point of view. A general discussion of the use of the tax file in SLID can be found in [1].

To provide a global measure of quality, surveys should permit calculation of mean square errors; that is the sum of the variance of a given variable and the square of the bias. This is usually hard to do because the bias can be very difficult to measure. Table 1 attempts to identify the potential advantages or drawbacks of each of collection methods (survey data and tax data) and tries to specify on which component of error it may have an impact.

Coverage is affected only by the use of tax data. Tax file coverage has improved over the years, now covering 94% of the population aged 20 and over. If the population of non filers is different from the population of filers, this could create some bias in the data. Students, for example, are a group that is likely going to be under represented amongst filers. Since they are usually associated with lower income, and that the filer students may be different from the non filer students, bias can be present.

Table 1. Comparison of the survey collection method vs using the tax data.

	Survey only	**Tax only**
coverage of population (bias)		↘ filers only
response rate (total)	↘ sensitivity response burden tracing	↗ all filers are "respondents"
(variance) (bias)		↘ not linked or wrongly linked
response error (bias)	↘ ↘ under-reporting of certain income sources (UI, interest...)	↘ under-reporting of certain income sources (underground economy)
	↘ rounding proxy reporting	
		↘ non taxable sources
time series consistency (bias)	↘ response error & longitudinal inconsistencies	
		↘ potential inconsistencies in definitions of income categories

↗ suggests an improvement ↘ suggests a disadvantage

Non-response can create problems both in terms of variance and bias. Income is a sensitive topic for some respondents, and it tends to have a "lower" response rate when it is collected by a survey. The fact that SLID is a longitudinal survey also impacts on the response rate; people move through the years and the inability to trace a person also decreases the response rate. The extent to which these non-respondents are different from the respondents will determine the magnitude of bias. The use of tax data should compensate for some of these problems in theory; as long as a person is a filer, it should be possible to locate their record on the tax file and this should increase the response rate. However, SLID does not collect the Social Insurance Number which is the unique link to the tax file. Other fields, which are described in Section 4, are used in a statistical matching procedure to link people in the SLID sample to the tax file of individuals. Some data quality control measures are done to improve the quality of the linkage but there is always the possibility of having a wrong linkage or that a person is not linked even if he/she is a filer. This also decreases the response rate.

Response error has been studied for certain income variables for which there existed an external source in order to validate the results and assess potential biases. Some studies have suggested that there is an under reporting of certain income sources when data are collected from a survey. SCF captures approximately 80% of UI

benefits compared to 94% in the tax system. Investment income is also prone to under-reporting. This creates bias in the results. In addition, there is a general feeling that tax data also suffer from some under-reporting of certain income sources that are related to the underground economy. However, because SLID asks respondents to consult their tax form to provide their income information, and because it is not clear that respondents will actually declare those kinds of income source through a survey, one could conclude that tax data may also be prone to bias, but not as severely as survey data.

A second source of response error is due to the rounding of income amounts reported in a survey. Rounding of a reported value for total income is problematic, but it is worse if rounding is done on the different sources of income since total income is derived as the sum of these components.

A third source of response error, affecting tax data, is the non availability of certain income sources on the tax form. Even though some non taxable income are reported on the tax file, it is limited to those sources that need to be reported for calculation of tax credits. Items such as lottery gains and inheritances are not reported, but are collected by the survey.

A fourth source of response error may arise because of changes in definitions and concepts in the tax environment. Time series may be affected because of changes in income tax regulations.

As can be seen, there are issues with both sources of data. The study wanted to see the impact of SLID's mixed collection on data quality. Because of the longitudinal nature of SLID, measures of change are important. Response errors create more problems in a longitudinal survey compared to a cross-sectional survey, since it is usually expected that correlation between the repeated measures will be larger than the correlation between the response errors. Because of the potential rounding and under reporting error, it is expected that income from administrative sources will be less prone to response error than survey data.

In particular, assume a person wanted to measure a variable X (income), but what is really measured is $x = X + u$, where u is the response error. In a regression, where one would like to predict: $Y = X\beta + \epsilon$,
what is really measured is: $y = x\beta' + \epsilon$

where β' is biased towards zero, under the regular assumptions of the independence and the normality of the errors. If one was interested in doing a regression on the measures of change: $\Delta Y = Y_{t+1} - Y_t$, it has been shown [1] that the measure of bias in the equation of change is bigger than that on the measure of level. Mathematically,

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < \frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2 + \sigma_{\Delta u}^2}$$

4. EMPIRICAL IDENTIFICATION OF THE SOURCES OF ERRORS

For its first year of data collection, SLID did not make use of the tax route and collected income data directly from the respondents. As a quality assurance activity, a comparison was performed between survey and tax data. This allowed an identification and quantification of error sources.

Tax data are obtained though matching but since the Social Insurance Number (SIN) is not asked of respondents, matching is done through a statistical procedure. Records were first linked though a direct match on name, postal code, data of birth, sex and marital status. This procedure linked 50% of the records. Records not matched were then run through a statistical matching process (allowing for missing values or discrepancy for one or more of the matching fields). This lead to an overall linkage of 85%. The study concentrated on response rate, coverage, linkage, and response errors. Special attention was devoted to the impact on yearly trends.

4.1 Response rates and potential biases

The SLID sample file was linked to the 1993 tax file using direct and statistical match approaches. Table 2 presents the distribution of the sample, by response status to the income interview and by the outcome of the linkage to tax data.

Table 2. Response status by linkage to tax data.

	Not linked to tax	Linked to tax	Total
Respondents	3,605	20,651	24,256 (76%)
Non-respondents	1,774	5,709	7,483 (24%)
Total	5,379 (17%)	26,360 (83%)	31,739

If everybody that was linked to the tax file had agreed to do so, there would actually be an increase in response rate. However, only 75% of respondents actually gave permission to use their administrative data. SLID also attempts to collect income for people who say they are non tax filers. Overall, there were two groups of people who could affect response rates: non-respondents to the income survey who gave permission to use their tax data and were linked would have a positive effect on the response rate while persons who gave permission to use their tax data but were not linked would have a negative effect. Approximately 1,700 persons were in each of the two groups. This means that in the end, the response rate remained the same with the mixed strategy.

However, because of potential biases due to a difference between linked and not linked respondents, these two groups were compared, using their income data from the first year survey collection. There were three subgroups with significant differences: single persons aged 15-19, single persons aged 20-24 and married women aged 45 years and more. Among these groups, there was a high percentage of records not linked and usually the incomes for the people linked and not linked were different (the non linked persons having a lower income). These groups of not linked persons were then compared based on whether permission was given or not. Five large categories were used in those comparisons : employment earnings (wages and salary plus self-employment income), investment income (taxable investment income including interests and dividends), government transfers (Unemployment Insurance, Social Assistance, Child Tax Benefits, Old Age Security, Canada Pension Plan, Workers' Compensation and Goods and Services Tax (GST) credits), and total income. The comparisons were done on a subset of records labelled "good" respondents. This was done to exclude potential effects due to imputation. Table 3 shows the results. A similar pattern was found for all income categories. This suggests that if a proper pool of recipients was defined, a fairly valid imputation model could be done for the unlinked persons, since there does not seem to be a difference between those who gave permission and those who did not.

Table 3. Comparisons of total income of "good" respondents (using survey data) for respondents who gave permission to use their administrative records vs. the ones who did not.

	"good respondents" who gave permission			"good respondents" who did not give permission		
	n	mean	median	n	mean	median
single 15-19	865	$ 2,624	$ 1,500	727	$ 2,458	$ 1,000
single 20-24	606	$ 10,987	$ 8,800	470	$ 9,623	$ 7,188
women/married 45+	1599	$ 12,771	$ 7,677	1183	$ 13,573	$ 8,160
others	7857	$ 25,657	$ 20,000	5509	$ 26,667	$ 21,567

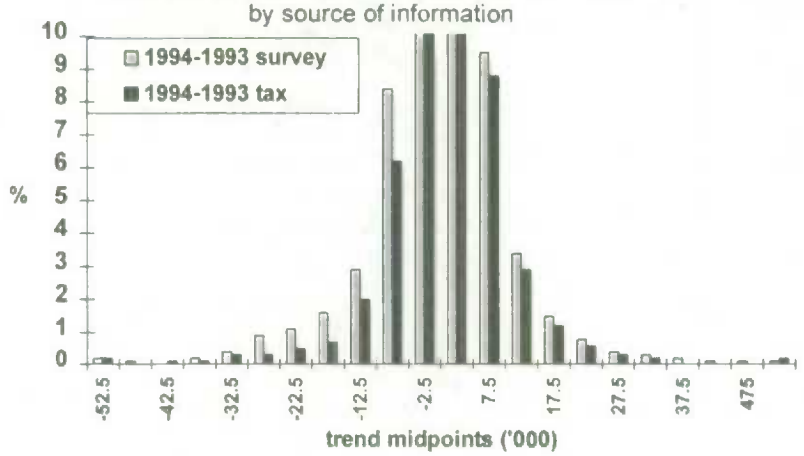
4.2 Response error

Income comparisons between survey and tax data have been done for a number of categories. A comparison between survey data and tax data for the "good" respondents, to see potential differences in income category definitions is given in [5]. These findings suggested that there were differences for self-employed income and social assistance. There were no differences in averages and medians for Wages and salaries and Unemployment Insurance benefits (UI). However there was still some under reporting of UI. The difficulty with this approach was to determine what was right. In particular, when there were differences in reporting of self-employment income, it was hard to see if this was representing income from the underground economy which would not be reported in tax, if it represented an income amount reported in some other income source or if it was wrongly reported. Since SLID is also interested mainly in longitudinal analysis, it was decided to study the differences in measures of change and to try to reconcile the micro-differences on records with two years of data. This also allows the study of response error with the longitudinal aspect in mind.

Only a subset of SLID respondents have two years of data from both the survey and tax. This limited the study to a sub-sample of 4274 respondents. This subsample is not quite representative of the whole sample; it has a slightly higher percentage of people in the 65 years of age and older group and smaller representation in the very young group 16-19. The differences were however not found to be important enough to invalidate the study. Of this subset, 86% of the records had been obtained through the direct match and so only that subset was kept, again to remove potential effects of incorrectly linked records. The comparisons were restricted to that subset of 3670 records. However, another 600 records were removed; 400 of these records had partial non-response in the second year, and most of the remaining 200 had no income in one or both years.

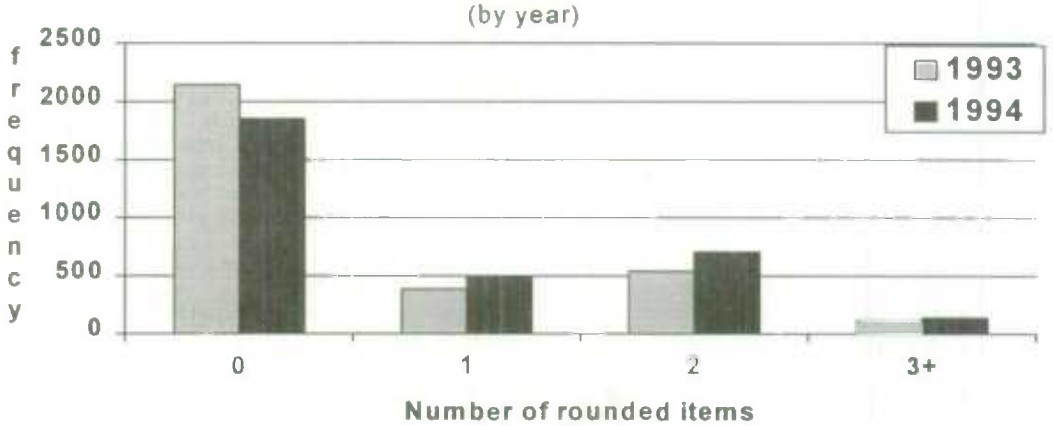
Particular attention was paid to the measures of change in total income between the two years from both survey and tax data. For each person, a change of total income --or micro-trend -- was calculated for the survey data and the tax data. Histogram 1 shows the distribution of the changes. The vertical axis scale has been cut to a maximum of 10% to allow a better view of the tails of the distribution (the scale should have gone up to around 30%) . The average change from the tax data was an increase of $498 while survey data suggested a decrease of $3 (the difference was significant at the 1% level). There also seemed to be more variability in the measures of change from the survey.

HISTOGRAM 1.
Distribution of yearly micro trends for total income
by source of information



The studies then compared the pattern of reporting. An initial look at the data suggested that there were two different behaviours depending on whether a person was giving approximate amounts (that was detected by looking at the income sources that were rounded) or exact amounts. So the study focused on the amount of rounding done. The study was restricted to the rounding of ten income categories only because other income categories were not reported in a similar way on both the survey data and the tax data. Rounding was defined to be when the two last digits were zero on survey but not on tax.

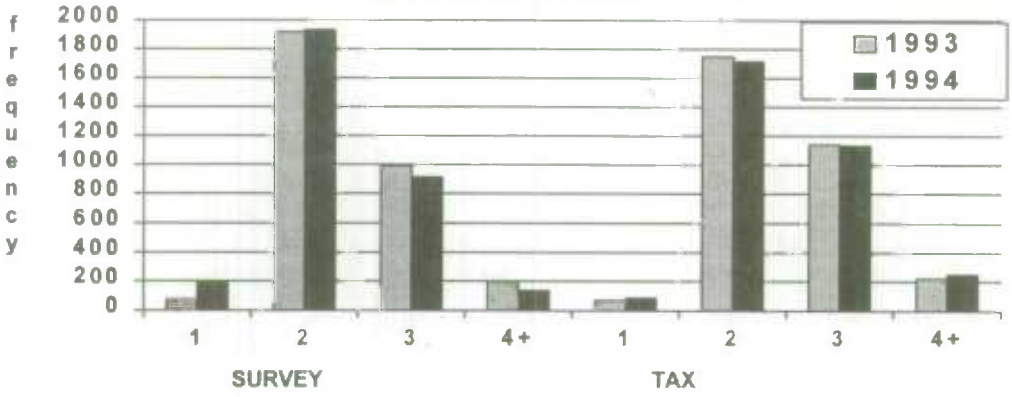
HISTOGRAM 2.
Distribution of the number of rounded items per individual
based on 10 items
(by year)



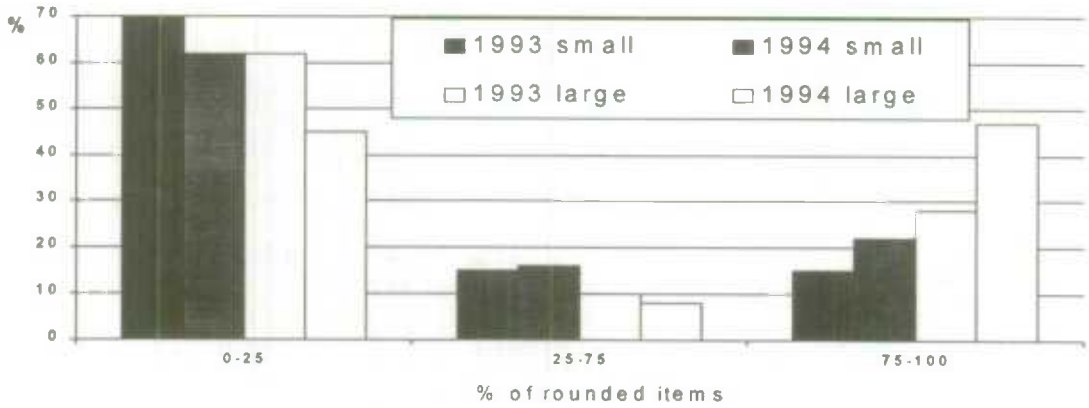
Histogram 2 shows the distribution of the respondents according to the number of items rounded in their survey data in 1993 and 1994 reference years. Rounding of income amounts happens frequently; only 1530 records, that is 47.5 % of the people do not round any of their income amounts in the two years. It also seems that the amount of rounding increases in the second year of collection.
Histogram 3 shows the distribution of the respondents according to the number of non-zero items reported to the survey. It is interesting to note that the average number of items reported to the survey slightly decreased in the second year of survey compared to the first year but this is not observed in the reporting on the tax for those same items.

It was of interest to know if rounding was different for different groups of respondents. Rounding behaviour over time was compared for various income groups. As indicated in Histogram 4, income seems to have an impact on rounding; those with large income tend to round more than those with small income. There also seemed to be less rounding for older people.

HISTOGRAM 3.
Distribution of non zero items per individual
(by year and source)



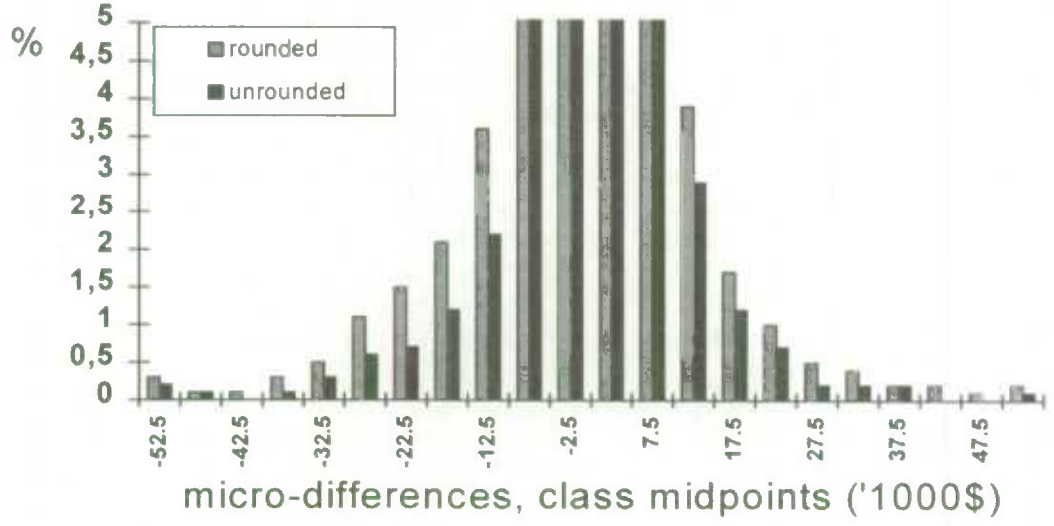
HISTOGRAM 4.
Distribution of individuals by proportion of rounded items
for small and large total income respondents
1993 and 1994 reference years



Once again total income was compared. The measures of change -- or micro-trend -- were compared using

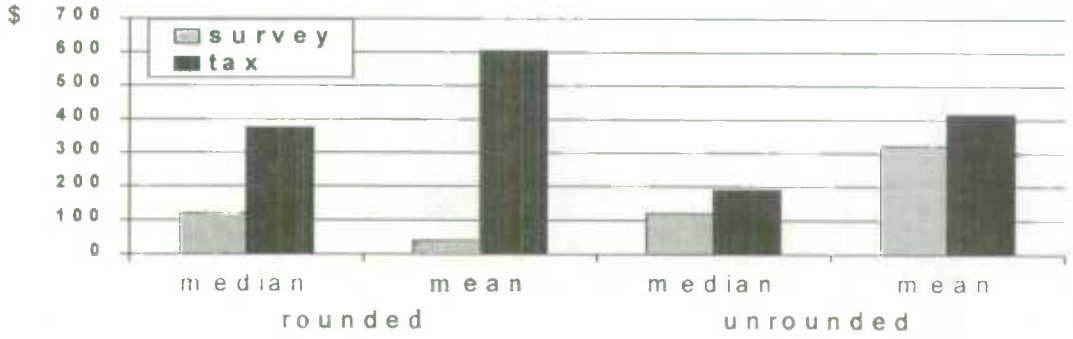
survey data, dividing respondents in two groups; respondents who rounded their reported income vs. the ones who did not. This is shown in Histogram 5. A respondent was assigned to the rounded group if at least one of the items was rounded. It is interesting to note that much of the variability of the measures of change observed in Histogram 1 is observed in the group of respondents who rounded at least one source.

HISTOGRAM 5.
Distribution of micro-differences
(yearly trends between 1994 and 1993)
for "rounded" and "unrounded" groups



To confirm this hypothesis, the measures of change were compared between survey data and tax data for the group of respondents who did not round the income amounts that were reported. Histogram 6 shows mean and median micro-trend separately for the "rounded" group and the "unrounded" group. For the group of people who did not round, the differences were not that large. As can be expected, the biggest discrepancies happen in the group of people who round their income.

HISTOGRAM 6.
micro-trends central tendencies
for rounded and unrounded groups



5. CONCLUSIONS

This was just a first look at the issue of response error. When some of the largest differences were examined by subject matter specialists, the differences were attributed to a response error in the survey in approximately 80% of the cases. Approximately 10% of the cases were attributed to an "error" in the tax data (a non taxable item was missing in one of the two years or there seemed to be an error in the tax field). Finally, the remaining 10% was not explainable.

There were some other interesting findings; approximately 30% of people provide exactly the same amounts (to the dollar) on both the survey and tax files, for at least one year of data. The rest have response error either from the survey or from tax files, or possibly both. Both sources of data have their limitations; the tax information has the problem of non tax filers and the underreporting of non taxable amounts while survey data seems to be prone to response error. The response error on the survey data also seems to increase with time. Based on the observed results, even if tax data is prone to error, the use of tax information in this mixed approach will probably improve the quality of the income data, especially because of the longitudinal nature of the survey. There are still things to investigate; the overall findings do not seem to hold quite as nicely for the self-employed. This group should be analysed further. In a similar fashion, the study should be refined to study reporting by income source. The impact of non taxable income amounts on the measure of change should also be evaluated in more detail. Finally, a number of techniques have been suggested to correct for response error [3], [6]. These techniques should be applied and tested to see if they can be incorporated to improve the quality of the measures of income.

The authors would like to acknowledge the contribution of Chantal Grondin, Martin Renaud, Carole Janelle and Elaine Fournier in preparing the study.

6. REFERENCES

Bound, J., Brown, C. Duncan., G. and Willard, R. (1991), "Measurement Error in Cross-Sectional and Longitudinal Labor Market Surveys: Validation Study Evidence", Panel Data and Labor Market Studies, J. Hartog, G. Ridder and J. Theeuwes (eds.), Elsevier Science Publishers

Dibbs, R., Poulin, S. and Webber, M. (1994), "The use of tax file data in the Survey of Labour and Income Dynamics: Summary report", SLID Research Paper Series, Cat. No. 94-11.

Fuller, W. (1987), Measurement error models, Wiley.

Groves, R. (1989), Survey Errors and Survey Costs, Wiley.

Michaud, S., Dolson, D. and Renaud, M. (1995), "Combining survey data and administrative data", SLID Research Paper Series, Cat. No.95-19.

Plewis, I. (1985), Analysing change, Measurement and Explanation using longitudinal data, Wiley.

Statistics Canada (1995), An overview of LAD, Longitudinal Administrative data. LAD report #94-20-01E, may 1995. Prepared by the Small Area and Administrative Data Division and Social Surveys Methods Division of Statistics Canada.

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

QUALITY ASSURANCE FOR THE CANADIAN CANCER REGISTRY

Leslie A. Gaudette, Tony LaBillois, Ru-Nie Gao and Heather Whittaker[1]

ABSTRACT

The Canadian Cancer Registry compiles data provided by the 12 provincial/ territorial registries (PTCRs), each of which collects data from a variety of administrative data sources. PTCRs vary widely in terms of the size of population covered, the number and complexity of data items collected and the sophistication in computer systems used. The CCR was developed as a person-oriented, longitudinal data file, comprising three modules: the core edits system, internal record linkage, and death clearance. The core module accepts only those records that have passed a stringent set of edits, with rejected records returned to the PTCRs for correction and resubmission. This, together with the update and deletion functions, ensures data contained in the CCR and in the dynamic PTCR data bases remain comparable. An important design component of the CCR is to develop and implement quality assurance processes. This initiative includes three basic components: establishing an infrastructure that includes development of approaches and tools, emphasis on communication, and phased implementation of strategies for performance measurement. Development of these components was guided by the six major attributes of quality assurance identified for cancer registries. A key feature of the initiative was the establishment of a Data Quality Committee that makes recommendations on data standardization and fosters regular communications. Acceptance sampling issues have been addressed through implementation of stringent edits at the PTCR level as defined in the CCR Data Dictionary and in the companion draft Procedures Manual. Process controls have been developed by setting guidelines for global monitoring, for key data items, of acceptable ranges of percentages of unknown values or of improbable combinations. Examples of guidelines for selected quality indicators and their importance in meeting requirements of users nationally and internationally are presented and discussed.

KEY WORDS: Quality assurance; Cancer registry; Administrative data; Nonsampling errors.

1. INTRODUCTION

Cancer is an increasingly important health problem in both developed and developing countries. In Canada, it is the second leading cause of death after cardiovascular disease. More than one in three Canadians can expect to develop cancer during their lifetime and about one in four will die of it (National Cancer Institute of Canada, 1996). Canada is one of the few countries, and one of the largest, that has developed a special data base called a cancer registry to provide information needed for effective cancer control (Statistics Canada, 1997; Band, et al., 1993). Cancer registries can be used to quantify the importance of the disease, describe the characteristics of the different types of cancer, and support a broad range of research applications, such as emerging developments in genetic origins of cancer, case-control and cohort studies, and economic analyses needed for health care planning.

Canada first developed the National Cancer Incidence Reporting System (NCIRS) in 1969 to collect population-based data on cancer incidence. In the mid 1980s, it was recommended that a new Canadian Cancer Registry (CCR) be built to meet the increasing demands of researchers for improved accuracy, quality and timeliness of cancer information. The CCR was developed as a standardized person-oriented updatable database to provide Canadian incidence and survival information required for cancer control. This new database operates with a

---

1 Leslie A. Gaudette, Senior Research Analyst, Health Statistics Division, Statistics Canada, Ottawa, K1A 0T6, e-mail, lgaudet@statcan.ca. Tony LaBillois, Senior Methodologist, Household Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, e-mail: labiton@statcan.ca. Ru-Nie Gao, Head Research Support Unit, Health Statistics Division, Statistics Canada, Ottawa, K1A 0T6. Heather Whittaker, Director of Records and Registry, Manitoba Cancer Treatment and Research Foundation, Winnipeg, Manitoba, R3E 0V9, e-mail, heather@mctrf.mb.ca

complete set of quality assurance concepts and procedures to ensure its reliability and comparability with other similar data bases in the world, as well as within our own country.

This paper describes the forces behind the development of the new CCR, describes its main components, reviews the quality assurance infrastructure, including examples of newly developed process controls, and concludes with some overall thoughts and future initiatives.

2. BACKGROUND

2.1 Brief History of Cancer Registration in Canada

Registries started in the provinces of Saskatchewan and British Columbia around 1930 are among the oldest in the world, and many other provinces set up registries during the 1940s and 1950s (Band, et al., 1993). Most provincial/territorial cancer registries (PTCRs) were established by the cancer agency responsible for cancer care in the province, while others were set up within the departments of health. By the mid 1960s, cancer registration in Canada had evolved to the point where a national system was feasible.

Thus, through the joint initiative of the National Cancer Institute of Canada and the Dominion Bureau of Statistics (now Statistics Canada), the National Cancer Incidence Reporting System (NCIRS) was established in 1969. The NCIRS, our first population-based national database on cancer, compiled data from the PTCRs until the diagnosis year 1991. In 1989, development of the CCR was undertaaken to further improve standardization, reliability and quality of cancer incidence data (Statistics Canada, 1997).

2.2 Data Sources Used by the Provincial and Territorial Cancer Registries (PTCRs)

In Canada, health care, and cancer registration, is a provincial or territorial responsibility. In the majority of provinces and territories, a legal requirement exists that cancer cases be reported to the appropriate agency, however registration remains voluntary in Newfoundland, New Brunswick and Ontario. Each registry attempts to record all new cases of cancer in its population by combining information from all available sources. These sources may include: cancer clinic files and radiotherapy reports; records from in-patient hospitals, out-patient clinics and private hospitals; pathology and other laboratory and autopsy reports; radiology and screening program reports; reports from physicians in private practice; and reports on cancer deaths from vital statistics registrars (Statistics Canada, 1997; Band, et al. 1993).

2.3 Overcoverage and Undercoverage

Comparability of data is affected by both under-and over-registration, which in turn are affected by the data sources and processing methods used. Coverage of registration at the Canada level is currently thought to be at or above the accepted standard of 95 per cent, (Howe, et al., 1996) although lower coverage has occurred in some provinces, especially in earlier years (Band, et al., 1993). Overcoverage occurs when two PTCRs unknowingly register the same case or where the identifiers on provincial files are insufficient to detect all duplicate reports. As the NCIRS was an incidence reporting system, cases were not linked together at the national level to identify patients with more than one primary tumour. Duplicate patients registered in more than one province were not routinely deleted (except for the two territories) before 1992. For more examples of over-reporting see Statistics Canada (1997) or Band, et al. (1993).

Undercoverage occurs when a registry does not use enough different sources to detect new cases of cancer. Some cancers are difficult to diagnose because of their location (or site) in the body. Others, such as lung, pancreas and stomach, quickly lead to deat. A registry that does not use death certificates as a source of information usually finds it has under-registered these cancers. Other sources of under-coverage include: lack of reporting of cancer cases treated in a province other than the province or territory of residence; use of different definitions of what constitutes a malignant neoplasm; and the lack of reporting of late registrations (or cases diagnosed after a registry has reported that year's cases) to Statistics Canada (Statistics Canada, 1997).

2.4 Other Differences Between PTCRs that Affect Data Quality

Comparability of registration, cancer incidence and survival rates, as well as accuracy and completeness of information provided will be affected by differences among registries coding practices, defintions of what is a cancer, and data entry or processing procedures. Some PTCRs collect more data items than others, and some provide more complete and accurate information for certain cancer-related variables. Timeliness is also an important issue: many PTCRs can report their data within 6 to 12 months after year end, others have only been able to report 3 to 4 years later. Some PTCRs have very sophisticated and flexible computer system. However, larger registries must use large complex systems that may require extensive re-programming, and smaller registries may have computer systems with insufficient capacity to incorporate all the computerized edits needed for the new CCR system. Quality of data at the national level therefore depends somewhat upon the extent to which PTCRs can implement the new reporting and editing requirements into their computer systems. Nonetheless, with the implementation of the new CCR, differences among the PTCRs are diminishing, particularly for items and practices of interest at the national level.

3. THE CANADIAN CANCER REGISTRY (CCR)

As of the 1992 diagnosis year, the CCR has been established as a standardized person-oriented longitudinal data base that includes mechanisms to update records, check for duplicates and (in the near future) provide national death clearance. The CCR is the culmination of many years of cooperation among PTCRs, Statistics Canada and many other organizations involved in cancer research in Canada.

3.1 Overview of the Canadian Cancer Registry

The CCR has three modules. The Core Edits module is the set of computerized procedures that read, verify, load, update and maintain the database. Internal Record Linkage detects potential duplicate patients or tumours on the CCR. The Death Clearance module matches or links patient records on the CCR with potential corresponding records on the Canadian Mortality Database (CMDB) (a data base that contains information taken from all Canadian death registrations). The CCR, after Death Clearance, can be used to conduct survival studies and to manage the size of the database by storing inactive cases separately..

3.2 Preparation and Submission of Data by the PTCRs for the CCR

Before data can be transmitted to the CCR, PTCRs must standardize them into a common format, and report information on each new patient and tumour to the CCR using standard record layouts, codes and edits defined by the CCR Data Dictionary. Unlike for the NCIRS, CCR data are not reformatted or recoded at Statistics Canada. Rather, the CCR applies a stringent set of validity and correlation edits; records failing to meet edits (usually less than 1%) are rejected and returned to the PTCR for correction and resubmission. Only data meeting all edit specifications are posted to the CCR data base. In cases where the level of errors on a submitted file is too high, the entire file is sent back to the province or territory for correction and resubmission. Data posted to the CCR must pass about 80 validity and correlation edits. A few warning edits are also used by the CCR and PTCRs to identify records containing combinations requiring review. Each PTCR annually reports Patient and Tumour records annually as part of either a regular or correction cycle. An electronic file containing the submission is sent to Statistics Canada and then used as an input to the CCR.

3.3 The Core Edits System

This module is the heart of the registry. It includes the database and all its associated functions. The CCR comprises two linked SAS databases, one for Patient Records and one for Tumours. As the CCR is person-oriented, data feeding into the CCR describe the individual with cancer on the Patient Record and the characteristics of that cancer on the Tumour Record. Each Patient Record has a unique CCR identifier and provincial Patient Identification Number that are used to link the corresponding Tumour Records stored on the other base.

The PTCRS may report the following types of Patient and Tumour Records to the CCR Core Edits module.

- *New record:* A new Patient Record is submitted for a patient who is not posted on the base and contains all personal information related to that individual. A new Tumour Record is sent for all new tumours diagnosed for a given patient .
- *Update record:* A complete update record must be sent each time a PTCR wants to change one or more fields on a patient or Tumour Record already posted on the CCR.
- *Delete record:* PTCRS can delete posted Patient or Tumour Records by submitting a delete record containing certain fields exactly identical to those on the posted record. To delete a Tumour, only one record need be sent, but to delete a patient, delete records corresponding to the Patient as well as all Tumour Records are needed.
- *Change-of-Ownership of a patient:* If a new tumour for an existing patient is diagnosed in a different province than the one where the previous tumour was diagnosed, the PTCR registry responsible for the registration of the new tumour may submit a change-of-ownership record to assume responsibility (and ownership) of the Patient Record.

The edit rules incorporated into this module are defined in the CCR Data Dictionary as the first step to ensure data quality. Validity edits are applied to verify the content of certain fields. Correlation Edits verify data items among the various combinations of Patient and Tumour Records reported for each patient. Input match edits ensure that all input records related to each patient are complete and make sense in terms of the operations to be performed (new, update, delete, change-of- ownership). Edit rules have been implemented to ensure that the changes to the CCR as a result of each submission respect its structure and internal logic.

The Core Edits system incorporates many new design features. The CCR operates in the second edition of the International Classification of Diseases for Oncology (ICDO). Multiple primary tumours are defined based on a set of rules that involves 4-digit ICDO codes for Topography (or site of cancer), groupings for ICDO Morphology (cell type) codes, and laterality. Automated conversions are in place at the national level between different versions of the codes. Warnings resulting from these conversions are returned to the PTCRs for review and submission of any update records required. PTCRs are also provided with ICDO and postal code conversion software to automate conversion to the Standard Geographic Classification down to the level of Census Metropolitan Area and Census Tract.

Within two weeks after receipt of a PTCR submission, Statistics Canada returns feedback reports to each registry to provide summary information as well as record-by-record descriptions of each record that failed a particular edit. Only for a few edits are warning messages given. A set of quality control reports monitor completeness of reporting of each item, as well as data quality indicators. Each PTCR is expected to correct or review its records based on review of these reports, and submit any resulting new or update records in the next submission to the CCR.

Duplicate reporting is controlled by the following procedures. Each PTCR is expected to return records for residents of other jurisdictions to the appropriate PTCR, as the CCR accepts Tumour Records only when the PTCR of reporting is the same as the province/ territory of residence. All incoming Patient and Tumour Records are checked by the Core Edits module for duplicates based on PTCR registration numbers. The Internal Record Linkage module conducts a final check for duplicate Patient and Tumour Records loaded onto the CCR.

3.4 The Internal Record Linkage Module

In addtion to reducing overcoverage, internal record linkage provides feedback to the PTCRs on other cancers involving their patients. One way duplicates can occur is if a tumour or patient is recorded in more than one source of information used by a PTCR and if that PTCR has assigned different patient identification numbers to the records. Duplicates can also occur if a tumour or patient are recorded in more than one province and each PTCR, if it has not previously exchanged information on that case, reports it to the CCR.

Each year, regular processing is stopped for two months while the Internal Linkage is run. The CCR is pre-processed and probabilistic record linkage is used to compare the database to itself. Probabilistic linkage uses complex rules that measure the degree of similarity of, and assign weights to, key variables that would not by themselves be unique identifiers. If the sum of weights for a given comparison is sufficiently high, the group of records is then examined by the PTCRs for potential duplicates. A complex set of programs is run to identify potential groups of duplicate patient and tumour records on the CCR data base. Feedback reports for each group are then produced by the system and sent to the appropriate PTCRs for duplicate resolution. About 4 weeks is allowed for the PTCRs to review the reports and to resolve the cases, which may involve consultation with other PTCRs. Reports are sent back to Statistics Canada with the decisions from the PTCRs as to whether or not the patient or Tumour Records were duplicated cases. The CCR database is updated accordingly and confirmation reports returned to the PTCRs.

3.5 Current Status of the CCR

The CCR has been in production since late 1994. As of December 1996, the CCR contained complete data for the 1992 and 1993 reporting years, data for most PTCRs for 1994 and from two PTCRs for 1995. Approximately 120,000 new Patient and 125,000 new Tumour Records are reported each year. The first Internal Linkage, performed in January and February 1996, found an internal duplication rate of 0.2%, with about half the duplicates occurring between PTCRs and the remainder within PTCRs. These results show that the PTCRs are efficient at detecting duplication before their data are sent to the CCR. However, the number of duplicates occurring between provinces may increase as patients have more opportunity over time to move between provinces. Development and testing of the Death Clearance module are now well under way and the first production run is scheduled for 1997.

4. QUALITY ASSURANCE

4.1 Quality Assurance, the Canadian Cancer Registry and the Data Quality Committee

Quality assurance (QA) is an important design component of the CCR and is needed to meet increasing user demands for timely, high quality data. Users comprise an increasingly diverse group ranging from medical geneticists through epidemiologists to health care planners. The QA approach use also aimed to optimize costs by moving responsibility for edit checks to the PTCRs, thereby reducing costs at Statistics Canada for routine production and minimizing response burden to providers. Automation of some coding functions has reduced costs to both PTCRs and Statistics Canada, and freed up time for other activities such as Internal Linkage and Death Clearance. This organized approach to QA has also reduced hidden costs associated with time-consuming clean-up of poor quality data.

Quality is managed for the CCR through a standing Committee on Data Quality (DQC) which reports to the Canadian Council of Cancer Registries (or Council). Statistics Canada co-chairs this Council, which has met at least annually over the past decade; members include representative of PTCRs and key user groups. The DQC makes recommendations and gives advice to Council on matters relating to the quality and standardization of data collected, coded and reported by the CCR. Consensus positions are developed based on periodic surveys of PTCRs conducted by the DQC to document current registration practices and procedures used across Canada. An infrastructure comprising a variety of *tools*, *communications* strategies, and methods for *performance measurement* addresses quality through processes aimed at improving various *attributes* or characteristics of the data.

4.2 Tools

The CCR Data Dictionary defines the standard record layouts, data items, validity edits, correlation edits, and reference and conversion files that must be used by PTCRs to report data to the CCR. Increased standardization of incoming data reduces operational costs by detecting errors earlier. PTCRs receive detailed feedback reports describing any problems occurring with each data submission. As medical care patterns change, ongoing vigilance is necessary to ensure that cases are registered from new sources, such as out-patient clinics and screening programs. A draft CCR Procedures Manual describes case definitions and reporting procedures needed to ensure case ascertainment reaches or exceeds the goal of 95% complete. The DQC also liaises with the North American Association of Central Cancer Registries (NAACCR) and the International Association for Research on Cancer (IARC) to coordinate implementation and use of standards across Canadian, American and other registries world-wide (American Association of Central Cancer Registries, 1994; Parkin et al., 1992; Parkin et al. 1994).

4.3 Communications

These technical approaches are balanced by a strong communications strategy. Registries may contact a qualified medical classification specialist, as well as other cancer registry staff at Statistics Canada by phone, fax, or e-mail, with any queries regarding coding and reporting issues. The *Cancer Record*, a newsletter produced two to three times annually, is the prime vehicle to communicate decisions about coding and reporting issues to technical staff located both in cancer registries and hospitals. It facilitates discussion of quality issues among data coders, registry technical staff, computer programmers, and medical doctors. Technical Workshops, organized and attended by cancer coders and data managers, have been held about every 18 months to provide training in new CCR procedures and to discuss issues related to improving quality and coding of data. Site visits to individual registries provide training in specific areas such as coding classifications or review of procedures to overcome obstacles in reporting data to the CCR. Finally, both the DQC and the Council communicate recommendations and decisions through minutes of conference calls and annual meetings. Implementation of these strategies has built an enthusiastic network across Canada that facilitates problem identification and resolution.

4.4 Performance Measures

The QA approach used for the CCR is based on statistical quality control methods: acceptance sampling, process control and designed studies (Hilsenbeck, 1994). *Acceptance sampling* involves inspection, and acceptance or rejection, of raw data through the CCR Data Dictionary and Core Edits modules. *Process controls* monitor production to identify when corrective action is needed (see also Section 5.2). *Designed studies*, the most sophisticated (and costly) method of quality control are used to assess completeness of case ascertainment (comparable to a reverse record check for the Census) or to reasbstract already registered cases to evaluate accuracy and completeness of reported data items.

4.5 Data Attributes

Quality of cancer registry data can be assessed by attributes that, in effect, represent the main sources of non-sampling errors. These are case ascertainment, accuracy, data completeness, timeliness and constancy (Hilsenbeck, 1994). *Case ascertainment* refers to both under- and over- registration of cases (see Section 2.3). Based on indirect measures, Canadian registries perform well, however, designed case ascertainment studies are needed to fully document problem areas. *Accuracy* of recorded data items is partially controlled in the CCR through the stringent edits in the CCR Data Dictionary, as well as global monitoring through quality control tables. However, designed recoding and reabstracting studies are also needed. The quality control tables also assess *data completeness*, by setting guidelines for acceptable percentages of unknown values (See also Section 5). *Timeliness* refers to the lag time in reporting and affects the usefulness of data, one attribute of quality. Improvements have been made to reach the 20 month reporting goal set by the Canadian Council of Cancer Registries, in part because the CCR system substantially reduces in-house processing time. *Constancy* of coding classifications and publication methods is addressed through agreement by Council regarding common implementation dates.

5. PROCESS CONTROLS FOR THE CCR

5.1 Overview of Process Controls for the CCR

The DQC has developed quality control reports and guidelines that provide feed back to PTCRs by monitoring the percentages both of unknown values appearing in key fields, and of records identified by applying algorithms for various quality indicators. Guidelines were developed in consultation with a number of sources and after thorough review of incoming data. Tabulations of data based on these guidelines are reviewed by Statistics Canada staff in consultation with the DQC to identify priorities for improvement and thereby contribute to improvements in Canada's system of cancer registration. PTCRs are expected to review cases for some indicators where the percentage of cases falls outside the suggested ranges. Where registries already have flags indicating the combination has already been reviewed, no further review is needed. Registries are also expected to provide explanations for results outside the ideal or acceptable range, and to take appropriate corrective action.

5.2 Examples

1. *Non-specific codes for cancer site/topography* measures the percentage of records reported to the CCR for which a specific cancer site (e.g., breast or lung) is either unknown or not specified. As all registries meet the proposed range of <5%, no further action is required (Figure 1).

2. *Non-specific codes for morphology (cell type) of cancer* measures the percentage of records for which a non-specific morphology code is reported to the CCR (e.g., although sarcoma is a cell type normally found in bone, in rare instances it may occur at other sites such as the breast). As results presented in Figure 2 are an average of all cancer sites, some cancer sites would have higher or lower percentages for unknown morphology. A level that is too high may indicate an unacceptable degree of imprecision in reporting or coding of diagnostic information, while a low level may indicate that a registry is not usingless reliable sources of information. Registries with results above the acceptable range of 0-8% are working towards improving the completeness of information provided on the source records (Figure 2).

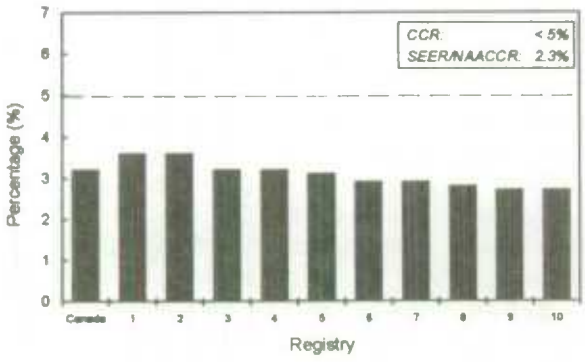


Figure 1. Percent Non-Specific Cancer Sites

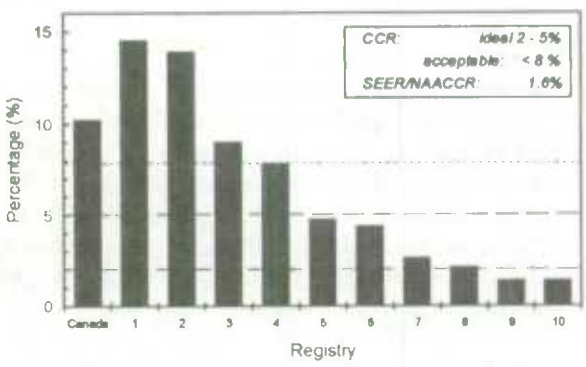


Figure 2. Percent Non-SpecificMorphology

In Figures 1 to 6, the first bar in each chart shows the Canada rate, with the ten provincial registries ranked in decreasing order for each indicator. As the rank order varies from indicator to indicator it is not possible to compare registries. The dashed lines show the ideal recommended levels, while the dotted lines on some graphs show acceptable levels or ranges. This analysis uses 1992 data from the Canadian Cancer Registry.

3. Improbable combinations of site-morphology-age. This indicator is an example of several CCR process controls that look at improbable or unusual combinations. For example, prostate cancer would rarely, if ever, occur in a five-year-old boy - the true age is probably 105! Each registry is expected to review all tumour records having an improbable site-morphology-age combination; registries not meeting the standard of <0.1% are expected to review their procedures, determine the reason, and take corrective action. Some registries edit these combinations

on-line and have a review flag that indicates unusual combinations have been verified. In all, about five combinations account for two-thirds of the improbable combinations (Gao and Gaudette, 1994) (Figure 3).

4. Percent Death Certificate Only (%DCO) measures the percentage of records where the only source of information on the diagnosis is the death certificate, an important indirect method of assessing completeness of case ascertainment. Improved consistency among registries seems to be occurring over time, and many registries are now at or close to the ideal range for %DCO. One registry does not yet perform death clearance, while DCO rates for another registry fell from over 10% to less than 5% in the past decade (Figure 4).

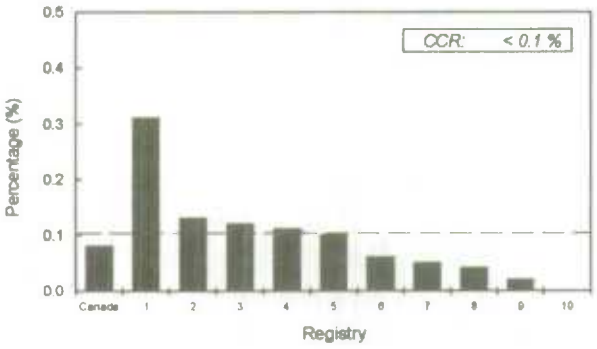


Figure 3. Percent Improbable Site/Morphology/Age Combinations

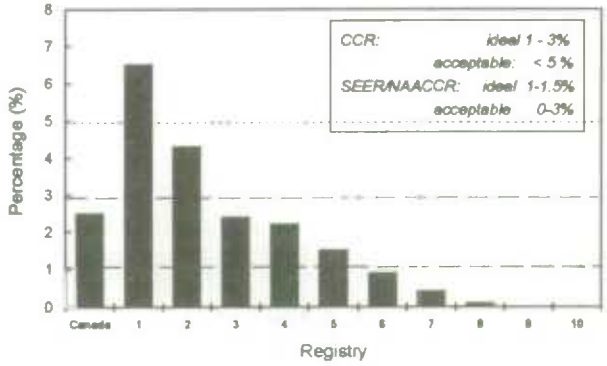


Figure 4. Percent of Death Certificate Only

5. Percent Unknown Census Division. For geographic information, a guideline of <5% missing was set based on analytical needs. PTCRs have access to automated geographic coding systems that reduce costs and improve completeness of reporting of geographic information. Registries with high levels of missing values can work towards accessing other provincial files containing postal code and other residence information to improve data completeness for this item (Figure 5).

6. Percent Unknown Dates (e.g., Day of Birth). The guideline for Day of Birth, originally proposed to be <1%, was increased to <5%, a level met by all but two registries. This change occurred because some PTCRs routinely impute unknown dates to be January 1, June 30 or July 1 of a given year, or unknown day to be the 15th of a known mont -- all valid dates in the CCR. The two registries with values >5% are known to identify imputed dates and convert them back to the correct values for unknown dates when reporting to the CCR. Thus, a low %unknown for Day of Birth, if based on imputed information, may not reflect the true %unknown values in a given registry, while a higher %unknown can reflect more accurate reporting (Figure 6).

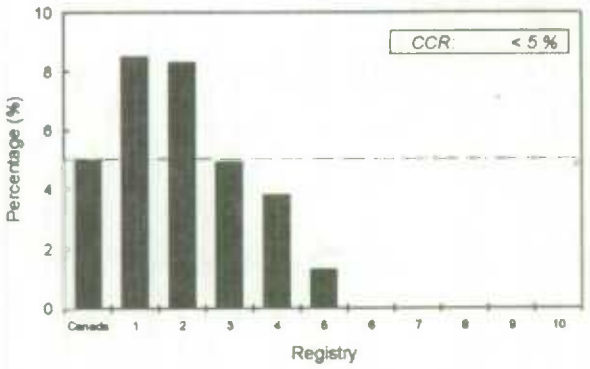


Figure 5. Census Division: Percent of Unknown Value

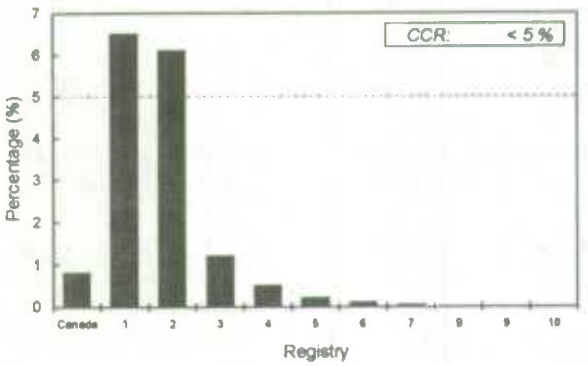


Figure 6. Date of Birth (Day): Percent of Unknown Value

6. SUMMARY AND FUTURE DIRECTIONS

Ongoing monitoring of data quality and implementation of new measures to resolve quality issues are needed to meet divergent user needs. While timeliness of data reporting has improved, reporting delay is still a concern. High quality data are needed by researchers in Canada and elsewhere to justify grant applications and document and justify validity of research results. At the same time, improved consistency of data at the international level is increasingly important to ensure meaningful interpretation of data in order that cancer control research and program development can be tackled as an international problem. Diminishing resources available to support cancer registry development and use need to be carefully managed and prioritized to optimize costs associated with quality assessment and production.

At present acceptance sampling is achieved relatively effectively using a number of tools, particularly the CCR Data Dictionary, to assure data quality. Feedback and quality control reports for the CCR have been effective in tracking edit problems and in identifying more global quality issues. The process control guidelines were approved by Council in May 1996. Implementation of this comprehensive QA approach has been accomplished through extensive consultation and involvement of staff from a variety of disciplines within both provincial/territorial and national registries and represents a significant step forward for quality of cancer registry data across Canada.

Case ascertainment remains an issue. The DQC is currently developing, in collaboration with NAACCR and key user groups, a plan for a designed study of completeness of registration of cancer incidence data based on that developed for NAACCR. At the same time, the CCR does not include information on staging and treatment, data that are increasingly needed to monitor cancer control efforts and to assess treatment results. And finally, the CCR and indeed, all cancer registries must be increasingly responsive to a changing environment, including political, medical, technological and economic trends.

7. REFERENCES

American Association of Central Cancer Registries. (1994). *Standards for Cancer Registries Volume III. Standards for Completeness, Quality, Analysis and Management of Data.* Seiffert, J.E. (ed.) Sacramento, CA: NAACCR Cancer Surveillance and Control Program.

Band, P.R., Gaudette, L.A., Hill, G.B., Holowaty, E.J., Huchcroft, S.A., Makomaski Illing, E.M., Mao, Y., and Semenciw, R.M. (1993). *The Making of the Canadian Cancer Registry: Cancer Incidence in Canada and its Regions, 1969-1988.* Ottawa: Canadian Council of Cancer Registries, Health and Welfare Canada, Statistics Canada.

Gao, R.N., and Gaudette, L.A. (1994). *Assessment of edits for improbable or impossible combinations of site, morphology and age for Canadian cancer registries.* Ottawa: Health Statistics Division, Statistics Canada.

Hilsenbeck, S.G. (1994). Quality Control. In: Menck H., and Smart C., (eds.), *Central Cancer Registries:Design, Management and Use, pp 131-177.* USA: Harwood Academic Publishers.

Howe, H.L., Lehnherr, M., Derrick, L., et al. (1996) *Cancer Incidence in North America 1988-1992.* Sacramento, CA: North American Association of Central Cancer Registries.

National Cancer Institute of Canada (1996). *Canadian Cancer Statistics, 1996.* Toronto: National Cancer Institute of Canada.

Parkin, D.M., Chen, V.W., Ferlay, J., Galceran, J., Storm, H.H., and Whelan, S.L. (1994). *Comparability and Quality Control in Cancer Registration.* Lyon: International Agency for Research on Cancer. IARC Technical Report No. 19.

Parkin, D.M., Muir, C.S., Whelan, S.L., Gao, Y.T., Ferlay, J., and Powell, J., (eds.), (1992). *Cancer Incidence in Five Continents, Volume VI.* Lyon: International Agency for Research on Cancer, IARC Scientific Publications No. 120.

Statistics Canada (1997). *Cancer Incidence in Canada 1969-1993.* Ottawa: Statistics Canada Catalogue 82-566 (occasional).

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

THE IMPACT OF LEGISLATION AND ADMINISTRATIVE PRACTICES IN THE FUNCTIONING OF A REGISTER-BASED STATISTICAL SYSTEM

- A CASE STUDY ON THE REGISTER OF UNEMPLOYED JOB-SEEKERS

Timo Koskimäki[1]

ABSTRACT

The most usual way to evaluate statistics based on administrative records is to compare results with a survey or a census on macro-level. In the article the quality of the administrative statistics is also assessed by micro-linkage with a large-scale survey and by studying the consistency of the register over time. Although the quality of a administrative statistics may seem acceptable on the basis of macro-level comparison, the checking of consistency and linkage with survey-data reveal that the quality of the register under study varies quite a lot. As errors to certain extent outnumber each other, the register of the unemployed job-seekers is acceptable as a macro-level indicator. When the data is used in micro-level analysis or as auxiliary information in survey estimation the varying quality of the register may, however, cause problems.

KEY WORDS: Administrative data; Quality; Register; Labour Force Survey.

1. BACKGROUND

The problem of quality in administrative statistics is often quite difficult to assess empirically. Three fundamentally different kinds of approaches can be used. The most common one is to try to benchmark administrative statistics with some kind of auxiliary information at a macro level. If empirical material is available and data protection issues can be solved, a micro-level comparison is as well possible. The third possibility is to study the administrative practices beyond the register and try to evaluate and measure their impact on the quality of register-based statistics.

This paper describes the factors affecting the contents and quality of the (Finnish) Register of the unemployed job seekers - the Claimant Count - from the three points of view. The monthly Labour Force Survey will be used as a benchmark and as a source of micro-level information.

The point of departure is to compare measures as short-term economical indicators. Other aspects of unemployment - e.g. unemployment as a social and individual problem - are in this paper left aside.

2. THE TWO MEASURES OF UNEMPLOYMENT

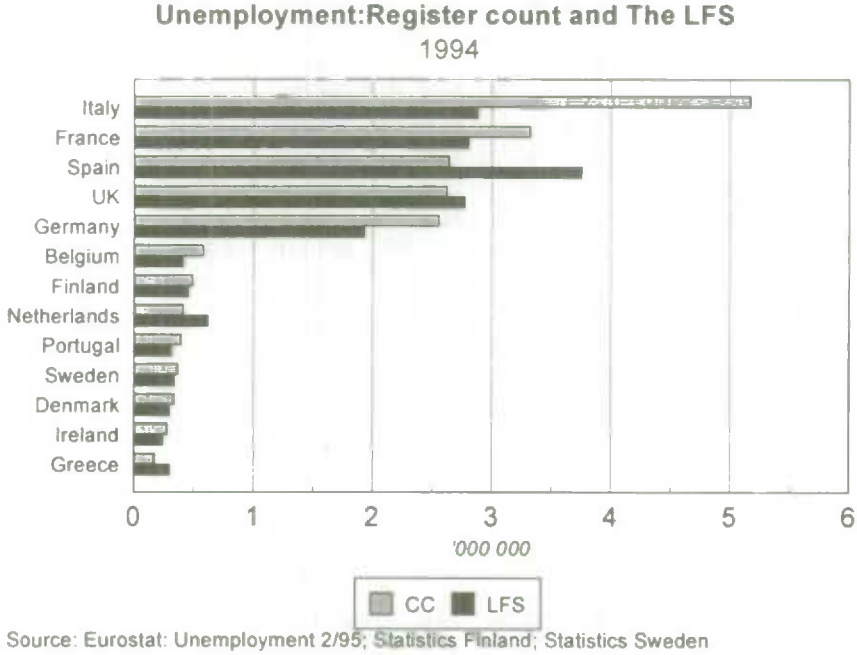
The existence of two official measures of unemployment - one based on administrative sources as The Register of Unemployed Job Seekers and another based on sample surveys like The Labour Force Survey - is a common situation in most countries. The two measures may give quite a similar picture of the level of unemployment, but most often they fail to do so. Figure 1 highlights that the interrelationship between the

[1] Timo Koskimäki, Statistics Finland, P.O.B. 5B FIN-00022 Statistics Finland E-mail: timo.koskimaki@stat.fi

register-based and LFS-based unemployment-estimate is controversial. In some cases LFS shows considerably lower level of unemployment than the register count. In other cases the situation may be the opposite LFS showing considerably higher level of unemployment.

The general reason for these differences is quite evident: the coverage, contents and quality of the register-based data depend on the infrastructure of the statistical system i.e. legislation, central and local administration. In essence, the origin of the standardised labour force surveys can be seen as an attempt to solve the problem of inter-regional - and later international - comparability of unemployment statistics (Innes 1990).

Figure 1:



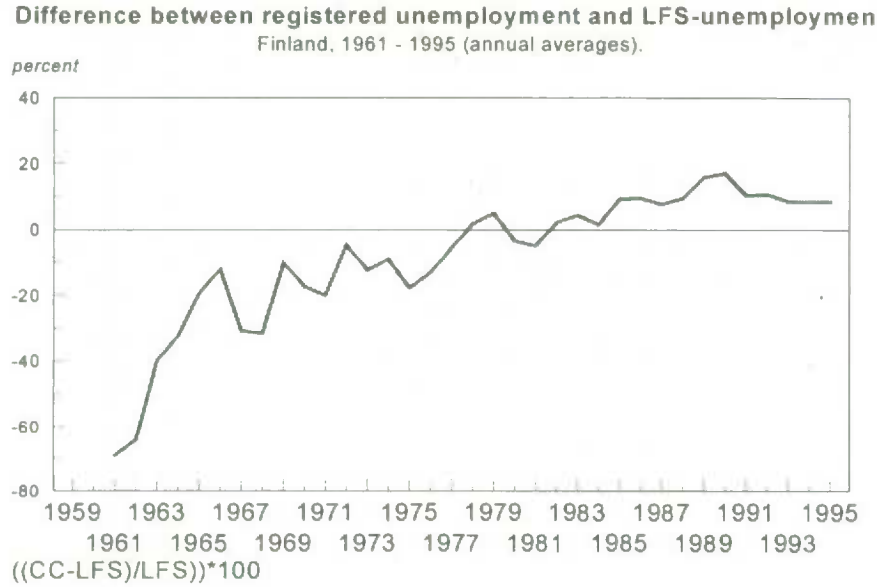
Source: Eurostat: Unemployment 2/95; Statistics Finland; Statistics Sweden

2.1 The case of Finland

In Finland both measures of unemployment are, at a first glance, astonishingly close to each other (see Annex 1). Should one take a more historical perspective one finds out that beyond the general unanimity of the measures the interrelationship between them does vary over time. The relatively high unanimity of the measures is most evident during a decade from mid 1970's to mid 1980's. Before that time unemployment was clearly higher according to The Labour Force Survey. Since mid 1980's the situation has been quite the opposite (Figure 2).

The reason for the one-decade unanimity lies in the fact that both the coverage of the unemployment insurance system and the resources of public employment administration could be characterised as sufficient during that period. The building of the general unemployment insurance system was completed and considerable unemployment-based benefits were channelled through employment authorities. At the same time the administrative system became more centralised and more effective, not only as a distributor of benefits but also as a supplier of jobs for the unemployed. Also, local authorities were able to keep relatively tight control over their clients. For example, all beneficiaries were obliged to contact the local office every fortnight. The sanction for not doing so was the loosing of the benefit.

Figure 2:



As a result of the above factors - and, at the same time, of a rapid economic growth that enabled the building of the unemployment services system - a situation emerged where most of the excess supply of labour - that is unemployment - actually did channel through the public employment administration.

2.2 Rolling Back the Statistics?

This paper from now on concentrates on the past eight years, the period where register-based data systematically has given somewhat higher estimates of unemployment than the Labour Force Survey. The general reason for the increasing difference between the two statistics was the diminishing official control of individuals - in this case the unemployed - and the role of state turning from control and administration -orientation towards service-orientation, in neoliberal political discourse often denoted as "rolling back the state".

One of the key indicators of the administrator's' control over its clients is the frequency of contacts client is obliged to take with the authority. In 1970's, every claimant had to report him/herself at the authorities every fortnight. The two-week re-registration interval was, during the 1980's, gradually lengthened to one month.

The decision on the time-span between claimants' obligatory contacts with the local office was decentralised from central administration to so-called employment services districts. Districts, in turn, delegated the decision further to local offices.

By 1988 every Local Employment Services Office got the right to decide for the length of the reporting period. In most offices the length of the period increased from one month to approximately two or three months. Later the reporting practice was further lightened and it is now the practice that the length of the reporting period is determined - by the employment service officer serving his/her client - individually for every registered person, case by case. This has resulted in reporting intervals round 6 to 12 months.

The sparse contact with unemployment agencies causes that register-based statistics lag behind, that is the share of persons that already have managed to get an employment although they are still formally registered as unemployed increases. This factor is also dependent on the general economic situation: lag becomes more severe when the economic situation is good and the probability of getting a job increase.

During an economic recession the lag diminishes. The effect of this can be seen in figure 3: during the very rapid economic growth of 1988 to 1990 the difference between the measures rocketed from 8 per cent to 17 per cent. As the recession started in 1991 the series moved again closer to each other the difference now being around 8 per cent (see also Koskimäki 1992).

3. RELIABILITY OF A REGISTER-BASED STATISTICAL SYSTEM

Our conceptual framework for analysing and describing register-based statistics is quite poor when compared to the rich methodological and conceptual tools we are able to use when evaluating survey-based statistics. As Myrskyla, Tauber and Knott (1995, 321) have stated, work on the quality issues of administrative data has so far been limited to consistency over time within the data-base and sample-surveys as quality checks.

I will in the following description make the analysis more systematic by making use of the commonly used ideas reliability and validity. Reliability refers quite straightforward to the statistical system's capability to measure 'what it is intended to measure'(Brackstone 1987, 32). The term validity refers here to the usefulness of the register as an indicator of labour-market behaviour.

3.1 Technical reliability

The use of the register of unemployed job-seekers for compiling statistics is quite simple. Once a month, at the end of the last trading-day, a cross-sectional register count of so-called "valid spells of unemployment" is compiled. The possible errors stem from the practices applied in registration of new spells and deletion of "completed spells" of unemployment. The possible sources of error are equally simple: they may derive either from the administrator's failures to record information correctly or from the client's failure to inform the administrator.

In the following I will use two methods to evaluate the performance of the register. The first method relies on quality-assurance system of the register-administrator. The second method is based on an attempt by Statistics Finland to recompile the claimant count from cleaned and more complete register-versions.

The register-administrator's quality assurance is made annually in the beginning of each statistical year. The practice is to compile count for the end-of the year situation from a register that has been left to update for 1 month as compared to only a couple of days for the published statistics. Another comparison is made with the register after an update period of thirteen months. Results of these comparisons are shown in Figure 3.

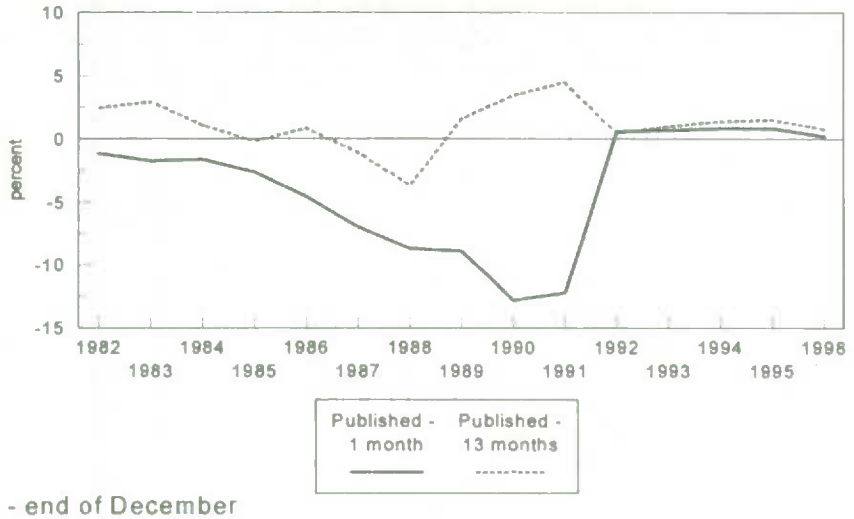
The behaviour of the register changed clearly in 1991. During pre-1991 period counts from different register versions oscillate: The first month updates are for the most additional cases of unemployment so that the +1 month versions show up to ten per cent more unemployed than the published version. During the succeeding 12 months the updates concentrate on "false valid spells" so that the final value of the register again shows somewhat less unemployed than the published count. After 1991 most updates have been deletions of false spells and both the +1 month version and the +13 month versions show less unemployed than the published count.

The second technical check was to compile a monthly register count from a register-version that had been let to update and which also contained information on the short-term lay-offs. Persons on lay-offs are classified as unemployed in the monthly claimant count publication but are not included in the yearly

quality check of the register administrator[2]. As an addition to the standard quality checks we also checked all months of the years under study, not only the end-of-year situation.

Figure 3:

Cross-sectional count compared with updated register:
register versions +1 month and +13 months after count



- end of December

The result of the second technical check is shown in figure 4. The updated register contains systematically less unemployed than the cross-sectional one (with one exception, December 1990). This indicates that the effect caused by delays in deletion of outgoing records has dominated over the delays in registration of new spells of unemployment. Also a considerable variation between different months of the year is apparent.

Since December 1990 the relative update -lag has been clearly lower than during the last years of 1980's. Pre -1990 the cross-sectional register contained about 30 per cent more unemployed than the updated one. The update-lag also seems to concentrate on summer months to certain extent and again is lower during the winter. After 1990 the general level of lag has reduced down to 5 per cent and no clear seasonal variation can be detected.

There are at least three explanations why the level and patterning of the register-lag changed. First, the economic situation changed dramatically. The number of unemployed was quadrupled in two years. This caused a severe jam in local employment offices. It is quite probable that the exceptional December 1990 when the updated register had slightly more unemployed than the cross-sectional one is due to the local office's incapability to handle their in-flowing new clients.

The severe economic recession also decreased the possibility of getting a job thus keeping the register more up to date simply because the share of completed spells of unemployment - and the failures in the registration of those spells - decreased.

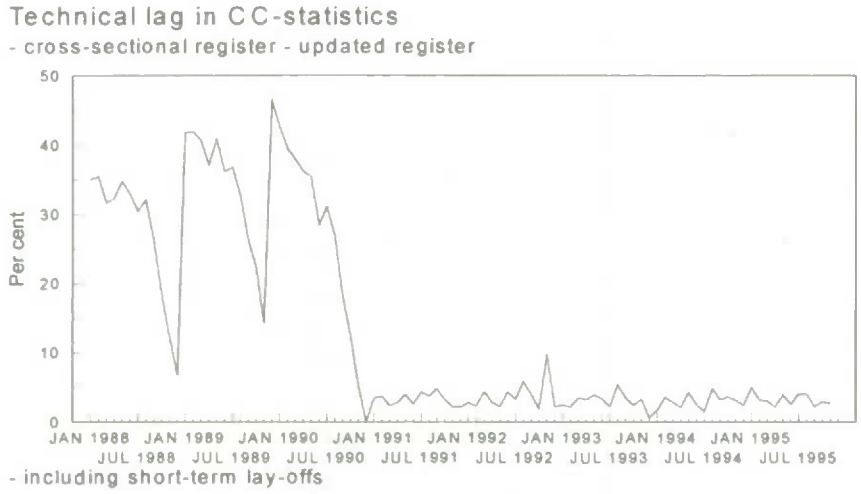
Second, the compilation of statistics was centralised: pre 1991 the information system was based on fourteen databases that were administrated by district authorities although the contents and structure of the databases were centrally planned. CC-statistics was compiled in two stages, first the district-wise counts

[2] About 5 to 25 percent - depending on month and economic situation - of published claimant count consisted of persons on short lay-offs.

and on the basis of district aggregates a national count. The 1991 information system consisted of a centralised database where the national total can be calculated directly.

The technical changes involved in this new system are probably of less importance than the fact that in combination with the technical change the local offices and districts were given new standards and regulations how to deal with the recording of the unemployed. Also, the possibilities of the central authority to control the district authorities' performance became easier in the centralised data-base environment. This seems to have a crucial effect on the quality of the register. Pre-1991 it can be characterised as being of poor administrative quality, whereas after the centralisation quality is quite acceptable.

Figure 4:



Third, due to the recession certain legislative changes that restricted the coverage of unemployment-related benefits occurred. The one explaining the change in seasonal pattern of the lag was the lengthening of the required engagement to labour market before one is entitled to unemployment-based benefits. In 1980's the required spell was only three months. Due to such a short period most students seeking casual work over the summer registered themselves as unemployed job-seekers.

This group, however, quite often did find a casual work and continued their studies in the fall but stayed in the register as unemployed job seekers to be removed not until they had failed to re-register themselves. In 1990's the required period of labour-market activity was lengthened to six months and students seeking casual work were not accepted as unemployed job-seekers, unless if they explicitly stated that they have no plans to continue their studies.

3.2 Substantial reliability

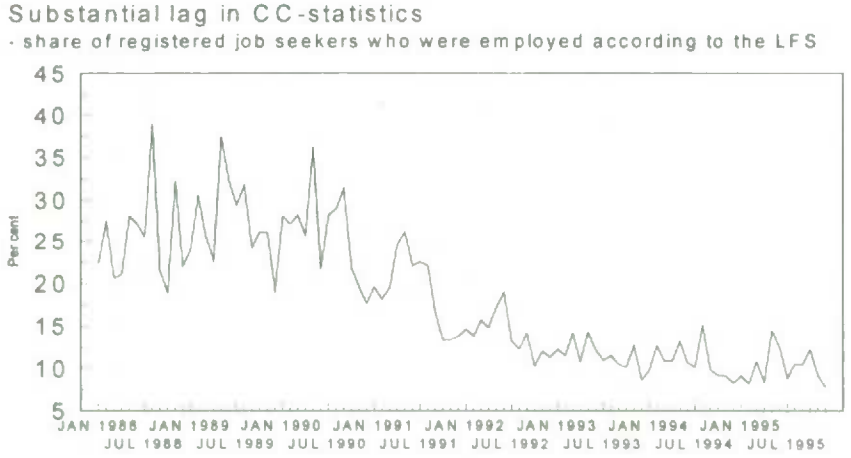
Previous chapter described errors caused by failures in the performance of the authority. Another set of potential errors stems from the behaviour of the clients in their contacts - or lack of contacts - with the authority. Two types of errors can be detected here. One is a situation where unemployed does not register him/herself as unemployed although it would be appropriate to do so. This type is close to validity issues discussed in the next chapter. The other, and more common type of error is a situation where no information on the true state of the client is obtained. The deletion from the register occurs after the agreed re-registration day passes. The client might also inform the authority about the change in his/her status with a delay. It depends on the authority's practices whether this type of unemployment-spell is recorded as being terminated in the day the information is reaches the authority or on the "true" termination day of the spell.

To measure this type of errors the information on whether a person is registered as unemployed or not was linked - at an individual level - to the monthly LFS-samples[i]. After doing this, a labour market status according to the LFS was cross-classified with the register-status of being unemployed job-seeker. As an estimate of this type of register-errors a share of those LFS-respondent who were in fact working according to LFS is given in figure 5.

Here the effect of the economic cycle is again apparent: In the years of growth close to 30 per cent of those that were registered unemployed were actually working according to the LFS. As the recession started the share of working claimants dropped down to round ten per cent.

As a coarse measure of reliability for a cross-sectional register count we can use the sum of technical and substantial errors: Hence, the technical reliability of the register count - the degree of its correctness in measuring unemployment according to administrative definitions - has varied considerably over time: during the growth of 1980's it was only about 0.4 but improved in the years of recession to over 0.8.

. Figure 5:



4. VALIDITY AS A LABOUR MARKET INDICATOR

The ILO recommendation on the statistics of unemployment states that, to be classified as unemployed, a person should have actually done something specific to obtain work, that is actively seek work. The other two criteria to be classified as unemployed are that a person has no work at all and that person state that he/she would be able to start to work within a short period of time. The specific activity to seek work - taking contact with an employer, contact with the employment services office e.t.c. - should have happened within four weeks preceding the reference period (Surveys of... 1990).

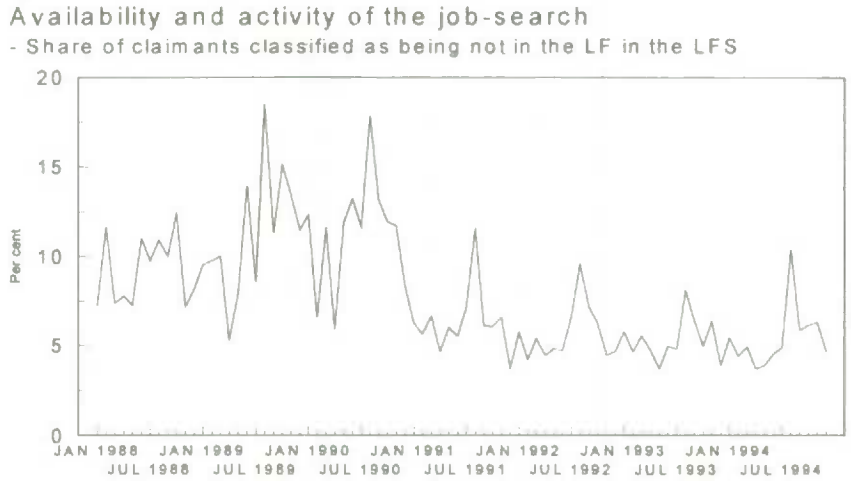
4.1 Activity of the job-search

It is obvious that this four week criterion has not been met in the Ministry of Labour statistics as such, because the obligatory re-registration intervals have been far longer than the required four weeks. From the point of view of the LFS this does not mean, however, that all claimants that have not had the contact with the employment office within the required time-span should be excluded from unemployment. Unemployed typically use multiple methods of job search so the claimant may in the LFS be classified as actively seeking work on the basis of other methods of job-search. The share of the passive claimants, i.e.

persons who were registered as unemployed job-seekers but according to the LFS were classified as being not in the labour force is given in figure 6.

The share of passive claimants has been quite stable slightly below 10 per cent throughout the period under study. The slight decrease after 1991 is probably due to changed legislation concerning students' rights to be registered as unemployed. A seasonal peak for the summer months probably stems from the public holidays: during the peak of the summer it is quite difficult to search for work and the LFS -respondents really do state that they have not searched for work or would not start the work within two weeks. Understandably, they still maintain their registration as unemployed job seekers.

Figure 6:



4.2 Coverage of the register

The measure of coverage of register of the unemployed job seekers - the share of those LFS-unemployed who were not registered as job-seekers - is shown in figure 7.

Here we find again a level shift that follows the economic cycle. In Late 1980's only about half of the LFS -unemployed were also registered as unemployed; in 1990's the share of those non-registered dropped down to slightly over 15 per cent.

Figure 7:

The plausible reason for higher non-registration during the economic growth is that the spells of unemployment tended to be relatively short - it was significantly easier to get a job than to wait for unemployment benefit. In 1990's situation was quite the opposite, resulting in a very high degree of registration.

The emerging seasonal pattern - a peak during the summer, again - is a parallel phenomenon to seasonal register-lag (see figure 4) in 1980's. As legislation concerning students' right to register themselves was changed the lag disappears but instead, there emerges a group of students not being able to find a casual job for the summer. As they are no longer accepted as claimants they appear as unemployed only through the LFS.

5. CONCLUSIONS

5.1 Strategies for assessing the quality of administrative statistics

As stated in the beginning, the assessment of quality in administrative statistics is often quite difficult to do empirically. In this study the problems were overcome by exceptional possibility to link the contents of the register to a large-scale survey for a long period of time. This type of analysis also brought to light the most severe source of error, i.e. the portion of claimants who in fact already had found a job.

Also the checking of consistency over time proved to be quite a useful tool to assess the quality of this type of register.

The picture we get from a macro-level comparison is quite different: As macro level indicators both sources give approximately the same result. The reason for this is that the errors in administrative register to certain extent did outnumber each other: as the coverage of the register was poor, the update lags were exceptionally high. The net effect was that the register showed the same level of unemployment as the LFS.

One of the basic findings in this study is that the errors in administrative statistics do depend on the phase of the economic cycle. In the case of LFS this result has been found to hold also in other cases of LFS than described here (Eurostat 1996) and in statistics on other areas, e.g. the statistics on the balance on payments.

Commonly used way to evaluate quality in register-based statistics is to incorporate a question on registration/program participation/receiving a benefit in a survey. This strategy is as well used to obtain national total of claimants in situation where no centralized register count for total is available. It will in theory work better as a validity-check than as a check for technical reliability of the register - the errors found to be significant here were for the most part due to shortcomings in the performance of the administrator and there is no way we can obtain that information from the respondent. There may also be risks for survey-type shortcomings[ii] like underreporting in sensitive issues like asking about received benefits (see Steel 1996).

5.2 Usability of the administrative statistics

Used as a general macro-level indictor on the level of unemployment it does not make much difference which series one chooses. Should one analyse the macro-series more thoroughly the instability of the claimant count may cause some inconvenience: Unemployment is (in Finland) clearly seasonally patterned and to enable month-to-month comparison seasonal adjustment has to performed. The standard X11-Arima performs worse in CC-series than in LFS-series: The error in Arima-forecasts tends to be higher and the detection of turning points of the trend is more difficult.

The register of unemployed job-seekers has also been used as a sample-frame for micro-level studies on unemployment. Here the risk of getting erroneous and misleading results due to shortcomings in the register is more acute. In late 1980's a sample drawn from the register would - taking the LFS-unemployment as a benchmark - have covered about 25 per cent of the total of LFS-unemployment. Same types of risks are also apparent when the content of the register - for example information on the duration of unemployment - is used in the evaluation of the labour market policy or research on labour market dynamics.

The use of administrative data as auxiliary information in survey-estimation is sometimes seen as promising effort to overcome both sampling- and non-sampling errors involved in surveys. To certain extent this is true, but, based on the experiences in using CC as auxiliary information in the LFS, the change of quality in register as a matter fact does penetrate through more technical procedures like weighting. The use of administrative data as auxiliary information in surveys should be carefully considered as well from the point of view of its quality.

6. REFERENCES

Alonso, William and Starr, Paul (editors): The Politics of Numbers. A Census Monographs Series, Russell Sage Foundation, New York 1987.

Brackstone, G.J. Issues in the Use of Administrative Records for Statistical Purposes. Survey Methodology, June 1987 Vol 13, No. 1 pp. 29-43. Statistics Canada.

Davidson, Roger. Official Labour Statistics: a Historical Perspective. J.R. Statist. Soc. A (1995) 158, part 1, pp. 165-173.

Innes, Judith Eleanor: Knowledge and Public Policy - The Search for Meaningful Indicators. Second Expanded Edition. Transaction Publishers, New Brunswick, New Jersey 1990.

Koskimäki, Timo: Rolling Back the Statistics? - The Unemployed in the Employment Service Register and in the Labour Force Survey. The Yearbook of the Finnish Statistical Association 1991, pp. 9 - 27. Helsinki, 1992 (In Finnish, with summary in English).

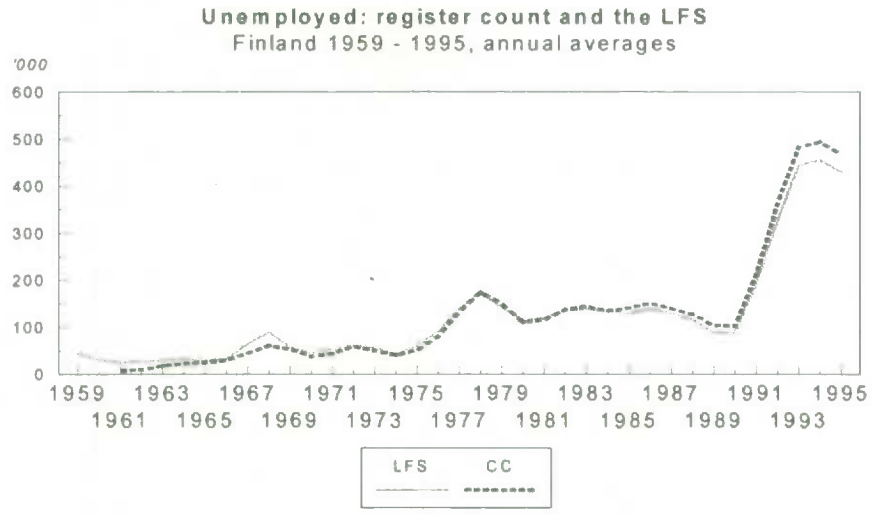
Myrskyla, Pekka; Taeuber, Cynthia and Knott, Joseph: Uses of Administrative Records For Statistical Purposes: Finland and the United States. Proceedings of the 1995 Annual Research Conference, March 19-21, Arlington, Virginia.

Steel, David. Options for Producing Monthly Estimates of Unemployment According to the ILO Definition - Report of the work undertaken by the Task Force. Central Statistical Office, United Kingdom, 1996.

Surveys of Economically Active Population, Employment, Unemployment and Underemployment - An ILO Manual on Concepts and Methods. ILO, Geneva, 1990.

Surveys of Economically Active Population, Employment, Unemployment and underemployment. An ILO manual on Concepts and Methods. International labour Office, Geneva, 1990.

Annex 1:



[i] The linkage was done using so called social security code which is a general unique identifier used in most administrative systems in Finland. This kind of information linkage is permitted only for statistical purposes.

[ii] Some basic information on the Finnish monthly LFS used here as a reference:Sample size about 12,000 individuals/month; Non-response rate round 7 percent. Non-response rate tends to be somewhat higher for the claimants, especially for those with longer spells of unemployment. Due to sample-size and very low level of unemployment in late 1980's some of the fuzziness in time series describing that period may stem from sampling error.

CLOSING KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 96
Nonsampling Errors
November 1996

NONSAMPLING ERRORS AND SURVEY ESTIMATION

Wayne A . Fuller[1]

ABSTRACT

Because nonsampling errors are pervasive in sample surveys, a multiple-faceted approach is required of survey statisticians. In designing the survey, practitioners attempt to identify sources of nonsampling errors and to minimize their effects subject to budget constraints. The presentations in this symposium illustrate the nature of these efforts in different areas of survey activity. Because all survey data contain nonsampling errors, it is important to devote resources to determining the properties of the measurement error, or (and) to include items in the collected data that can be used to reduce the effect of nonsampling errors in the analysis. Some estimation methods for data observed subject to measurement error are discussed.

KEY WORDS: Measurement error; Quantiles; Regression calibration.

This symposium can be considered an exposition on current best practice in survey methodology. Claes Anderson used the term *current best methods*. We understand the term to mean not only using the best currently known procedures, but to consciously design experiments and data collection activities to improve upon current procedures. The opening address by Groves set the tone by emphasizing the necessity of quantifying the effects of measurement error in order to improve survey procedures. The first session reminded us of the importance of reporting information on measurement errors and sampling errors. Sessions three, four, and five emphasized techniques designed to reduce nonsampling errors. Sessions two, six, seven, and eight concentrated on procedures to estimate the properties of measurement error.

While several of the presentations contained illustrations of situations where observed properties of measurement and nonresponse errors led to revised data collection or processing procedures, few illustrated the use of properties of measurement error in estimation. The topics addressed in this symposium reflect the fact that the use of estimation techniques that adjust for measurement error is not common in the survey statistics community. This symposium and other similar conferences make it clear that survey statisticians are aware that the data are error prone. Also, the literature on estimation techniques for data with measurement error is now quite extensive. See Fuller (1987), Carroll, Ruppert, and Stefanski (1995), and references cited by these authors. Some of the techniques in those texts address relatively complex problems in a relatively complex way. Other procedures are relatively simple by modern computing standards.

What are the reasons that use of measurement error procedures is not common in survey statistics? One is the custom that determines how we proceed as survey statisticians. Survey statisticians assign certain responsibilities to themselves. Many of us see our tasks to be designing a survey procedure, including frame designation, sample design, instrument construction, data collection and data processing to produce a final data set complete with weights. We describe the nature of these activities in our report to the sponsor. With the exception of design, which is explicitly excluded, these are the topics of this symposium. We also

[1] Wayne A. Fuller, Professor, Department of Statistics, Iowa State University, Ames, Iowa, USA, 50011.

generally calculate estimated sampling variances or provide a procedure to do so. If we do all of these things well, we are held in high esteem by the members of our particular subgroup of the scientific community. It is the responsibility of those who analyze the data to worry about the effect of measurement error on the analyses.

Survey statisticians have not rushed to apply analysis procedures including adjustment for measurement error for additional reasons. The procedures are relatively complex and, perhaps most important, almost all measurement error estimation procedures require the specification of a model. Because sample randomization is the cornerstone of our discipline, we are reluctant to adopt procedures whose performance rests upon model assumptions. It is not that we will not adopt such procedures. Witness the adjustment for nonresponse, imputation for missing data, and small area estimation. The adjustment for the two types of missing data are now part of the responsibility that most of us accept as survey statisticians. If one examines old studies, I think one will find that the techniques for nonresponse adjustment were not always as common as they are now. At one time, many surveys simply reported the number of nonrespondents. The importance of nonresponse and the fact that survey statisticians possessed the information about the nature of nonresponse led to the current practice of nonresponse adjustment.

I feel that adjustments for response variance and response bias will ultimately be in the same category as adjustments for nonresponse. In cases where the same individuals are closely associated with both the survey operation and the analysis, techniques that recognize measurement error are more often used. See the special issue of *Statistics in Medicine* (1989), Nusser et al. (1996), and references cited by Carrol, Ruppert, and Stefanski (1995).

In the remainder of this discussion, I will suggest some procedures that might be adopted by survey statisticians as an aid to mildly sophisticated users. While not always fully efficient, the procedures provide ways to improve simple analyses such as two-way tables, regressions, and quantile estimation.

Consider the situation in which we can identify a variable or vector of variables that are of considerable interest in the study and that are measured with error. Denote this variable or vector by X_t or $\boldsymbol{X}_t$, respectively. There are other types of variables of interest. Variables of the *Y*-type are variables that might be dependent variables in an analysis containing *X* as an explanatory variable. Variables of the *Z*-type are other explanatory variables that might be included in an analysis. Examples of *Z*-variables might be age, region of country, or urbanization of residential area. A variable such as education is on the margin between *X* and Y. In many analyses, this would be an explanatory variable, but there are analyses where it might be a dependent variable.

Assume that a subsample of the original sample is selected for study of the measurement error. In one case, we are able to determine the true value of x_t in the subsample. (We consider the scalar case to simplify the presentation.) As a further subcase, assume that all variables (x, Y, Z) are observed on the subsample. Then, under the assumption that the subsample is a probability sample, the configuration is that of classical two-phase sampling and most in the survey statistics community would use one of the standard two-phase procedures. If the subsample is a small fraction of the total sample and the total sample contains many variables, one might choose an alternative estimation procedure based on techniques outlined below. With such procedures, the entire sample is available to the analyst.

Consider now a situation in which the second sample is not a subsample of the large sample, but is a separate sample. We call such a sample a calibration sample. Furthermore, assume that only the vector (x, X, Z) is observed on the subsample. In this case, it is possible to estimate the joint distribution of (x, X, Z) using two-phase methodology, but it is not possible to directly estimate the joint distribution of (Y, x, X, Z). Assume first that x has a density. Using the subsample, we estimate the cumulative

distribution function (CDF) of x, denoted by $\hat{F}_x\{\cdot\}$, and the cumulative distribution function of X, denoted by $\hat{F}_X\{\cdot\}$. Also assume that we calculate the regression of x on (X, Z). That is, we estimate the conditional expected value of x given (X, Z). Let the estimated conditional expected value be denoted by

$$\hat{x} = g\left(X, Z, \hat{\theta}\right),$$

where $\hat{\theta}$ is a vector of estimated parameters. In some cases, g will be a linear regression model, but it can be a nonlinear function. In order to estimate the conditional expected value, we must be able to treat the calibration sample as a probability sample from the population of interest.

Using the results from the calibration sample, an analysis data set is created from the large sample. The vector of variables in the analysis data set is $(\ddot{x}, \hat{x}, X, Y, Z)$, where

$$\ddot{x}_t = \hat{F}_x^{-1}\left\{\hat{F}_X\left(X_t\right)\right\} = T\left(X_t\right) \tag{1}$$

and $\hat{x}_t = g\left(X_t, Z_t, \hat{\theta}\right)$. There are alternative ways to define $\hat{F}_x^{-1}\{\cdot\}$. One method is to define $\hat{F}_x\{\cdot\}$ as a continuous function by joining the midpoints of the rises in the empirical step function CDF with line segments. A number of estimators can be constructed from data set of $(\ddot{x}, \hat{x}, X, Y, Z)$ using simple procedures. It is obvious that $\ddot{x}_t$ can be used to construct univariate statistics. It is equally clear that standard errors for, say, quantiles, calculated as if the $\ddot{x}_t$ were a sample of observations will be too small. The estimators are two-phase estimators and proper standard errors can be calculated by replication or other procedures. See Rao and Sitter (1995), Särndal and Swensson (1987), Kott (1990), Breidt and Fuller (1993), and Fuller (1996).

The $\hat{x}_t$ can be used to construct estimators of the regression of Y on x and Z. The regression need not be linear for the application of ordinary techniques to $(Y, \hat{x}, Z)$ in order to estimate the regression of Y on (x, Z). The method requires the assumption that the measurement error is independent of the error in the model. An example where the procedure is applicable is the linear model

$$Y_t = \beta_0 + x_t\beta_1 + e_t \tag{2}$$

$$X_t = x_t + u_t,$$

where (e_t, u_t, x_t) is distributed with mean $(0, 0, \mu_x)$ and covariance matrix $\text{diag}(\sigma_{ee}, \sigma_{uu}, \sigma_{xx})$.

In this case, the regression of Y_t on $\hat{x}_t = E\{x_t | X_t\}$ will give a consistent estimator of β_1.

The procedure of replacing x by $\hat{x}$ is called *regression calibration* by Carroll, Ruppert, and Stefanski (1995). They give the three steps of the procedure:

1. Estimate the regression of x on (X, Z) to obtain $\hat{x}$.
2. Replace x with $\hat{x}$ in the standard analysis.
3. Adjust the standard errors of the standard procedure.

The third step makes clear that the standard errors produced by the simple analysis treating $\hat{x}_t$ as x_t are not correct. The bias in the standard errors calculated by the standard analysis depends on the size of the calibration sample. If the calibration sample is a small fraction of the basic sample, the true standard error could be considerably larger than that calculated using standard programs applied to the basic sample. While one cannot recommend the use of the simple incorrect variance estimator, such a procedure will generally produce less biased estimators of the mean square error than does ordinary least squares. Proper estimators of variance are available in Carroll, Ruppert and Stefanski (1995), and in Fuller (1987). Also replication procedures can be used.

Example. Consider a variable X with a reliability ratio of $\kappa_{xx} = 0.85$ and the regression model

$$Y_t = \beta_0 + \beta_1 x_t + e_t, \tag{3}$$

$$X_t = x_t + u_t,$$

where $(e_t, u_t, x_t) \sim \text{NI}\left(0, \text{diag}\left[\sigma_{ee}, \sigma_{uu}, \sigma_{xx}\right]\right)$, $(\sigma_{ee}, \sigma_{xx}) = (2.00, 0.85)$, and $\beta_1 = 1.00$. Assume a simple random sample from the population. Then

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} \sim \left(\begin{pmatrix} \beta_0 + \beta_1 \mu_x \\ \mu_x \end{pmatrix}, \begin{pmatrix} 2.85 & 0.85 \\ 0.85 & 1.00 \end{pmatrix} \right)$$

and $R^2_{XY} = 0.2535$. The variance of the ordinary least squares coefficient, $\hat{\gamma}_{1\ell}$, for samples of size 1000 is

$$V\left\{\hat{\gamma}_{1\ell}\right\} = 0.00213$$

and $E\left\{\hat{\gamma}_{1\ell}\right\} = 0.85$. If $\kappa_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$ is known, we let

$$\ddot{x}_t = \bar{X} + (0.85)^{1/2}\left(X_t - \bar{X}\right)$$

$$\hat{x}_t = \bar{X} + 0.85\left(X_t - \bar{X}\right).$$

Then the sample distribution function of $\ddot{x}_t$ is an estimator of the distribution function of x_t. The $\hat{x}_t$ is the conditional expected value of x_t given X_t. The simple forms for $\ddot{x}_t$ and $\hat{x}_t$ are appropriate only if the distribution of (x_t, u_t) is normal. Note that the variance of $\hat{x}_t$ is less than the variance of the population of true values.

If κ_{xx} is estimated from an independent sample of 100 as the correlation between replicate measurements on X, the variance of $\hat{\kappa}_{xx}$ is, approximately,

$$V\left\{\hat{\kappa}_{xx}\right\} \doteq n^{-1}\left(1 - \kappa_{xx}^2\right)^2 = 0.00077.$$

The regression of Y_t on $0.85X_t$ has a variance of $(0.85)^{-2}(0.00213)$ and an expected value of 1.00 for a sample of size 1000. If the reliability ratio is estimated from an independent sample of 100, the approximate variance of the estimator of β_1 obtained by regressing Y_t on $\hat{x}_t = \bar{X} + \hat{\kappa}_{xx}\left(X_t - \bar{X}\right)$ is

$$V\left\{\tilde{\beta}_1\right\} = V\left\{\hat{\kappa}_{xx}^{-1}\hat{\gamma}_{1\ell}\right\} \doteq (0.85)^{-2}(0.00213 + 0.00077)$$

$$= 0.00401.$$

In this case, estimating κ_{xx} increases the variance 36% relative to the variance with known κ_{xx}. The mean square error of the ordinary least squares estimator is

$$E\left\{\left(\hat{\gamma}_{1\ell} - \beta_1\right)^2\right\} = (0.15)^2 + 0.00213 = 0.02463.$$

Thus, the estimator $\tilde{\beta}_1$ is essentially unbiased with a MSE one sixth that of ordinary least squares.

The same general approach is applicable when binomial variables, measured as binomial variables are included in the vector $\boldsymbol{X}_t$. There is no need to create the variable $\ddot{x}_t$ for binomial variables because the marginal distribution of x_t is the same as that of X_t. The regression prediction procedure requires the error in the equation and measurement error in the dependent variable to be independent of the measurement error in the explanatory variable. If x_t and X_t are binomial random variables, there is a negative correlation between u_t and X_t. However, the regression of Y_t on $E\left\{x_t | X_t\right\}$ yields a consistent estimator of β_1 if e_t of model (2) is independent of X_t.

If both the dependent and independent variables are binomial random variables, an extension of the approach is required. Consider the model

$$y_t = \beta_0 + x_t\beta_1 + q_t,$$

$$Y_t = y_t + e_t,$$

$$X_t = x_t + u_t,$$

$y_t \sim B_i\left(1, p_y\right)$, $Y_t \sim B_i\left(1, p_Y\right)$, $x_t \sim B_i\left(1, p_x\right)$, $X_t \sim B_i\left(1, p_X\right)$, $E\left\{e_t\right\} = 0$, and $P\left\{X = i | \mathrm{x} = \mathrm{j}\right\} = \kappa_{ij}$. Because the variables $\left(Y, X, x\right)$ are constrained to $\left(0, 1\right)$, there is a negative covariance between u_t and x_t. It follows that, in general, there is a covariance between e_t and u_t even though κ_{ij} does not depend on Y_t. Thus, for the pair $\left(Y_t, X_t\right)$ of binomial variables observed with error, it seems necessary to construct the expected values for all possible configurations. Let Z be the r-dimensional multinomial row vector defined by the $\left(Y, X\right)$ pairs $\left[(1, 1), (1, 0), (0, 1), (0, 0)\right]$. That is, $\boldsymbol{Z}_t = \left(1, 0, 0, 0\right)$ if $\left(Y_t, X_t\right) = \left(1, 1\right)$ and $\boldsymbol{Z}_t = \left(0, 1, 0, 0\right)$ if $\left(Y_t, X_t\right) = \left(1, 0\right)$. Let z_t be the corresponding vector defined by the $\left(y_t, x_t\right)$ pairs. (The variable Y_t may also be measured with error.) Given response probabilities, one can define the expected value of z_t given $\boldsymbol{Z}_t$.

We now consider a particular example of the estimation of the distribution function of a variable observed subject to measurement error. Nusser et al. (1996) is a particular application to the estimation of the distribution of usual intakes for a dietary component. A simplified version of the empirical model used by Nusser et al. (1996) is

$$
\begin{aligned}
Y_{ti} &= g^{-1}\left(X_{ti}\right), \\
X_{ti} &= x_t + u_{ti}, \\
\begin{pmatrix} x_t \\ u_{ti} \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & 0 \\ 0 & \sigma_{uu} \end{pmatrix}\right),
\end{aligned}
\tag{4}
$$

and Y_{ti} is the observed intake on day i for the t-th individual. The variable u_{ti} represents day-to-day variability in intakes plus the response variance. In this case, the sampling (day-to-day) variance is judged to be the most important. The usual intake is

$$
y_t = E\left\{Y_t | \mathrm{x}_t\right\} = E\left\{g\left(x_t + u_{ti}\right) | \mathrm{x}_t\right\}. \tag{5}
$$

Denote the conditional expected value function by $h(\quad)$. Thus,

$$
y_t = h\left(x_t\right). \tag{6}
$$

Equation (6) defines the distribution of y_t because $x_t \sim \mathrm{N}\left(0, \sigma_{xx}\right)$. In practice, we cannot always expect a single transformation of the observations, $g\left(Y_t\right)$ to produce observations, X_t, that can be decomposed into two normal components. However, in the application to dietary components, transforming the original observations to approximate normality seems to result in data for which model (4), or a modest extension of model (4), can be used as a working model.

The primary data in the nutrition studies are reported food intakes for a day. On the basis of the composition of the foods, the reported intake of a component such as protein is constructed. The distribution of individuals' long run average intake, called *usual intake* is of interest. These data are characterized by large day-to-day variation. About two thirds of the observed variation in daily intakes is due to within-person variation and about one third is due to person-to-person variability. If two observations are made on each individual, about 50% of the variation in the individual means is due to within person variability. Even in studies with four observations per individual, the within person variability represents one third of the observed variability in individual means.

In the Nusser et al. (1996) procedure, the transformation g is constructed in two steps. First, a power is chosen that moves the observations closer to normality. Then a spline function is used to approximate the transformation of the observed quantile function into that of the quantiles of the normal distribution.

The model in the transformed variables is estimated by analysis of variance procedures. Thus,

$$\hat{\sigma}_{uu} = [n(r-1)]^{-1} \sum_{t=1}^{n} \sum_{i=1}^{r} \left(X_{ti} - \overline{X}_{t.} \right)^2 ,$$

$$\hat{\sigma}_{xx} = (n-1)^{-1} \sum_{t=1}^{n} \left(\overline{X}_{t.} - \overline{X}_{..} \right)^2 - \hat{\sigma}_{uu} ,$$

where

$$\overline{X}_{t.} = r^{-1} \sum_{i=1}^{r} X_{ti} \text{ and } \overline{X}_{..} = n^{-1} \sum_{t=1}^{n} \overline{X}_{t.} .$$

Then the transformation h is estimated to define the estimated distribution of usual intakes. In practice, different distributions are estimated for different subgroups. (The variables Z define subgroups.) The data set required for analyses by a user interested in looking at relationships with usual intakes as an explanatory variable requires the estimated expected value. An example might be the possible relationship between cancer incidence and fat intake.

Figure 1 contains a plot of the estimated expected value of y_t given (Y_{t1}, Y_{t2}) plotted against $0.5(Y_{t1} + Y_{t2}) = \overline{Y}_t$ for energy for a two-day per individual data set for 737 women aged 25-50 from the 1985 Continuing Survey of Food Intakes by Individuals conducted by the U.S. Department of Agriculture (1987). The dots do not fall on a line because the expected value is not a simple function of $\overline{Y}_t$. The plotted values are $h(\hat{x}_{t.})$, where

$$\hat{x}_{t.} = \overline{X}_{..} + (\hat{\sigma}_{uu} + \hat{\sigma}_{xx})^{-1} \sigma_{xx} (\overline{X}_{t.} - \overline{X}_{..}) .$$

Figure 2 contains a plot of the representative values for the same data set where the representative values are

$$T(Y_{t1}, Y_{t2}) = h(\tilde{x}_{t.}) ,$$

and

$$\tilde{x}_{t.} = \overline{X}_{..} + \left[(\hat{\sigma}_{uu} + \hat{\sigma}_{xx})^{-1} \sigma_{xx} \right]^{1/2} (\overline{X}_{t.} - \overline{X}_{..}) .$$

The sample cumulative distribution function of the representative values is an estimator of the cumulative distribution function of usual intakes. The dots for the estimated expected value have a smaller slope than the data for the representative values. In the x-scale the variance of the representative values is $\hat{\sigma}_{xx}$, while the variance of the estimated expected values is $(\hat{\sigma}_{uu} + \hat{\sigma}_{xx})^{-1} \hat{\sigma}_{xx}^2$. The relationship between the estimated expected value and the observations is curved because the original distribution is skewed. The functions for other dietary components, such as vitamins, is more curved because the distributions are more skewed.

Carriquiry, Goyeneche, and Fuller (1996) have extended the estimation of the usual intake distributions to the bivariate case. They employ a multiple step transformation of $\boldsymbol{Y}_{ti} = \left(Y_{1ti}, Y_{2ti}\right)$ into $\boldsymbol{X}_{ti} = \left(X_{1ti}, X_{2ti}\right)$ such that $\boldsymbol{X}_{ti} = g\left(\boldsymbol{Y}_{ti}\right)$ is approximately normally distributed. Figure 3 is a plot of the 50% and 90% contours for the estimated joint distribution of total calories and calories from fat. The contours are for the original observations and for the usual intake distribution for the sample of 4,734 women aged 20-59 from the 1994 Continuing Survey of Food Intakes by Individuals. The 50% contour for the original data is not greatly different from the 90% contour for the usual intake distribution. The lines in the plots are the estimated conditional means of total calories given calories from fat for the two distributions. The two lines would differ more if the correlation structure of Σ_{uu} was less similar to that of Σ_{xx}.

ACKNOWLEDGEMENTS

This research was partly supported by Research Agreement No. 58-3198-9-032 and Cooperative Agreement 58-3198-2-006 between the Agricultural Research Service, U.S. Department of Agriculture, and the Center for Agricultural and Rural Development, Iowa State University. I thank Kevin Dodd and Juan Jose Goyeneche for the analyses of the nutrition data.

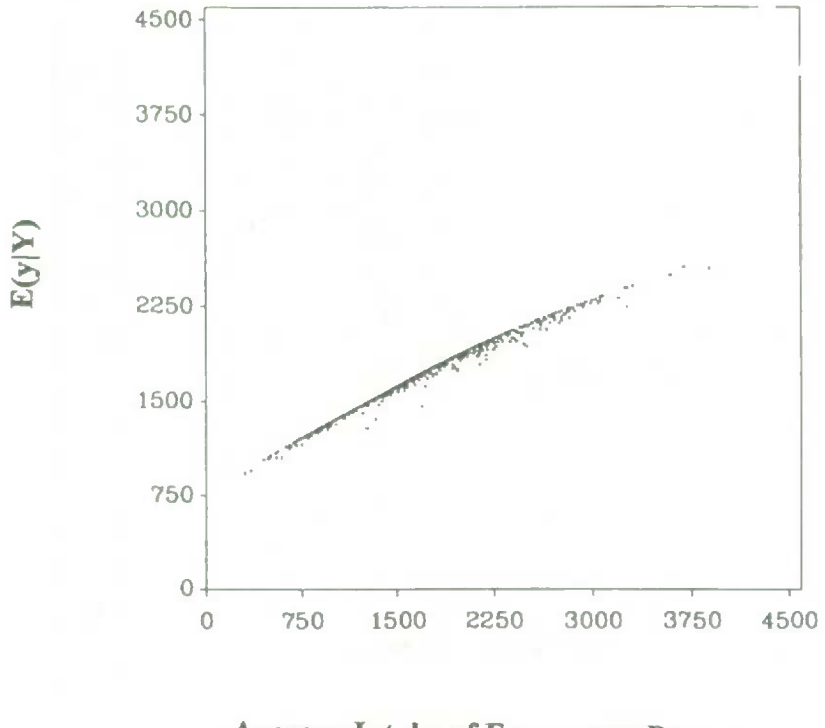


Figure 1. Estimated expected value of usual intake of energy plotted against two-day mean intake.

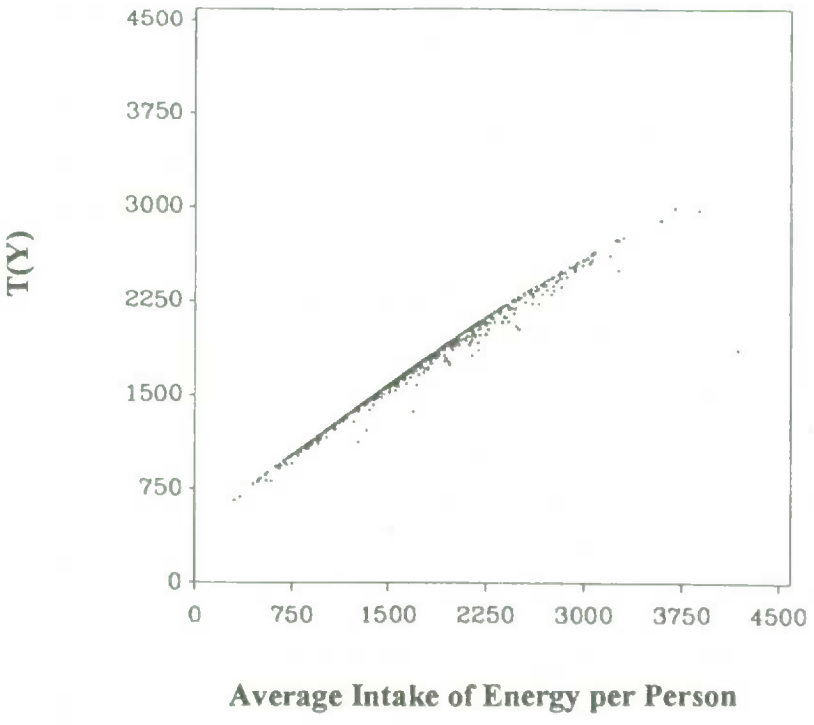


Figure 2. Representative values of energy usual intake plotted against two-day mean intake.

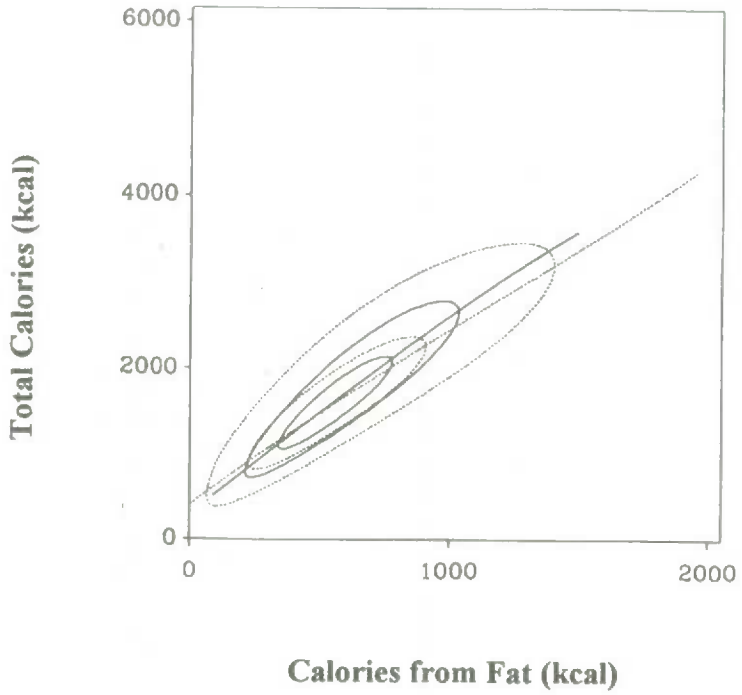


Figure 3. Contours for the bivariate distribution of total calories and calories from fat. The dashed lines are 50% and 90% contours for one-day intakes. The solid lines are 50% and 90% contours for usual intakes.

REFERENCES

Breidt, F. J. and Fuller, W. A. (1993). Regression weighting for multiphase samples. *Sankhyā Services B* **55**, 297-309.

Carriquiry, A. L. Goyeneche, J. J. and Fuller, W. A. (1996). Estimation of bivariate usual intake distributions. Unpublished report. Iowa State University, Ames, Iowa.

Carrol, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman and Hall, New York.

Chua, T. C. and Fuller, W. A. (1987). A model for multinomial response error applied to labor flows. *J. Amer. Statist. Assoc.* **82**, 46-51.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. John Wiley, New York.

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. *Proc. ASA Section on Survey Research Methods*, Washington, D.C., 212-217.

Fuller, W. A. (1996). Replicate variance estimation for two-phase sample. Unpublished manuscript. Iowa State University, Ames, Iowa.

Kott, P. S. (1990). Variance estimation when a first-phase area sample is restratified. *Survey Methodology* **16**, 99-103.

Lin, L. I.-K. and Vonesh, E. F. (1989). An empirical nonlinear data-fitting approach for transforming data to normality. *The American Statistician*, **43**, 237-243.

Mendelsohn, J. and Rice, J. (1982). Deconvolution of Microfluorometric histograms with ph{B}-splines. *J. Amer. Statist. Assoc.* 77, 748-753.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*. To appear.

Nusser, S. M., Fuller, W. A. and Guenther, P. M. (1996). Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-hour food intake data. In *Survey Measurement and Process Quality*, eds. L. Lyberg, M. Collins, E. DeLeeuw, C. Dippo, W. Schwartz, and D Trewn. John Wiley, New York.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Statist.* **10**, 153-165.

Rao, J. N. K. and Shao, J. (1966). On balanced half-sample variance estimation in stratified random sampling. *J. Amer. Statist. Assoc.* To appear.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**, 453-460.

Särndal , C. E. and Swensson, B. (1987). A general view of estimation for two-phases or selection with application to two-phase sampling and nonresponse, *Int'l. Statist. Review* **55**, 279-294.

Särndal , C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Stefanski, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statistics and Probability Letters* **9**, 229-235.

Stefanski, L. A. and Bay, J. M. (1996). Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. *Biometrika*. To appear.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169-184.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

CLOSING REMARKS

Proceedings of Statistics Canada Symposium 96
Nonsampling errors
November 1996

CLOSING REMARKS

David Binder[1]

I would like to add some personal remarks on the topic of nonsampling errors. First of all, what is *nonsampling error* as opposed to *sampling error*. Perhaps this terminology is poor. After all, when we talk about *response errors* versus *nonresponse errors*, we mean errors due to people who don't respond. Why wouldn't nonsampling error be errors due to not being in the sample? To non-statisticians, this would make sense. When we pick up a typical book on sampling theory, we start off with a population of a finite number of units, a sample is selected and there is a discussion of estimation, and other aspects of the sampling strategy. It is then explained that anything that differs between the true value and our estimators is the sampling error. We statisticians think of nonsampling errors as everything else. However, are all errors, except for the particular ones due the sampling, nonsampling errors? For example, if, while driving, you misjudge your left turn and have an accident, is this a nonsampling error? We need some clearer, well-understood definitions.

The field of nonsampling errors, even in the statistical sense, is large and we tend to define it as a collection of subfields. We talk about frame errors, non-response errors, response errors, process errors, and so on. The fact is that most of the processes that we go through in survey-taking involve some kind of human intervention and humans are error prone. What we have been discussing during this symposium are those errors due mostly to human errors. For this reason, it is advisable to consult with experts in human behaviour. Human cognitive processes can be a guide to a better understanding of the causes of nonsampling errors.

In 1978, Statistics Canada published 'A Compendium of Methods of Error Evaluation in Censuses and Surveys'. One might ask how far we have progressed over the last 20 years. Looking at the main topics in the 'Table of Contents', after the introduction, we have sections on coverage, response, non-response, coding, data capture, edit and imputation and others. It is interesting to see that we knew the problems back in 1978 and we had some ideas on how to measure and evaluate these errors, but when we compare the contents with this symposium, it is clear we have made substantial progress. We still have much to learn.

What has become obvious during the sessions is that there is a very strong interaction among the survey management, the process of trying to measure its errors and evaluate the errors, and the necessary feedback. Often during this symposium, we have asked about the cost-benefits of measuring and evaluating the nonsampling errors. However, the problem is that we do not ask many questions about the nonsampling errors unless things go wrong. A case in point is the Smoking Survey discussed by Lecily Hunter, where the data seem to indicate something that was counter- intuitive to the expert, so the questions arose about the sources of nonsampling errors. We must be sensitive to these needs, particularly in these days of limited budgets. It is too easy to dismiss the need to measure these errors, but, in fact, in the case where when the estimates appear questionable, we need to be able to answer the concerns of the users.

Recently there has been discussion in the press about what went wrong with the pre-election polls for the U.S. presidential race. What happened was that there was a much wider gap between Clinton and Dole in the polls than what took place in the actual vote, and there have been questions about what could have been the cause of this. Not only were there questions on the appropriateness of non-probability sampling, but also on the various sources of sampling errors. On the other hand, for the polling that took place before the Quebec referendum October of 1995, the results of the polls were almost identical to the final results. Here, there was no general concern in the press. The point is that concerns over nonsampling errors arise when these errors are appear to be large.

---

[1]David Binder, Director, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Another example is the issue of measurement and adjustment for Census undercoverage. If the coverage appears very poor, there is a stronger case for adjusting the counts, since the nonsampling errors are excessive. On the other hand, if the errors are small, one has to think what additional errors are being added by adjusting for the nonsampling errors. Admittedly, the issues here are quite complex and I am focussing on only one aspect.

Many other examples could be cited. Survey managers must carefully assess the benefits of measuring the nonsampling errors against its costs, taking into account the risks associated with obtaining data that are questionable.

I would now like to thank the organizers of this symposium for having run such a highly successful meeting. In total, there are probably around 30 people who were involved in the organization in one way or another in this symposium. The main organizing committee, headed by Eric Rancourt, included Ann Brown, Jane Burgess and Johane Dufour. We had a number of subcommittees. Subcommittee chairs included Michelle Simard who looked after logistics, Josée Morel who looked after the round table discussions; Hew Gough who looked after the workshops and Lynn Savage and Micheline Sabourin who provided many of the support functions over the last few months. Much of the correspondence was handled by Sophie Dionne and Guylaine Dubreuil.

Next year's symposium is entitled 'New Directions in Surveys and Censuses'. Jack Gambino is the Chair of the Organizing Committee for that conference.

Finally, again, I would like to thank all the participants, all the Session Chairs, all the people who contribute to the discussion and all of you for helping to make this such a successful conference.