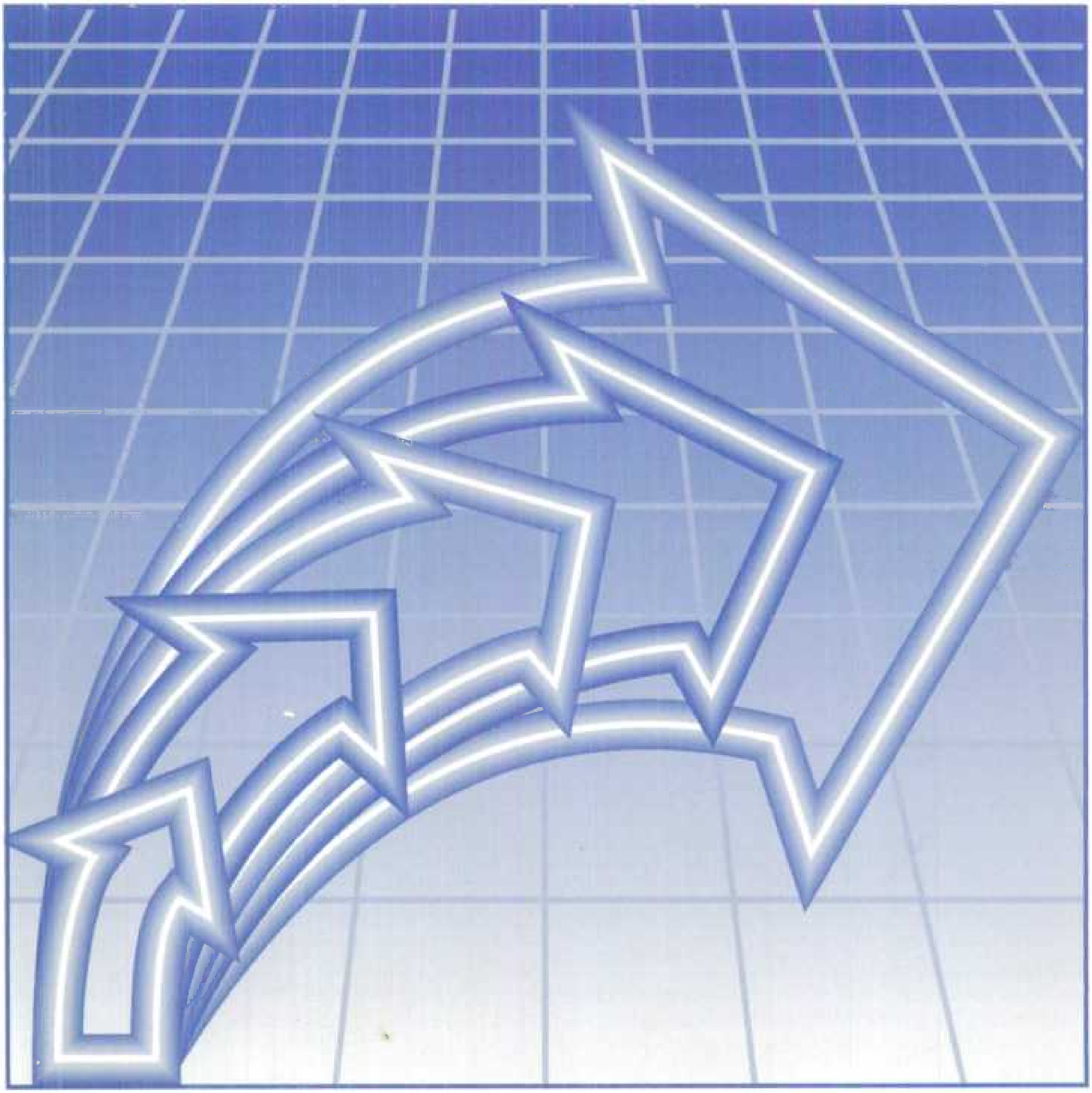
November 1992

SYMPOSIUM 92

Design and Analysis of Longitudinal Surveys

PROCEEDINGS



Statistics Canada Statistique Canada

Canada

SYMPOSIUM 92

Design and Analysis of Longitudinal Surveys

November 2-4, 1992

Ottawa, Ontario, Canada

PROCEEDINGS

August 1993

Symposium 92 Organizing Committee

John Armstrong Nancy Darcovich Pierre Lavallée

Published by authority of the Minister
responsible for Statistics Canada

PREFACE

Symposium 92 was the ninth in the annual series of conferences on issues in survey methodology sponsored by Statistics Canada. Each year the symposium focuses on a particular theme. The 1992 theme was design and analysis of longitudinal surveys.

The 1992 symposium was attended by more than 400 persons, representing nine countries, who met over three days in the Simon Goldberg Conference Centre in Ottawa to listen to experts from government agencies, universities and private industry. A total of 31 papers were presented. All papers were invited. Aside from translation and formatting, the papers submitted by the authors have been reproduced in these proceedings.

The organizers of Symposium 92 would like to acknowledge the contributions of the many persons involved in the preparation of this volume.

Naturally, the presenters at Symposium 92 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 Christine Larabie and Carmen Lacroix with the assistance of Myra Kent. Papers were translated by Michel Charuest, Pierre Desautels, Maryse Montpetit and Josée René de Cotret. Proofreading was done by a number of Statistics Canada methodologists: Yanick Beaucage, Jean-René Boudreau, René Boyer, Marcel Bureau, James Croal, André Cyr, Sylvie DeBlois, Johanne Denis, Johane Dufour, Gerrit Faber, Lyne Guertin, Michael Hidiroglou, Wisner Jocelyn, Guy Laflamme, Danielle Lalande, Normand Laniel, Danielle Lebrasseur, Josée Morel, Christian Nadeau, Stephen Rathwell, Martin Renaud, Laurent Roy, Craig Seko, Michelle Simard, Hélène St-Jean, Pierre St-Martin, Alain Théberge, Brad Thomas, Jocelyn Tourigny and Johanne Tremblay. Production was co-ordinated by Christine Larabie.

Statistics Canada's tenth annual symposium, "International Conference on Establishment Surveys", was held June 27-30, 1993 in Buffalo, New York. The eleventh annual symposium will be held in the fall of 1994. The theme will be re-engineering for statistical agencies.

Symposium 92 Organizing Committee

August 1993

STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data

1985 - Small Area Statistics

1986 - Missing Data in Surveys

[illegible] – 1987 - Statistical Uses of Administrative Data

1988 - The Impact of High Technology on Survey Taking

(R/L [illegible] Aug [illegible]) CSA [illegible] Ottawa

1989 - Analysis of Data in Time

? 1990 - Measurement and Improvement of Data Quality

R/L Vancouver – 1991 - Spatial Issues in Statistics

(R/L [illegible] July 1992) IACA [illegible]

1992 - Design and Analysis of Longitudinal Surveys

1993 - International Conference on Establishment Surveys (Buffalo)

1994 - Re-engineering for Statistical Agencies

(R/L D April 1994 AACCR [illegible])

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

Use the order form on this page to order additional copies of the proceedings of Symposium 92: Design and Analysis of Longitudinal Surveys. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 92 PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 6th 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 - Symposium 92 Proceedings").

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	____ @ $10 EACH
1987 -	Statistical Uses of Administrative Data - FRENCH	____ @ $10 EACH
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	____ @ $12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	____ @ $10 EACH
1989 -	Analysis of Data in Time - BILINGUAL	____ @ $15 EACH
1990 -	Measurement and Improvement of Data Quality - ENGLISH	____ @ $18 EACH
1990 -	Measurement and Improvement of Data Quality - FRENCH	____ @ $18 EACH
1991 -	Spatial Issues in Statistics - ENGLISH	____ @ $20 EACH
1991 -	Spatial Issues in Statistics - FRENCH	____ @ $20 EACH
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH	____ @ $22 EACH
1992 -	Design and Analysis of Longitudinal Surveys - FRENCH	____ @ $22 EACH

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 ______________

Please note: Each Symposium 92 registrant not employed by Statistics Canada receives one free copy of the Symposium 92 Proceedings.

DESIGN AND ANALYSIS OF LONGITUDINAL SURVEYS

TABLE OF CONTENTS[1]

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

SESSION 4: Using Longitudinal Structure in Estimation
Chairperson: C. Särndal, Université de Montréal

SESSION 5: Longitudinal Studies in Health Research
Chairperson: D. Krewski, Health and Welfare Canada

SESSION 6: General Applications I
Chairperson: G. Hole, Statistics Canada

SESSION 7: General Applications II
Chairperson: P. Lavallée, Statistics Canada

SESSION 8: Data Analysis I

Chairperson: D. Kasprzyk, National Center for Education Statistics, U.S.A.

SESSION 9: Data Analysis II

Chairperson: D. Binder, Statistics Canada

SESSION 10: Quality Issues

Chairperson: M. Colledge, Statistics Canada

SPECIAL INVITED LECTURE

Chairperson: J.N.K. Rao, Carleton University

OPENING REMARKS

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

OPENING REMARKS

G.J. Brackstone[1]

On behalf of Statistics Canada, welcome to Symposium 92, welcome to Ottawa, and for many of you, welcome to Canada. This is number nine in the annual series of Methodology Symposia held at Statistics Canada. In organizing these Symposia we have usually benefitted from help and support from other organizations. On this occasion, we are very pleased that the Laboratory for Research in Statistics and Probability of Carleton University and the University of Ottawa and the Environmental Directorate of Health and Welfare Canada are sharing with us the sponsorship of this Symposium.

As I said, this is the ninth in a series of Symposia which have covered a range of topics of relevance to survey methodologists, statistical analysts, and government statisticians. The full list of topics of previous Symposia was printed in the brochure so I won't read them all out. The topic for this year's Symposium is Design and Analysis of Longitudinal Surveys. Before saying a little about the reasons for the choice of this topic, let me say a few things about our motivation in sponsoring these conferences.

I could say that our objectives are to provide a forum for the discussion and exchange of ideas; to contribute to the discipline of survey methodology by fostering professional co-operation; to bring together practitioners and theoreticians from diverse backgrounds who bring different experiences and perspectives to problems of common interest. An so on and so on. In other words, the advancement of statistical science and survey practice. All that is absolutely true. Those are admirable objectives and we hope this Symposium will move us towards them. But let's be honest about our motivation: it is not pure altruism; there is a large degree of self-interest.

The first element of self-interest is that we have real practical problems and challenges that we are facing in the programs of Statistics Canada and we want help from the best and the most experienced in these problems. And I am pleased to see that many of the best and most experienced in longitudinal surveys are here this week. We are very grateful to those of you who have come here, some travelling long distances, to share your experiences with us. I know we will benefit; I hope you will too.

A second element of self-interest in organizing these Symposia is that they give many of our staff the opportunity to improve and broaden their understanding of issues related to their work. Far more of our staff can come down here and listen to world-class experts than could ever hope to attend professional conferences held elsewhere.

So there is a strong element of self-interest in our sponsorship of these Symposia but, judging by the support and attendance we always have, I feel confident that many others are benefitting too. Let me turn now to the topic for this Conference: Design and Analysis of Longitudinal Surveys.

Last year our topic was Spatial Issues in Statistics. So we have moved from the space dimension to the time dimension. Last year we dealt with geographic structures, with spatial and areal concerns. Our focus was on latitude and longitude. This year we are focussing just on the longitude - or at least on longitudinal surveys. Yet we are giving ourselves a lot of latitude in how we interpret longitudinal.

In some ways this Symposium can be seen as the third wave of a series of conferences on related topics. The Conference on Panel Surveys in Washington in 1986 began the series. That conference resulted in the Wiley book on Panel Surveys. Three years later, in 1989, our Symposium here focussed on the Analysis of Data in

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26-J, R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Time and included a lot of material related to longitudinal surveys. Now three years later again, we are back to Longitudinal Surveys. There has been some attrition over the three waves but I see that many of the participants in the 1986 conference are here to give us the benefit of their experience again.

Survey designs involving measurements on the same sample at different points in time are not new. Sample overlap has been a component of the design of repeated surveys for many years. Use of sample overlap has been motivated by the need to have efficient estimates of change, and by operational considerations such as the cost of introducing new respondents to a survey.

Recent years have seen increasing interest in the design of surveys for which repeated measurement of characteristics for the same set of units is a survey objective in itself, rather than just a design choice. This interest has been fostered by the recognition that many important questions of economic and social policy require information about the impact of dynamic processes on individual persons or businesses, and not just on their aggregate or macro effects. Prime examples which we will hear more about over the next few days are poverty and health status, where understanding how and why people move in and out of poverty or poor health is probably much more important to policy and program design than just knowing how the overall level of poverty or poor health in the population is changing. Understanding these dynamics requires measurements for the same units over extended time periods.

Such measurements require, it seems, either a longitudinal survey design to obtain responses from the same units on different occasions, or a retrospective survey. In fact, any longitudinal survey has a certain retrospective element in it, so that, in practice, survey designs have to compromise between these two approaches through the choice of an optimum recall period. But we shouldn't forget that there is also a third important possibility for obtaining longitudinal data and that is through longitudinal linkage of administrative records. We have some experience of that in Statistics Canada.

In the past decade large-scale multi-wave longitudinal surveys of socio-economic conditions and household behaviour have been established in several European countries. In the United States, in 1983, the Survey of Income and Program Participation began collection of information related to income and eligibility and participation in various income support programs.

Having had the privilege of serving on a Census Bureau advisory committee on SIPP, along with our keynote speaker I should add, I have witnessed first hand the tremendous challenges faced by the Census Bureau in that undertaking. I used to be smug that we didn't have to face such problems at Statistics Canada. But now we do - and that is partly why this Symposium is being held. I hope we have learned a lot from the pioneering work of the Census Bureau on SIPP.

Our modest start in longitudinal surveys began with the Labour Market Activity Survey in the late 1980s. A more ambitious survey, the Survey of Labour and Income Dynamics, or SLID, will be launched next year. This survey will follow individuals and families for five or six years collecting information about labour market experiences, income, and family circumstances.

We are also planning a major new longitudinal survey of health that will be in the field for the first wave of data collection in 1994. The survey is intended to provide information about the effects over time of socio-economic and lifestyle factors on individual well-being and use of the health care system.

Longitudinal surveys give rise to many methodological issues. The program that has been arranged for the next three days reflects the diversity of these issues as well as the international interest in longitudinal surveys. Papers will address questions of data collection, sample selection and weighting, non-response, analysis of longitudinal data (an area that opens up methods not traditionally used in survey applications), and the quality of data from longitudinal surveys. We will also hear presentations on applications of longitudinal survey methods.

The program includes speakers from the United Kingdom, France, Germany and Luxembourg as well as Canada and the United States. The private sector, academia and government are all represented.

So, again, thank you all for coming to share your expertise and experience with us. I hope by the end of the Symposium you will feel that you have benefitted as much as we at Statistics Canada undoubtedly will.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

PANEL SURVEYS: ADDING THE FOURTH DIMENSION

G. Kalton[1]

ABSTRACT

Surveys across time can serve many objectives. The first half of the paper reviews the abilities of alternative survey designs across time - repeated surveys, panel surveys, rotating panel surveys and split panel surveys - to meet these objectives. The second half concentrates on panel surveys. It discusses the decisions that need to be made in designing a panel survey, the problems of wave nonresponse, time-in-sample bias and the seam effect, and some methods for the longitudinal analysis of panel survey data.

KEY WORDS: Panel surveys; Rotating panel surveys; Repeated surveys; Panel attrition; Time-in-sample bias; Seam effect; Longitudinal analysis.

1. INTRODUCTION

Survey populations are constantly changing over time, both in composition and in the characteristics of their members. Changes in composition occur when members enter the survey population through birth (or reaching adulthood), immigration, or leaving an institution (for a noninstitutional population) or leave through death, emigration, or entering an institution. Changes in characteristics include, for example, a change from married to divorced, or from a monthly income of $2,000 to one of $2,500. These population changes give rise to a range of objectives for the analysis of survey data across time. This paper reviews survey designs that produce the data needed to satisfy these various objectives.

The paper is divided into two parts. The first part contains a review of the general issues involved in conducting surveys across time, including the objectives of such surveys and the types of survey design that may be employed. This part is to be found in Section 2. The second, and main, part of the paper discusses one particular survey design, a panel survey that follows the same sample of units through time. The considerations involved in designing, conducting, and analyzing a panel survey are reviewed in Section 3. Section 4 provides some concluding remarks.

2. SURVEYS ACROSS TIME

This section presents an overview of analytic objectives across time, of designs for surveys across time, and of the extent to which different designs can satisfy the various objectives. The discussion relies heavily on Duncan and Kalton (1987), which contains a more detailed treatment of these issues.

Changes in population characteristics and composition over time lead to a variety of objectives for surveys across time. These objectives include the following:

(a) The estimation of population parameters (*e.g.*, the proportion of the population in poverty) at distinct time points;

(b) The estimation of average values of population parameters across time (*e.g.*, the daily intake of iron averaged across a year);

[1] G. Kalton, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850.

(c) The estimation of net changes, that is changes at the aggregate level (*e.g.*, the change in the proportion of unemployed from one month to the next);

(d) The estimation of gross changes and other components of individual change (*e.g.*, the proportion of persons who were in poverty one year and were not in poverty in the following year);

(e) The aggregation of data for individuals over time (*e.g.*, the summation of twelve monthly incomes to give annual income);

(f) The collection of data on events occurring in a specified time period (*e.g.*, becoming unemployed), and on their characteristics (*e.g.*, duration of spells of unemployment);

(g) The cumulation of samples over time, especially samples of rare populations (*e.g.*, women who become widowed);

(h) The maintenance of a sample of members of a rare population that was identified at one point of time (*e.g.*, scientists and engineers identified from a large-scale survey at one point of time).

A number of survey designs have been developed to provide the data needed to address these objectives. These designs are:

- *Repeated survey*. A repeated survey is a series of separate cross-sectional surveys conducted at different time points. No attempt is made to ensure that any of the same elements are sampled for the individual surveys. The elements are sampled from a population defined in the same manner for each individual survey (*e.g.*, the same geographical boundaries and age-limits) and many of the same questions are asked in each survey.

- *Panel survey*. A panel survey collects the survey data for the same sample elements at different points of time.

- *Repeated panel survey*. A repeated panel survey is made up of a series of panel surveys each of a fixed duration. There may be no overlap in the time period covered by the individual panels, for instance one panel may start only as (or after) the previous one ends, or there may be an overlap, with two or more panels covering part of the same time period.

- *Rotating panel survey*. Strictly, a rotating panel survey is equivalent to a repeated panel survey with overlap. Both limit the length of a panel, and have two or more panels in the field at the same time. However, it seems useful to distinguish between the two designs because they have different objectives. Rotating panel surveys are widely used to provide a series of cross-sectional estimates and estimates of net change (*e.g.*, of unemployment rates and changes in such rates), whereas repeated panel surveys with overlaps also have a major focus on longitudinal measures (*e.g.*, durations of spells of unemployment). In consequence, repeated panel surveys tend to have longer durations and have fewer panels in operation at any given time than rotating panel surveys.

- *Split panel survey*. A split panel survey is a combination of a panel survey and a repeated survey or rotating panel survey.

The choice of design in a particular case depends on the objectives to be satisfied. Some designs are better than others for some objectives but poorer for other objectives. Some designs cannot satisfy certain objectives at all. For a detailed discussion, see Duncan and Kalton (1987).

The strength of a repeated survey is that it selects a new sample at each time point, so that each cross-sectional survey is based on a probability sample of the population existing at that time. A panel survey is based on a sample drawn from the population existing at the start of the panel. Although attempts are sometimes made to add samples of new entrants to a panel at later time points, such updating is generally difficult to do and is done imperfectly. Moreover, nonresponse losses from a panel as it ages heighten concerns about nonresponse bias when the panel sample is used to estimate parameters for later time points. For these reasons, repeated surveys are stronger than panel surveys for producing cross-sectional and average cross-sectional estimates (objectives (a) and (b)). With average cross-sectional estimates, another factor to be considered is the correlation between the values of the survey variables for the same individual at different time points. When this correlation is positive, as it generally is, it increases the standard errors of the average cross-sectional estimates from a panel survey. This factor thus also favours repeated surveys over panel surveys for average cross-sectional estimates.

The superior representation of the samples for a repeated survey at later time points also argues in favour of a repeated survey over a panel survey for estimating net change (assuming that the interest in net change relates to changes in both population composition and characteristics). However, in this case the positive correlations of the values of the survey variables for the same individuals across time decreases the standard errors of estimates of net change from a panel survey. Hence the presence of this correlation operates in favour of the panel design for measuring net change.

The key advantages of the panel design are its abilities to measure gross change, and also to aggregate data for individuals over time (objectives (d) and (e)). Repeated surveys are incapable of satisfying these objectives. The great analytic potential provided by the measurement of individual changes is the major reason for using a panel design.

Repeated surveys can collect data on events occurring in a specified period and on durations of events (*e.g.*, spells of sickness) by retrospective questioning. However, retrospective questioning often introduces a serious problem of response error in recalling dates, and the risk of telescoping bias. A panel survey that uses a reference period for the event that corresponds to the interval between waves of data collection can eliminate the telescoping problem by using the previous interview to bound the recall (*i.e.*, an illness reported at the current interview can be discarded if it had already been reported at the previous one). Similarly, a panel survey can determine the duration of an event from successive waves of data collection, limiting the length of recall to the interval between waves.

Repeated data collections over time can provide a vehicle for accumulating a sample of members of a rare population, such as persons with a rare chronic disease or persons who have recently experienced a bereavement. Repeated surveys can be used in this manner to generate a sample of any form of rare population. Panel surveys, however, can be used to accumulate only samples of new rare events (such as bereavements) not of stable rare characteristics (such as having a chronic disease). If a sample of members with a rare stable characteristic (*e.g.*, persons with doctoral degrees) has already been identified, a panel survey can be useful for maintaining the sample over time, with suitable supplementation for new entrants at later waves (for an example, see Citro and Kalton 1989).

Rotating panel surveys are primarily concerned with estimating current levels and net change (objectives (a) and (c)). As such, elements are usually retained in the panel for only short periods. For instance, sample members remain in the monthly Canadian Labour Force Survey for only six months. The extent to which individual changes can be charted and aggregation over time can be performed is thus limited by the short panel duration. A special feature of rotating panel surveys is the potential to use composite estimation to improve the precision of both cross-sectional estimates and estimates of net change (see Binder and Hiridoglou 1988).

By combining a panel survey with a repeated survey or a rotating panel survey, a split panel survey can provide the advantages of each. However, given a constraint on total resources, the sample size for each component is necessarily smaller than if only one component had been used. In particular, estimates of gross change and other measures of individual change from a split panel survey will be based on a smaller sample than would have been the case if all the resources had been devoted to the panel component.

In comparing alternative designs for surveys across time, the costs of the designs need to be considered. For instance, panel surveys avoid the costs of repeated sample selections incurred with repeated surveys, but they face costs of tracking and tracing mobile sample members and sometimes costs of incentives to encourage panel members to continue to cooperate in the panel (see Section 3). If two designs can each satisfy the survey objectives, the relative costs for given levels of precision for the survey estimates need to be examined.

3. PANEL SURVEYS

The repeated measures over time on the same sampled elements that are obtained in panel surveys provide such surveys with a key analytic advantage over repeated surveys. The measurements of gross change and other components of individual change that are possible with panel survey data form the basis of a much greater understanding of social processes than can be obtained from a series of independent cross-sectional snapshots. The power of longitudinal data derived from panel surveys has long been recognized (see, for instance, Lazarsfeld and Fiske 1938; Lazarsfeld 1948), and panel surveys have been carried in many fields for many years. Subjects of panel surveys have included, for example, human growth and development, juvenile delinquency, drug use, victimizations from crime, voting behaviour, marketing studies of consumer expenditures, education and career choices, retirement, health, and medical care expenditures. (See Wall and Williams (1970) for a review of early panel studies on human growth and development, Boruch and Pearson (1988) for descriptions of some U.S. panel surveys, and the Subcommittee on Federal Longitudinal Surveys (1986) for descriptions of U.S. federal panel surveys.) In recent years, there has been a major upsurge in interest in panel surveys in many subject-matter areas, and especially in household economics. The ongoing U.S. Panel Study of Income Dynamics began in 1968 (see Hill 1992, for a description of the PSID) and similar long-term panel studies have been started in the past decade in many European countries. The U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) in 1983 (Nelson *et al.* 1985; Kasprzyk 1988; Jabine *et al.* 1990), and Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) in 1993. The growth in interest in panel surveys has also given rise to an increase in literature about the methodology of such surveys, including such recent texts as Kasprzyk *et al.* (1989), Magnusson and Bergman (1990), and Van de Pol (1989).

This section reviews the major issues involved in the design and analysis of panel surveys. The treatment is geared towards repeated panel surveys of fixed duration like the SIPP and SLID, but most of the discussion applies more generally to all forms of panel survey.

3.1 Design Decisions for a Panel Survey

The time dimension adds an extra dimension of complexity to a panel survey as compared with a cross-sectional survey. In addition to all the decisions that need to be made about the design features of a cross-sectional survey, a wide range of extra decisions need to be reached for a panel survey. Major design decisions include:

- *Length of the panel*. The longer the panel lasts, the greater is the wealth of data obtained for longitudinal analysis. For instance, the longer the panel, the greater the number of spells of unemployment starting during the life of the panel that will be completed before the end of the panel, and hence the greater the precision in estimating the survival function for such spells. On the other hand, the longer the panel, the greater the problems of maintaining a representative cross-sectional sample at later waves, because of both sample attrition and difficulties in updating the sample for new entrants to the population.

  It can sometimes be beneficial to vary the length of the panel between different types of panel members. Thus, for instance, when the analytic objectives call for it, panel members with certain characteristics (*e.g.*, members of a minority population) or who experience certain events during the course of the regular panel (*e.g.*, a divorce) can be retained in the panel for extended periods of observation.

- *Length of the reference period*. The frequency of data collection depends on the ability of respondents to recall the information collected in the survey over time. Thus, the PSID, with annual waves of data collection, requires recall of events occurring in the previous calendar year, whereas SIPP, with four-monthly waves of data collection, requires recall for the preceding four months. The longer the reference period, the greater the risk of recall error.

- *Number of waves*. In most cases the number of waves of data collection is determined by a combination of the length of the panel and the length of the reference period. The greater the number of waves, the greater the risk of panel attrition and time-in-sample effects, and the greater the degree of respondent burden.

- *Overlapping or non-overlapping panels*. With a repeated panel survey of fixed duration, a decision needs to be made as to whether the panels should overlap across time. Consider, for instance, the proposal of a

National Research Council study panel that the SIPP should be a four-year panel (Citro and Kalton 1993). One possibility is to run each panel for four years, starting a new panel when the previous one finishes. Another possibility is run each panel for four years, but starting a new panel every two years. Yet another possibility is to run each panel for four years, starting a new panel every year.

The design of nonoverlapping panels has the benefit of simplicity, since only one panel is in the field at any one time. It also produces a large sample for longitudinal analysis; for instance, the panels with the nonoverlapping design can be roughly twice the size of those with the design that has two overlapping panels at any one time. However, this increase in sample size for nonoverlapping panels does not apply for cross-sectional estimates, since the data from the panels covering a given time point can be combined for cross-sectional estimation. Also, the cross-sectional estimates for a time period near the end of a panel with the nonoverlapping design are at greater risk of bias from attrition, time-in-sample bias, and failure to update the sample fully for new population entrants than is the case with an overlapping design, in which one panel is of more recent origin. Moreover, the overlapping design permits the examination of such biases through a comparison of the results for the two panels for a given time period, whereas no such examination is possible with a nonoverlapping design. Another limitation of the nonoverlapping design is that it may not be well positioned to measure the effect of such events as a change in legislation. For instance, if legislation takes effect in the final year of a nonoverlapping panel, there will be little opportunity to evaluate its effect by comparing the situations of the same individuals before and for some period after the legislation. With overlapping panels, one of the panels will provide a wider window of observation.

- *Panel sample size*. For a given amount of annual resources, the sample size for each panel is determined by the preceding factors. A larger panel for longitudinal analysis can be achieved by lengthening the reference period and by employing a nonoverlapping design. The sample size for cross-sectional estimates can be increased by lengthening the reference period, but not by using a nonoverlapping design.

The above list determines the major parameters of a panel survey design, but there still remain a number of other factors that need to be considered:

- *Mode of data collection*. As with any survey, a decision needs to be made as to whether the survey data are to be collected by face-to-face interviewing, by telephone, or by self-completion questionnaire, and whether computer assisted interviewing (CAPI or CATI) is to be used. With a panel survey, this decision needs to be made for each wave of data collection, with the possibility of different modes for different waves (for instance, face-to-face interviewing at the first wave to make contact and establish rapport, with telephone interviewing or mail questionnaires at some of the later waves). When modes may be changed between waves, consideration needs to be given to the comparability of the data across waves. Sometimes a change in mode may involve a change in interviewer, as for instance would occur with a change from face-to-face interviewing to a centralized CATI operation. Then the effects of a change of interviewer between waves on the respondent's willingness to continue in the panel and on the comparability of responses across waves also need to be carefully considered.

- *Dependent interviewing*. With panel surveys there is the possibility of feeding back to respondents their responses at earlier waves of data collection. This dependent interviewing procedure can secure more consistent responses across waves, but risks generating an undue level of consistency. The ease of application of dependent interviewing depends on the length of the interval between waves and the mode of data collection. Processing the responses from one wave to feed back in the next is easier to accomplish if the interval between waves is a long one and if computer assisted interviewing is employed.

- *Incentives*. Monetary or other incentives (*e.g.*, coffee mugs, calculators, lunch bags) may be offered to sampled persons to encourage their participation in a survey. With a panel survey, incentives may be used not only to secure initial participation but also to maintain cooperation throughout the duration of the panel. There is an issue of when are the best times to provide incentives in a panel survey (*e.g.*, at the first wave, at an intermediate wave, or at the last wave of the panel). Panel survey researchers often send respondents a survey newsletter, frequently giving some recent highlights from the survey findings, at regular intervals, both to generate goodwill for the survey and to maintain contact with respondents (see below). Birthday cards sent at the time of the respondents' birthdays are also often used for these purposes.

- *Respondent rules*. Survey data are often collected from proxy informants when respondents are unavailable for interview. With a panel survey, this gives rise to the possibility that the data may be collected from different individuals at different waves, thus jeopardising the comparability of the data across waves. The respondent rules for a panel survey need to take this factor into account.

- *Sample design*. The longitudinal nature of a panel survey needs to be considered in constructing the sample design for the initial wave. Clustered samples are commonly employed for cross-sectional surveys with face-to-face interviewing in order to reduce fieldwork travel costs and to enable frame construction of housing unit listings to be performed only for selected segments. These benefits are bought at the price of the increase in the variance of survey estimates arising from the clustering. The optimum extent of clustering depends on the various cost factors involved and the homogeneity of the survey variables in the clusters (see, for instance, Kish 1965). With a panel survey, the use and extent of any clustering should be determined in relation to the overall panel with all its waves of data collection. In particular, the benefit of reduced fieldwork costs disappears for waves of data collection that are conducted by telephone interviewing or mail questionnaire. Also the migration of panel members to locations outside the original clusters reduces the benefit of the initial clustering for fieldwork costs at later waves. (However, some benefits of the initial clustering still operate for the large proportion of mobile persons who move within their own neighbourhoods.)

  Oversampling of certain population subgroups is widely used in cross-sectional surveys to provide sufficient numbers of subgroup members for separate analysis. Such subgroups may, for instance, comprise persons with low incomes, minority populations, persons in a specified age-group, or persons living in certain geographical areas. Such oversampling can also be useful in panel surveys, but caution is needed in its application. With long-term panels, one reason for caution is that the objectives of the survey may change over time. Oversampling to meet an objective identified at the start of a panel may prove harmful to objectives that emerge later. Another reason for caution is that many of the subgroups of interest are transient in nature (*e.g.*, low income persons, persons living in a given geographical area). Oversampling persons in such subgroups at the outset of the panel may be of limited value for later waves: some of those oversampled will leave the subgroup while others not oversampled will join it. Thirdly, the definition of the desired subgroup for longitudinal analysis needs to be considered. For instance, SIPP data are used to estimate durations of spells on various welfare programs. Since such estimates are usually based on new spells starting during the life of the panel, it may not be useful to oversample persons already enrolled on welfare programs. See Citro and Kalton (1993) for a discussion of oversampling for the SIPP.

- *Updating the sample*. When the sole objective of a panel survey is longitudinal analysis, it may be sufficient to adopt a cohort approach that simply follows the initial sample selected for the first wave. However, when cross-sectional estimates are also of interest, it may be necessary to update the sample at each wave to represent new entrants to the population. Updating for all types of new entrants is often difficult, but it is sometimes possible to develop fairly simple procedures to account for certain types of new entrants. For instance, in a panel of persons of all ages, babies born to women panel members after the start of the panel can be included as panel members. The SIPP population of inference comprises persons aged 15 and over. By identifying in initial sampled households persons who are under 15 years old but who will attain that age before the end of the panel, by following them during the panel, and by interviewing them after they reach 15 years of age, a SIPP panel can be updated for this class of new entrants (Kalton and Lepkowski 1985).

  Attention also needs to be paid to panel members who leave the survey population. For some the departure is clearly permanent (*e.g.*, deaths), but for others it may be only temporary (*e.g.*, going abroad or entering an institution). If efforts are made to keep track of temporary leavers, they can be readmitted to the panel if they return to the survey's population of inference.

  Panel surveys such as SIPP and PSID collect data not only for persons in original sampled households, but also for other persons - nonsampled persons - with whom they are living at later waves. The prime purpose of collecting survey data for nonsampled persons is to be able to describe the economic and social circumstances of sampled persons. The issue arises as to whether any or all nonsampled persons should remain in the panel after they stop living with sampled persons. For some kinds of analysis it is useful to follow them. However, to follow them would eat significantly into the survey's resources.

- *Tracking and tracing.* Most panel surveys encounter the problem that some panel members have moved since the last wave and cannot be located. There are two ways to try to handle this problem. First, attempts can be made to avoid the problem by implementing procedures for tracking panel members between waves. One widely-used procedure when there is a long interval between waves is to send mailings, such as birthday cards and survey newsletters, to respondents between waves, requesting the post office to provide notification of change of address if applicable. Another tracking device is to ask respondents for the names and addresses or telephone numbers of persons close to them (*e.g.*, parents) who are unlikely to move and who will be able to provide locating information for them if they move.

  The second way to deal with lost panel members is to institute various tracing methods to try to locate them. With effort and ingenuity, high success rates can be achieved. Some methods of tracing may be specific for the particular population of interest (*e.g.*, professional societies for persons with professional qualifications) while others may be more general, such as telephone directories, computerized telephone number look-ups, reverse telephone directories for telephone numbers of neighbours, mail forwarding, marriage licence registers, motor vehicle registrations, employers, and credit bureaus. It can be useful to search death records for lost panel members, particularly for long-term panel surveys. Panel members found to have died can then be correctly classified, rather than being viewed as nonrespondents. Methods of tracing are discussed by Burgess (1989), Clarridge *et al.* (1978), Crider *et al.* (1971) and Eckland (1968).

3.2 Problems of Panel Surveys

Panel surveys share with all surveys a wide range of sources of nonsampling error. This section does not review all these sources, but rather concentrates on three sources that are unique to panel surveys, namely wave nonresponse, time-in-sample bias and the seam effect.

3.2.1 Wave nonresponse

The nonresponse experienced by panel surveys at the first wave of data collection corresponds to that experienced by cross-sectional surveys. The distinctive feature of panel surveys is that they encounter further nonresponse at subsequent waves. Some panel members who become nonrespondents at a particular wave do not respond at any subsequent wave while others respond at some or all subsequent waves. The former are often termed attrition cases and the latter non-attrition cases. The overall wave nonresponse rates in panel surveys increase with later waves, but with well-managed surveys the rate of increase usually declines appreciably over time. For example, with the 1987 SIPP panel, the sample loss was 6.7% at wave 1, 12.6% at wave 2, and it then increased slowly to 19.0% at wave 7 (Jabine *et al.* 1990). The tendency for the nonresponse rate to flatten off at later waves is comforting, but nevertheless the accumulation of nonresponse over many waves produces high nonresponse rates at later waves of a long-term panel. For instance, in 1988, after 21 annual rounds of data collection, the PSID nonresponse rate for individuals who lived in 1968 households had risen to 43.9% (Hill 1992).

The choice between the two standard general-purpose methods for handling missing survey data - weighting adjustments and imputation - is not straightforward for wave nonresponse in panel surveys. For longitudinal analysis, the weighting approach drops all records with one or more missing waves from the data file and attempts to compensate for them by weighting adjustments applied to the remaining records. This approach can lead to the loss of a substantial amount of data when the data file covers several waves. On the other hand, the imputation approach retains all the reported data, but requires conducting wholesale imputations for missing waves. A compromise approach uses imputation for some patterns of wave nonresponse (*e.g.*, those with only one missing wave, where data are available from both adjacent waves), and weighting for others (see, for example, Singh *et al.* 1990). For cross-sectional analysis, separate data files may be created for each wave. These files can comprise all the respondents for that wave, with either weighting adjustments or imputations for the wave nonrespondents. Methods for handling wave nonresponse are discussed by Kalton (1986) and Lepkowski (1989).

3.2.2 Time-in-sample bias

Time-in-sample bias, or panel conditioning, refers to the effect that panel members' responses at a given wave of data collection are affected by their participation in previous waves. The effect may reflect simply a change in reporting behaviour. For example, a respondent may recognize from previous interviews that a "Yes" response to a question leads to follow-up questions, whereas a "No" answer does not. The respondent may therefore give a "No" answer to avoid the burden of the extra questions. Alternatively, a respondent may learn from previous interviews that detailed information on income is needed, and may therefore prepare for later interviews by collecting the necessary data. The time-in-sample effect may also reflect a change in actual behaviour. For example, a respondent may enrol in the food stamp program as a result of learning of its existence from the questions asked about it at earlier waves of data collection.

A recent experimental study of panel conditioning in a four-year panel study of newlyweds found some evidence that participation in the study did affect marital well-being (Veroff *et al.* 1992). However, that study used in-depth interviewing techniques that are more intrusive than those used in most surveys. A number of studies of panel conditioning that have been conducted in more standard survey settings have found that conditioning effects do sometimes occur, but they are not pervasive (Traugott and Katosh 1979; Ferber 1964; Mooney 1962; Waterton and Lievesley 1989).

A benefit of rotating and overlapping panel surveys is that they enable estimates for the same time period obtained from different panels to be compared. Such comparisons have clearly identified the presence of what is termed "rotation group bias" in the U.S. and Canadian labour force surveys (*e.g.* Bailar 1975, 1989, and U.S. Bureau of the Census 1978, for the U.S. Current Population Survey; Ghangurde 1982, for the Canadian Labour Force Survey). Rotation group bias may reflect nonresponse bias and conditioning effects. In analyses comparing the overlapping 1985, 1986 and 1987 SIPP panels, Pennell and Lepkowski (1992) found few differences in the results from the different panels.

3.2.3 Seam effect

Many panel surveys collect data for subperiods within the reference period from the last wave of data collection. The SIPP, for instance, collects data on a monthly basis within the four-month reference period between waves. The seam effect refers to the common finding with this form of data collection that the levels of reported changes between adjacent subperiods (*e.g.*, going on or off of a welfare program from one month to the next) are much greater when the data for the pair of subperiods are collected in different waves than when they are collected in the same waves. The seam effect has been found to be pervasive in SIPP, and to relate to both recipiency status and amounts received (see, for example, Jabine *et al.* 1990; Kalton and Miller 1991). It has also been found in PSID (Hill 1987). Murray *et al.* (1991) describe approaches used to reduce the seam effect in the Canadian Labour Market Activity Survey.

3.3 Longitudinal Analysis

There is a substantial and rapidly expanding literature on the analysis of longitudinal data, including a number of texts on the subject (*e.g.*, Goldstein 1979; Hsiao 1986; Kessler and Greenberg 1981; Markus 1979). This treatment cannot be comprehensive, but rather identifies a few general themes.

- *Measurement of gross change*. As has already been noted, a key analytic advantage of a panel survey over a repeated survey is the ability to measure gross change, that is change at the individual level. The basic approach to measuring gross change is the turnover table that tabulates responses at one wave against the responses to the same question at another wave. The severe limitation to this form of analysis is that changes in measurement errors across waves can lead to serious bias in the estimation of the gross change (see Kalton *et al.* 1989, and Rodgers 1989, for further discussion).

- *Relationship between variables across time*. Panel surveys collect the data necessary to study the relationships between variables measured at different times. For instance, based on the data collected in the 1946 British birth cohort, the National Survey of Health and Development, Douglas (1975) found that children who were hospitalized for more than a week or who had repeated hospitalizations between the ages of 6 months and

3½ years exhibited more troublesome behaviour in school and lower reading scores at age 15. In principle, cross-section surveys may employ retrospective questions to collect the data needed to perform this type of analysis. However, the responses to such questions are often subject to serious memory error, and potentially to systematic distortions that affect the relationships investigated.

- *Regression with change scores.* Regression with change scores can be used to avoid a certain type of model misspecification. Suppose that the correct regression model for individual i at time t is

$$Y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \epsilon_{it}$$

where x_{it} is an explanatory variable that changes value over time and z_{it} is an explanatory variable that is constant over time (*e.g.*, gender, race). Suppose further than z_{it} is unobserved; it may well be unknown. Then β can still be estimated from the regression on the change scores:

$$Y_{i(t+1)} - Y_{(t)} = \beta (x_{i(t+1)} - x_{it}) + \epsilon_{i(t+1)} - \epsilon_{it}$$

(Rodgers 1989; Duncan and Kalton 1987).

- *Estimation of spell durations.* The data collected in panel surveys may be used to estimate the distribution of lengths of spells of such events as being on a welfare program. In panel surveys like the SIPP, some individuals have a spell in progress at the start of the panel (initial-censored spells), some start a spell during the panel, and some spells continue beyond the end of the panel (right-censored spells). Thus, not all spells are observed in their entirety. The distribution of spell durations may be estimated by applying survival analysis methods, such as the Kaplan-Meier product-limit estimation procedure to all new spells (including right-censored new spells) starting during the life of the panel (*e.g.*, Ruggles and Williams 1989).

- *Structural equation models with measurement errors.* The sequence of data collection in a panel survey provides a clear ordering of the survey variables that fits well with the use of structural equation modelling for their analysis. This form of analysis can make allowance for measurement errors, and with several repeated measures can handle correlated error structures (*e.g.*, Jöreskog and Sörbom 1979).

4. CONCLUDING REMARKS

The data sets generated from panel surveys are usually extremely rich in analytic potential. They contain repeated measures for some variables that are collected on several occasions, and also measures for other variables that are asked on a single wave. Repeated interviewing of the same sample provides the opportunity to collect data on new variables at each wave, thus yielding data on an extensive range of variables over a number of waves. A panel data set may be analyzed both longitudinally and cross-sectionally. Repeated measures may be used to examine individual response patterns over time, and they may also be related to other variables. Variables measured at a single wave may be analyzed both in relation to other variables measured at that wave and to variables measured at other waves.

The richness of panel data is of value only to the extent that the data set is analyzed, and analyzed in a timely manner. Running a panel survey is like being on a treadmill: the operations of questionnaire design, data collection, processing and analysis have to be undertaken repeatedly for each successive wave. There is a real danger that the survey team will become overwhelmed by this process, with the result that the data are not fully analyzed. To avoid this danger, adequate staffing is needed and a well-integrated organization needs to be established.

In addition it is advisable to keep the panel survey design simple. The survey design should be developed to meet clearly-specified objectives. Adding complexities to the design to enhance the richness of the panel data set for other uses should be critically assessed. Although persuasive arguments can often be made for such additions, they should be rejected if they threaten the orderly conduct of any stage of the survey process.

As noted earlier, measurement errors have particularly harmful effects on the analysis of individual changes from panel survey data. The allocation of part of a panel survey's resources to measure the magnitude of such errors is therefore well warranted (Fuller 1989). Measurement errors may be investigated either by validity studies (comparing survey responses with "true" values from an external source) or by reliability studies (*e.g.*, reinterview studies). The results of such studies may be then used in the survey estimation procedures to adjust for the effects of measurement errors.

REFERENCES

Bailar, B.A. (1975). The effects of rotation group bias on estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bailar, B.A. (1989). Information needs, surveys, and measurement errors. *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 1-24.

Binder, D.A., and Hidiroglou, M.A. (1988). Sampling in time. *Handbook of Statistics, Vol. 6*, eds. P.R. Krishnaiah and C.R. Rao, New York: North Holland, 187-211.

Boruch, R.F., and Pearson, R.W. (1988). Assessing the Quality of Longitudinal Surveys, *Evaluation Review*, 12, 3-58.

Burgess, R.D. (1989). Major Issues and Implications of Tracing Survey Respondents. *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 52-74.

Citro, C.F., and Kalton, G. (1989). *Surveying the Nation's Scientists and Engineers*, Washington DC: National Academy Press.

Citro, C.F., and Kalton, G. (1993). *The Future of the Survey of Income and Program Participation*, Washington DC: National Academy Press.

Clarridge, B.R., Sheehy, L.L., and Hauser, T.S. (1978). Tracing Members of a Panel: a 17-year Follow-up. *Sociological Methodology*, Ed. K.F. Schuessler, San Francisco: Jossey-Bass, 389-437.

Crider, D.M., Willits, F.K., and Bealer, R.C. (1971). Tracking Respondents in Longitudinal Surveys. *Public Opinion Quarterly*, 35, 613-620.

Douglas, J.W.B. (1975). Early Hospital Admissions and Later Disturbances of Behaviour and Learning. *Developmental Medicine and Child Neurology*, 17, 456-480.

Duncan, G.J., and Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97-117.

Eckland, B.K. (1968). Retrieving Mobile Cases in Longitudinal Surveys. *Public Opinion Quarterly*, 32, 51-64.

Subcommittee on Federal Longitudinal Surveys (1986). *Federal Longitudinal Surveys*, Statistical Policy Working Paper 13, Washington DC: Office of Management and Budget.

Ferber, R. (1964). Does a Panel Operation Increase the Reliability of Survey Data: the Case of Consumer Savings. *Proceedings of the Social Statistics Section, American Statistical Association*, 210-216.

Fuller, W.A. (1989). Estimation of Cross-Sectional and Change Parameters: Discussion, *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 480-485.

Ghangurde, P.D. (1982). Rotation Group Bias in the LFS Estimates. *Survey Methodology*, 8, 86-101.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*, New York: Academic Press.

Hill, D. (1987). Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 210-215.

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*, Newbury Park, CA: Sage Publications.

Hsiao, C. (1986). *Analysis of Panel Data*, New York: Cambridge University Press.

Jabine, T.B., King, K.E., and Petroni, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*, Bureau of the Census, Washington DC: U.S. Department of Commerce.

Jöreskog, K.G., and Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models*, Lanham MD: University Press of America.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys. *Journal of Official Statistics*, 2, 303-314.

Kalton, G., Kasprzyk, D., and McMillen, D.B. (1989). Nonsampling Errors in Panel Surveys. *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 249-270.

Kalton G., and Lepkowski, J.M. (1985). Following Rules in SIPP. *Journal of Economic and Social Measurement*, 13, 319-329.

Kalton, G., and Miller, M.E. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation. *Journal of Official Statistics*, 7, 235-245.

Kasprzyk, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*, SIPP Working Paper No. 8830, Washington DC: U.S. Bureau of the Census.

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

Kessler, R.C., and Greenberg, D.F. (1981). *Linear Panel Analysis*, New York: Academic Press.

Kish, L. (1965). *Survey Sampling*, New York: John Wiley.

Lazarsfeld, P.F. (1948). The Use of Panels in Social Research. *Proceedings of the American Philosophical Society*, 42, 405-410.

Lazarsfeld, P.F., and Fiske, M. (1938). The Panel as a New Tool for Measuring Opinion. *Public Opinion Quarterly*, 2, 596-612.

Lepkowski, J.M. (1989). Treatment of Wave Nonresponse in Panel Surveys. *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 348-374.

Magnusson, D., and Bergman, L.R. (Eds.) (1990). *Data Quality in Longitudinal Research*, New York: Cambridge University Press.

Markus, G.B. (1979). *Analyzing Panel Data*, Beverly Hills, CA: Sage Publications.

Mooney, H.W. (1962). *Methodology in Two California Health Surveys*, Public Health Monograph No. 70, Washington DC: U.S. Department of Health, Education, and Welfare.

Murray, T.S., Michaud, S., Egan, M., and Lemaitre, G. (1991). Invisible Seams? The Experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Bureau of the Census Annual Research Conference*, Washington DC: U.S. Department of Commerce, 715-730.

Nelson, D., McMillen, D., and Kasprzyk, D. (1985). *An Overview of the SIPP, Update 1*, SIPP Working Paper No. 8401, Washington DC: U.S. Bureau of the Census.

Pennell, S.G., and Lepkowski, J.M. (1992). Panel Conditioning Effects in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.

Rodgers, W.L. (1989). Comparisons of Alternative Approaches to the Estimation of Simple Causal Models from Panel Data. *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 432-456.

Singh, R., Huggins, V., and Kasprzyk, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys*, SIPP Working Paper No. 9009, Bureau of the Census, Washington DC: U.S. Department of Commerce.

Traugott, M., and Katosh, K. (1979). Response Validity in Surveys of Voting Behavior. *Public Opinion Quarterly*, 79, 359-377.

U.S. Bureau of the Census (1978). *The Current Population Survey: Design and Methodology*, Bureau of the Census Technical Paper No. 40, Washington DC: U.S. Government Printing Office.

Van de Pol, F.J.R. (1989). *Issues of Design and Analysis of Panels*, Amsterdam: Sociometric Research Foundation.

Veroff, J., Hatchett, S., and Douvan, E. (1992). Consequences of Participating in a Longitudinal Study of Marriage. *Public Opinion Quarterly*, 56, 315-327.

Wall, W.D., and Williams, H.L. (1970). *Longitudinal Studies and the Social Sciences*, London: Heinemann.

Waterton, J., and Lievesley, D. (1989). Evidence of Conditioning Effects in the British Social Attitudes Panel. *Panel Surveys*, Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, 319-339.

SESSION 1

Questionnaire Design and Collection Issues

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

USING CAPI IN A LONGITUDINAL SURVEY: A REPORT FROM THE MEDICARE CURRENT BENEFICIARY SURVEY

W.S. Edwards, S. Sperry and B. Edwards [1]

ABSTRACT

The Medicare Current Beneficiary Survey (MCBS) is a continuous panel survey of persons receiving Medicare in the United States. Interviews are conducted three times a year to collect information about the use and cost of health care services. All household interviews are conducted in person with computer-assisted personal interviewing (CAPI). In addition to the usual features of a computer-assisted interview, the MCBS CAPI design includes extensive use of rosters, or lists of items, which interviewers select from, add to, or correct responses from the current or a previous interview. The MCBS CAPI design also incorporates previously collected information in several other ways: (1) presenting it to respondents to aid in bounding and stimulating recall, (2) offering it as a baseline and asking about change since baseline, and (3) probing for previously unavailable details. While this approach has not been problem-free, it has proved successful for collection of longitudinal data about complex behavior.

KEY WORDS: CAPI; Dependent Interviewing; Health Care Survey.

1. INTRODUCTION

1.1 Special Methodological Issues in Longitudinal Surveys

Longitudinal or "panel" surveys face a variety of methodological issues unique to reinterviews with the same respondents. Because respondents are not always consistent across rounds or waves of a longitudinal survey, data cleaning and/or analysis must deal with issues such as which of two values for a static variable given in different interviews is correct, and whether an apparent change in some variable over time is true change or merely a different way of saying the same thing. Longitudinal surveys of behavior also face the issue of whether a particular event occurred within the current survey reference period or whether it is the same as an event reported in a previous interview.

Neter and Waksberg (1964), in methodological research on the U.S. Consumer Expenditure Survey, found that "bounding" of a reference or recall period with previously reported information reduced error due to "telescoping,"or reporting events from outside the reference period as if they were within it. Other longitudinal surveys have used previously reported information to reduce the "seam effect," or the tendency of longitudinal surveys to overestimate change at the "seam,"or boundary, of reference periods between panel interviews. (The "seam effect" is described in, for example, Moore and Kasprzyk 1984.) The use of previously reported data in an interview is commonly called "dependent interviewing."

The use of information reported in a previous interview has typically been limited by the technical difficulties of design and interviewers' difficulties with dependent interviewing in a structured questionnaire. Dependent interviewing has typically required at least two separate documents: one with questions or a question guide, and one with the previously collected data. References from one to the other are often awkward, and require some creativity on the part of the interviewer to form intelligible sentences or questions.

[1] W.S. Edwards, S. Sperry and B. Edwards, Westat, Inc., Rockville, MD, U.S.A.

1.2 Potential Advantages of Computer-assisted Interviewing in Longitudinal Surveys

As Saris (1991) points out, computer-assisted interviewing (CAI) has the potential to make much greater use of previously collected information than have paper-and-pencil interviews. Saris identifies the following advantages of CAI, including:

- range and consistency checks can be built into an interview (consistency can mean within or across interviews);
- error correction can be built in through the display of cumulative information on what Saris calls "summary and correction" screens, which include both current and previous interview's data;
- CAI can schedule questions about regular behaviors based on information about the expected intervals;
- CAI can use chronological data as a recall aid in life history interviewing;
- CAI can be used to remind respondents of previous answers for bounding of recall.

1.3 Purpose of this Paper

Many of the advantages of computer-assisted interviewing for longitudinal surveys have been employed in designing the Medicare Current Beneficiary Survey (MCBS), a panel survey of persons in the U.S. Medicare program that employs computer-assisted personal interviewing (CAPI) for most data collection. This paper will describe some of the aspects of the MCBS design that highlight uses of previously collected information.

2. OVERVIEW OF THE MEDICARE CURRENT BENEFICIARY SURVEY

The MCBS is a continuous panel survey of Medicare beneficiaries being conducted by Westat, Inc., for the Health Care Financing Administration (HCFA), the federal government agency responsible for administering the Medicare program. The MCBS collects information about the use and costs of health care services, about health insurance coverage, and about other health-related matters. The sample comprises about 12,000 individuals, selected from Medicare rolls, of whom about 11,000 reside in households and 1,000 in nursing homes or other institutions at any given time. All household interviews are conducted in person by interviewers using Westat's CAPI system.

2.1 The Medicare Program

Medicare is a Federal health insurance program that covers most persons 65 and older and certain disabled persons in the United States. Medicare includes both hospital and medical insurance. Medicare pays for many, but not all, of beneficiaries' health care expenses. Beneficiaries are responsible for deductibles and co-insurance for many covered services, and are responsible for the entire cost of noncovered services. Most prescribed medicines, dental care, and long-term nursing home care are not covered by Medicare. Beneficiaries may have other public or private health insurance coverage that pays some or all of what Medicare doesn't.

2.2 Purpose of the MCBS

The Medicare program maintains massive claims files that can be used to analyze the use of covered services by beneficiaries and the amounts that Medicare pays. However, Medicare files include only limited demographic and health status information, very little information on how beneficiaries pay deductibles, co-insurance, and other charges for covered services, and no information on the use and cost of services not covered by Medicare. The MCBS is designed to provide analysts with this missing information for improved program management, to assess the effects on beneficiaries of changes in the program, and to model the effects of proposed changes.

2.3 The Complexity of Health Care Financing in the United States

Obtaining and paying for medical care in the United States can be a complicated set of activities. People may go to many different doctors and other medical providers, who may be organized in a variety of different ways. A single medical event, particularly if it is a hospital stay, may involve several different medical services, several providers, and several different bills. Alternatively, many services may be received from a single provider and included in one bill. As described in the previous paragraph, several different sources may wind up paying parts of a particular medical bill. All of this complexity makes design of a survey to describe the use and costs of, and payments for, medical care services a challenging task.

2.4 The MCBS Interview

The MCBS includes three interviews per year with each respondent. Each interview covers the period of time between the previous interview and the current interview. The core interview, administered in each round, includes questions on household composition, health insurance coverage, the use of a wide variety of medical services, and the charges and sources and amounts of payment for these services. Each interview also includes supplementary questions on a variety of topics.

3. CROSS-SECTIONAL ASPECTS OF MCBS CAPI DESIGN

This section will describe features of the MCBS CAPI design that are not specifically related to the longitudinal nature of the survey, but are essential building blocks for the panel-specific features.

3.1 Basic Features of CAPI Design

The MCBS questionnaire includes the usual features of CAI: automatic skips based on responses to previous questions; automatic word choices such as insertion of the sample person's name, gender-appropriate pronouns, insertion of medical provider names, and so on; range checks on numeric and other entries to reduce keying errors.

In addition, the MCBS CAPI relies extensively on the use of rosters, or lists of items in a particular category mentioned in interviews. The MCBS CAPI includes rosters of people (*e.g.*, sample person, household members, contact persons), medical providers (*e.g.*, doctors, hospitals, home health agencies), health insurance plans (*e.g.*, Medicare, Medicaid, Blue Cross/Blue Shield), dates of visits to each medical provider, and sources of payment for medical care (*e.g.*, insurance plans, family, professional courtesy). The rosters are developed by the interviewer during the interview. Roster screens come up automatically when appropriate in the interview flow, or they may be called up by the interviewer at any time for review. Generally, interviewers can select an item already on a roster, can add items to a roster, and can make corrections to items on a roster.

3.2 Details of MCBS Rosters

Exhibit 1 presents an MCBS CAPI screen -- a probe for private health insurance coverage. The screen includes identifying numbers across the top, a question text to be read by the interviewer, and answer categories at the bottom of the screen. The interviewer enters the number of the appropriate response (the cursor resides within the parentheses above the answer categories) and strikes the "Enter" key to complete recording the response.

Exhibit 1: MCBS CAPI Screen -- Health Insurance Probe.

```
3.17    HI17B                         10000027 9112181803 911219

        I would like to ask about other types of health insurance.

        At any other time since August 20, 1991,
        have you been covered by
        private health insurance, that is, a plan that
        pays hospital or doctor bills or covers the cost of
        prescribed medicines?

                          ( )

                          1. YES
                          2. NO
```

If the response to the question in Exhibit 1 is "yes," the CAPI program displays the Health Insurance Roster, presented in Exhibit 2.

Exhibit 2: MCBS CAPI Screen -- Health Insurance Roster.

```
3.200   HI20                          10000027 9112181803 911219

        What is the name of each of the other plans that provide
        your medical insurance coverage?
        [ENTER ALL PRIVATE PLANS.]

      TO ERASE AN X, PRESS SPACE BAR.  TO ADD A PLAN, PRESS CTRL/A.
      TO LEAVE SCREEN, PRESS ESC.

        |--- PLAN NAME------------------------- STATUS----------------- |
        |    MEDICARE                           CURRENT                  |
        |    MEDICAL ASSISTANCE                 CURRENT                  |
        |    MARYLAND PHARMACY ASST.            STOPPED 11/30/91         |
        |    THE RAINBOW COVERAGE PLAN          BEING ADDED              |
        |--------------------------------------------------------------- |
```

In Exhibit 2, the roster of health insurance plans appears within the box. The question text appears above the box in upper and lower case, and instructions to the interviewer are in all upper case. The interviewer can select a plan already entered by placing an "X" to the left of the plan name, can add a new plan, or can correct the name of an existing plan. In this case, the interviewer is adding the plan "The Rainbow Coverage Plan." Apparently the respondent reported earlier that the coverage for "Maryland Pharmacy Assistance" had stopped. Following this screen, the program would bring up questions eliciting details of the Rainbow Plan coverage.

Exhibits 3-5 present more examples of how rosters are used in the MCBS. Exhibit 3 is a probe for doctor visits within the reference period (August 20, 1991, to the date of the interview).

Exhibit 3: MCBS CAPI Screen -- Doctor Visit Probe.

```
10.991  MP1                    10000027 9112181803 911219


     Since August 20, 1991,
     have you seen
     any medical doctors?
     [INCLUDE ANY VISITS FOR TESTS/X-RAYS.]

                       ( )

                       1. YES
                       2. NO
```

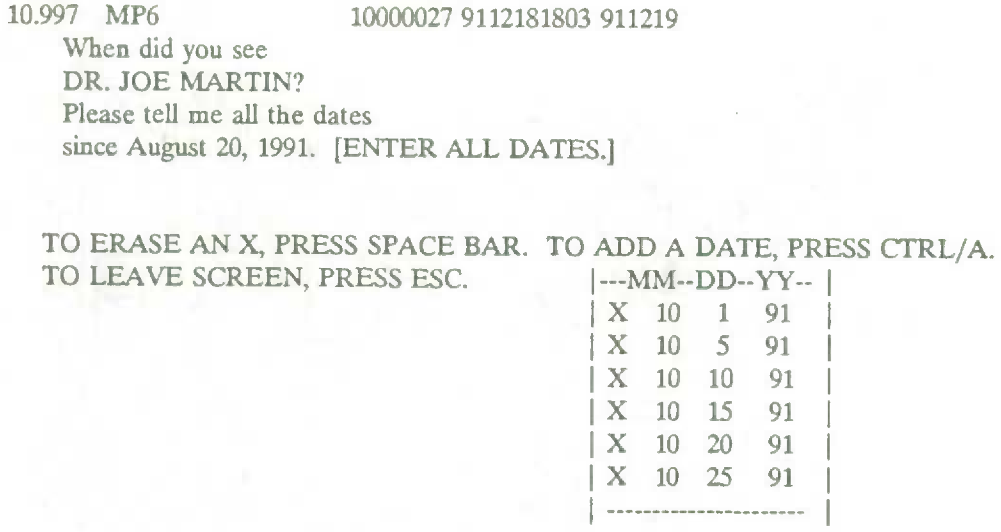
If the response to the probe in Exhibit 3 is "Yes," CAPI displays the medical provider roster as shown in Exhibit 4. The interviewer selects a previously entered medical provider or enters a new one, just as with the Health Insurance Roster in Exhibit 2. Following selection of a medical provider, CAPI then presents the Visit Roster for that provider, shown in Exhibit 5.

Exhibit 4: MCBS CAPI Screen -- Medical Provider Roster.

```
10.992   MP2                    10000027 9112181803 911219

     Who did you see?
     [ENTER ONLY ONE PROVIDER.]

TO SELECT A PROVIDER, USE ARROW KEYS, PRESS X, PRESS ENTER.

TO ERASE AN X, PRESS SPACE BAR.  TO ADD A PROVIDER, PRESS CTRL/A.
TO LEAVE SCREEN, PRESS ESC.

    |--- PROVIDER NAME----------------------------------- |
    |    SARA DALE                                        |
    |    BOB DALE                                         |
    |    ROCKVILLE CENTER                                 |
    |    DR. JOE MARTIN                                   |
    |---------------------------------------------------- |
```

In Exhibit 5, the interviewer has added six dates on which Dr. Joe Martin, the provider selected from the provider roster, was seen during the reference period. The CAPI program would now bring up a series of questions about each of the visits. The roster boxes on the screens shown in Exhibits 2, 4, and 5 scroll up and down to allow entry and review of a large number of items.

Exhibit 5: MCBS CAPI Screen -- Visit Roster.



```
10.997   MP6                    10000027 9112181803 911219
     When did you see
     DR. JOE MARTIN?
     Please tell me all the dates
     since August 20, 1991.  [ENTER ALL DATES.]


  TO ERASE AN X, PRESS SPACE BAR.  TO ADD A DATE, PRESS CTRL/A.
  TO LEAVE SCREEN, PRESS ESC.          |---MM--DD--YY-- |
                                       | X  10   1   91  |
                                       | X  10   5   91  |
                                       | X  10  10   91  |
                                       | X  10  15   91  |
                                       | X  10  20   91  |
                                       | X  10  25   91  |
                                       | --------------- |
```

3.3 Summary of CAPI Advantages for Cross-sectional Design in MCBS

This section has presented a few examples of how CAPI allows a complex questionnaire design to capture details of the complex behavior of obtaining and paying for medical care services. As is typical of CAI systems, the MCBS CAPI includes complex word choice options within a question, including insertion of the sample person's name, the name of medical providers and other items from rosters, and a customized reference period for each interview.

The heart of the MCBS CAPI is the system of rosters, which allows the interviewer to build lists of items throughout the interview, and to select items from the rosters that reappear later. The MCBS CAPI tracks complex relationships between items on one roster and items on another. In the examples shown above, a health insurance plan listed in Exhibit 2 might be one source of payment for a visit in Exhibit 5 to a provider in Exhibit 4.

A final important advantage of CAPI is the ability to employ complex, interdependent routing schemes, where several different data items may be reviewed before the correct next item is selected. For example, the rosters typically begin loops through question series. In the example from Exhibit 5, the program would present a series of questions about each of the visits entered in the roster, then a probe for additional visits to the same provider, and then a probe for additional providers. If additional visits or providers are mentioned, the loops start again.

4. LONGITUDINAL FEATURES OF THE MCBS CAPI

4.1 Overview

As noted earlier, the MCBS is a longitudinal survey. Many of the features described in the preceding section have longitudinal applications. In particular, the rosters are maintained across as well as within interview rounds, so that, for example, a provider mentioned in one interview will appear on the Provider Roster in the following interview. Some rosters have an additional column to designate in which interview round the item was reported. Other information routinely carried from one interview to the next includes household composition, the name of the respondent and, if a proxy, his/her relationship to the sample person, locating information, and so on.

In a sense, these applications constitute "dependent interviewing," or the use of previously collected information in a current interview. However, the MCBS interview includes other dependent interviewing features that make more dramatic use of previous interviews' data. These are described in the following sections.

4.2 "Passive" Dependent Interviewing

One way in which previous interviews' data is used in the MCBS is what we have labeled "passive" dependent interviewing. It is passive in that information is presented to the respondent, but the questionnaire includes no specific questions about the data, and the data are not used explicitly in questions to probe for further information. The presentation is in the form of an annotated calendar, with icons representing doctor visits, hospital stays, and purchases of medical equipment, supplies, and related items or services.

The purposes of presenting data in this way are (1) to reduce "telescoping," or reporting of medical events from outside of the current round's reference period, and (2) to provide recall cues to the respondent -- the sample person may have seen the same medical provider or purchased the same medicine in the current round's reference period as in the previous period. The respondent is encouraged to peruse the summary and to refer to it during the interview. The CAPI script does not solicit additions or corrections to the summary data, but if the respondent offers them the interviewer may make additions or corrections. Of more than 110,000 medical events reported in the first round of the core interview, just over 1,000, or about one percent, were deleted during the next interview. An additional 768 medical events, representing 0.7 percent of the total, were added during the next interview.

4.3 "Baseline/Change" Approach to Dependent Interviewing

One of the two "active" dependent interviewing techniques used in the MCBS is to present previously collected information as a "baseline" during the current interview, and then to ask whether there has been any change. This approach is used for continuous information, notably household composition, health insurance coverage, and home health care.

For health insurance coverage, the interviewer introduces the topic and hands the respondent a printed summary of the plans in effect in the previous round. The interviewer then verifies the coverage, as shown in Exhibit 6.

Exhibit 6: MCBS CAPI Screen -- Verification of Health Insurance Baseline Information.

```
3.01   HIS1                    10000027 9112181803 911219


          You had Medicare coverage and you were
          also covered by (READ PLAN NAMES BELOW)
          on February 19, 1992.
          Is that correct?

          MEDICARE              CONNPACE
          BLUE CROSS            BLUE SHIELD

                                 ( )

                           1. YES, ALL CORRECT AS SHOWN
                           2. NO, PLAN MISSING
                           3. NO, PLAN NAME INCORRECT
                           4. NO, PLAN NEEDS DELETION
```

Unlike the "passive" presentation of utilization information, here the interviewer explicitly probes for corrections to the previous information. The reason for this probe is to prevent corrections to the previous information from being misinterpreted as change. Of 11,804 non-Medicare health insurance plans reported in the first MCBS interview, 308 (2.6 percent) were deleted in the next interview. An additional 298 plans (2.5 percent) were added as a result of this probe. Under other circumstances, these corrections might have been counted as changes between interview rounds.

The interview continues by asking whether, for each of the plans (other than Medicare) reported as in effect at the time of the previous interview, the plan is still in effect. Of the approximately 10,950 plans in effect at the end of the first interview round, 358 (3.3 percent) ended during the second round's reference period. An additional 849 plans (7.8 percent of the Round 1 total) were added during the second round. These should represent actual change, as opposed to correction of previous response.

4.4 Probing for Newly Available Information

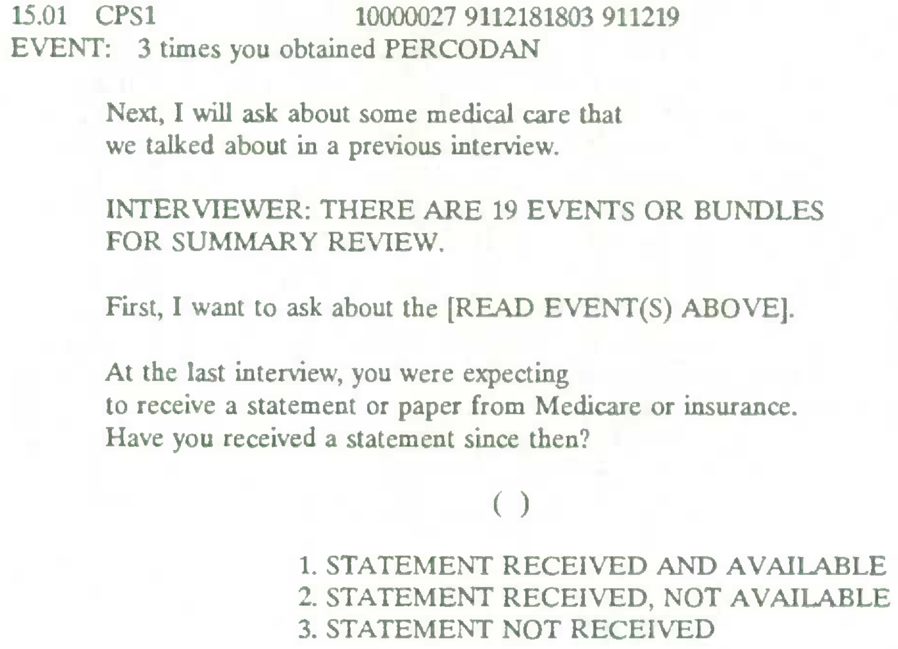
Although most medical events take one day or less (hospital and nursing home stays are notable exceptions), the payment process for a medical event often takes several months. Thus, for many medical events reported in the MCBS interview over the approximately four-month reference period the respondent cannot report all the details of who paid and how much. The interview design takes this delay into account. For example, if the sample person has not received a statement from Medicare (usually the first payer), but expects one, the interviewer asks no further questions about charges or payment. This saves some time in the interview, and presumably some respondent annoyance at being asked to provide information they do not have. In other cases, one source may have paid, but another not, or the sample person may have paid himself or herself and be awaiting reimbursement from Medicare or private insurance.

In such cases, the MCBS questionnaire carries the information over from one interview to the next, and follows up on the basis of what was missing in the previous interview. In these kinds of situations, the basic question is, "Do you now know?" If, for example, a statement was expected from Medicare for a particular event in the first interview, a flag is set in the database to bring this event up in the "Charge/Payment Summary" section of the next interview. However, before that section, the interviewer asks the respondent for any Medicare or insurance statements that have been received since the previous interview, and enters charge and payment data from them. A flagged event may be covered in this way, in which case the flag is removed before the summary section. For events that make it through to the summary, Exhibit 7 shows the probe screen in the Charge/Payment Summary, here for an "event" consisting of three purchases of the medicine Percodan.

If a statement had been received, the program would proceed to collect information about charges and payments, with different screens depending upon whether the statement was available. If the statement had not been received, the next question would be, "Do you still expect to receive a statement . . .?" An event could be carried over one more round if a statement was still expected.

Of 19,031 "statements expected" from the first core interview round, statements for 26 percent were picked up before reaching the Charge/Payment Summary. Another 11 percent were picked up (received and available) in the probe shown in Exhibit 7, and 32 percent were coded as "statement received, not available" in the probe. This approach clearly captures additional information that would not be obtained as accurately otherwise (assuming the statements provide more accurate information than the respondents could from memory). However, some of the "statement received, not available" cases were undoubtedly not "real" "statement expected" cases in the previous interview. As yet the tradeoff between saving time and respondent burden by delaying the charge/payment questions and improving the accuracy of reporting on the one hand and delaying questions that will require the respondent to search his or her memory on the other hand is not clear.

Exhibit 7. MCBS CAPI Screen -- Follow-up Probe for Expected Statement



```
15.01   CPS1                    10000027 9112181803 911219
EVENT:   3 times you obtained PERCODAN

            Next, I will ask about some medical care that
            we talked about in a previous interview.

            INTERVIEWER: THERE ARE 19 EVENTS OR BUNDLES
            FOR SUMMARY REVIEW.

            First, I want to ask about the [READ EVENT(S) ABOVE].

            At the last interview, you were expecting
            to receive a statement or paper from Medicare or insurance.
            Have you received a statement since then?

                                        ( )

                              1. STATEMENT RECEIVED AND AVAILABLE
                              2. STATEMENT RECEIVED, NOT AVAILABLE
                              3. STATEMENT NOT RECEIVED
```

5. DISCUSSION -- LESSONS LEARNED FROM THE MCBS

5.1 Caveats

This paper has presented only the beginning of the complexity of the MCBS CAPI questionnaire design. While we are convinced that capturing complex behavior accurately requires a complex design and that CAPI has greatly facilitated achieving such a design for the MCBS, the road has not been without significant obstacles.

The design of the MCBS was developed over the course of one year, from contract award to start of the pretest, much longer than had originally been scheduled. The design work also required the integration of skills that are traditionally separated in survey operations, notably questionnaire design and database design. The questionnaire designers had to understand the complex database design, and the database designers had to understand the objectives and techniques of questionnaire design. Staff with these disparate skills met on the field of the CAPI programming.

As we have reported elsewhere (Edwards *et al.* 1992), the interviewers took well to using CAPI. However, not all interviewers mastered all aspects of the questionnaire program. Interviewer training was long (seven-day, in-person sessions before Rounds 1 and 2, and five days before Round 3, the first round with the complex longitudinal features), and focused on problem-solving as well as straightforward questionnaire administration. Unlike a paper-and-pencil field study, in which interviewers can get around problems by writing long notes and finding the next applicable question, or CATI studies, in which floor supervisors can be summoned in a jam, a CAPI study requires interviewers to solve problems within the rules in order to proceed with the interview. The MCBS, as we have described, attempts to capture complex behavior that is often documented in complicated paperwork. While the CAPI instrument made the questionnaire easier to administer, it could not necessarily overcome the difficulty of the subject matter for both interviewer and respondent.

Finally, the extensive use of previously reported data means that the database becomes "live" for long periods of time while it is in the field. The design of such an application must include numerous tight controls on what an interviewer can and can't do to affect the database. For example, insurance plans or medical events that we referred to as being "deleted" in the dependent interviewing applications are in fact merely flagged for deletion. The data are not actually removed from the database.

5.2 Advantages of CAPI for Longitudinal Surveys

Despite the caveats listed above, CAPI does offer many advantages for designing complex longitudinal surveys, the most dramatic of which involve use of previously collected information. First, with a CAPI program, interviewers can review, correct, and add to previously collected information with relative ease, regardless of whether the information was collected in the current interview or a previous interview. Important design decisions in this kind of application include whether and when to allow corrections and deletions, what kinds of constraints to impose on these activities, and how to handle old and new entries in the database.

A second use of previously collected data in a CAPI design is to program precise and complex skip patterns based on previous responses. Again, the responses may be from the current or a previous interview. A third possibility is precise scripting of questions based on previous responses, including the insertion of names, gender- and person-appropriate pronouns, dollar amounts, and so on. This feature may include calculations or other manipulations of previous responses, such as (in an MCBS example) calculating the amount of a bill remaining to be paid after totaling several payments, or sorting events in chronological order.

These examples certainly do not exhaust the possibilities for CAPI design of longitudinal surveys. The MCBS design demonstrates that interviewers can handle data from a previous interview virtually as easily as data from a current interview in a CAPI application.

REFERENCES

Edwards, B., Edwards, W.S., Gay, N., and Sperry, S. (1992). CAPI on the medicare current beneficiary survey: a report on round 1. Paper presented at the Annual Conference of the American Association of Public Opinion Research, St. Petersburg, FL.

Moore, J., and Kasprzyk, D. (1984). Month-to-month recipiency turnover in the ISDP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Washington, D.C.

Neter, J., and Waksberg, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

Saris, W.E. (1991). Computer-assisted interviewing. Sage University paper series on quantitative applications in the social sciences, series no. 07-080. Newbury Park, CA: Sage.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

A "COGNITIVE" INTERVIEWING APPROACH FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION: DEVELOPMENT OF PROCEDURES AND INITIAL TEST RESULTS

J.C. Moore, K. Bogen and K.H. Marquis[1]

ABSTRACT

This paper describes the development and initial testing of experimental data collection procedures for the Survey of Income and Program Participation (SIPP). The new procedures derive from prior research which has revealed serious levels of measurement error in some of SIPP's basic statistics, the important implications of the errors for standard analytical uses of the data, and which has suggested the cognitive bases of the errors. The key features of the redesigned procedures are a clear and consistent message to *all* participants that accuracy is the primary goal, and an emphasis on the use of records to assist income reporting. Initial results from small-scale tests of the new procedures indicate a high rate of record use to report income flows, and decreased response error (as indicated by a reduction in underreport errors and in the "seam bias"); on the negative side, the initial tests have suffered substantially greater nonresponse than does standard SIPP, and possibly increased per-case costs.

KEY WORDS: Cognitive research; Measurement error; Questionnaire design; Record use; Seam bias.

1. INTRODUCTION

1.1 The Survey of Income and Program Participation

The Survey of Income and Program Participation (SIPP) is a major, continuing demographic survey program of the U.S. Census Bureau, and an important source of key social and economic indicators for the United States. This large-scale survey provides the most comprehensive information ever assembled on the economic situation of persons and families in the United States. SIPP data contribute to a wide range of policy decisions - health insurance and pension coverage, tax reform, Social Security costs, the effectiveness of state and federal assistance programs, *etc*.

In its current design, a new SIPP panel is introduced every year, and has a life of about 2½ years; households in each panel are interviewed eight times at four month intervals. All household members aged 15 and older are eligible for interview. Each interview (or "wave") gathers monthly data for the four calendar months preceding the interview month. Self-response is the preferred reporting mode, but proxies are accepted for persons not available at the time of the interviewer's visit. Until recently, all SIPP interviews were conducted by personal visit. Starting in February 1992, however, interviews 3, 4, 5, 7 and 8 were designated as telephone interviews.

1.2 Overview of the Paper

This paper presents an interim report on a research program still in progress. It describes the SIPP Cognitive Research (SIPP-CR) Project, the goal of which is to develop and test alternative measurement procedures for the SIPP to reduce important measurement errors. Section 2 presents a brief description of prior research leading up to the current study. Section 3 describes the major features of the new procedures, and the ways in

[1] J.C. Moore, K. Bogen and K.H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, Washington, DC, USA, 20233-4700.
The views expressed herein are the authors' and do not necessarily represent the official views of the Census Bureau.

which they differ from standard SIPP. Section 4 outlines the research plan for the SIPP-CR Project, and Section 5 summarizes results of initial pretests using the new procedures. Section 6 offers some brief conclusions and thoughts about next steps. Additional information on these topics is presented in Marquis, Moore and Bogen (1991).

2. PRIOR RESEARCH ON SIPP MEASUREMENT ERRORS

2.1 The "Seam Bias"

Previous research has revealed important measurement error problems in SIPP. In an early study, Burkhead and Coder (1985) identified a "seam bias" in the measurement of month-to-month change using SIPP data. The seam bias is the tendency for many more changes (for example, from "on" to "off" participation in some transfer program) to appear between adjacent months at the "seam" between two interview waves, than between two adjacent months within the reference period of a single interview wave. No reasonable scenario for true month-to-month change could produce this sort of pattern; thus, the seam bias is a clear indicator of problems in SIPP's measurement of change.

2.2 The SIPP Record Check Study

The SIPP Record Check Study (Moore and Marquis (1989), Marquis and Moore (1990)) was implemented to investigate the nature and extent of response error in SIPP, and, specifically, to better understand the nature of the seam bias and its causes. The study used a full-design record check to assess measurement quality for reports of participation in and income received from eight government transfer programs, in four states, for the first two waves of the 1984 SIPP panel.

The Record Check Study showed that reporting errors in SIPP are quite rare; overall, fewer than 2% of the reports about program participation or changes in program participation were found to be in error. The study also showed, however, that even low levels of response error can have severe effects on important estimates, both univariate estimates and measures of association. For program participation rates, Marquis and Moore (1990) report net underestimates in the 10-40% range. Program participation change rates are underestimated by even greater amounts within a wave (off-seam), while change rates on the interview seam are severely *over*estimated.

While the record check research has permitted detailed descriptions of SIPP response errors, it has proved much less useful in identifying causes of the response errors. Marquis and Moore (1990) examined several of the traditional hypotheses about the causes of survey response errors - forgetting, memory decay, confusion, proxy bias, *etc.* - and found that none was strongly supported by the data.

2.3 Exploratory SIPP Cognitive Research

In a further search for causal insights, the Census Bureau implemented a small-scale, exploratory cognitive research project to look for clues to the survey's response error difficulties in respondents' understanding of SIPP tasks and questions, and their thought processes in answering those questions. Census Bureau headquarters staff received training in "cognitive interviewing" techniques, and accompanied experienced SIPP interviewers while they administered the standard SIPP interview. The observers were free to interject questions to find out how the respondent interpreted a general task or a specific question, formulated a response, *etc.*, or simply to observe the interviewer-respondent interaction. This project and its results are summarized in Marquis (1990).

The exploratory cognitive research yielded several important insights into the response error dynamics of the SIPP interview. One key insight was the limited role that memory plays in respondents' supposed "recall" of their reference period income. As a substitute for detailed, direct recall of the 4-month payment history, respondents tend to rely on very simple rules, combined with a few recalled facts, to construct a plausible (though not necessarily accurate) story about their income. Furthermore, this shortcut, "story-telling" strategy appears to be not only tolerated by current SIPP procedures, but encouraged in subtle but important ways. For example, interviewers' performance evaluations are primarily based on their response rates and efficiency. This environment can work against response quality, because it can discourage interviewers from "pushing too hard" for accurate answers, either through record use or through difficult, more complex recall strategies; from following up "don't know's" and item refusals; or from responding to, or even recognizing, respondents' elaborations to too-simple answers. The evaluation environment can also encourage interviewers to rush through the interview, and even to actively assist respondents in finding easy approximations to complex truth.

The cognitive research also pointed out many ways in which the current SIPP questionnaire contributes to response quality problems. In many areas, the questionnaire presents demands on memory that are simply unreasonable (for example, asking for highly unrealistic detail, or for the recall of material that is unlikely to have been stored in memory in the first place); in other areas respondents are effectively denied an opportunity to report accurately in the interest of processing efficiency (for example, in the requirement to report all income in monthly "chunks," even income which comes on a schedule not immediately translatable to monthly). These shortcomings force respondents away from reporting accuracy, and toward "story-telling." The SIPP questionnaire also falls short by failing to provide clear and consistent information to respondents about the nature of their task. The cognitive interview observers often found that respondents did not understand the point of an entire question series. This was sometimes due to the lack of explanatory transition statements between major topic areas; in other instances the complexity of the instrument, with its myriad of screeners and check items (often read aloud by interviewers, further disrupting the flow and context), was clearly at fault. The instrument also fails to provide adequate or consistent information about the level of accuracy or effort expected of the respondents.

3. AN ALTERNATIVE MEASUREMENT DESIGN FOR SIPP

The exploratory cognitive research provided important insights into the likely causes of SIPP's response error problems, and led directly to many of the changes incorporated into a set of alternative measurement procedures for the survey. The major components of the new procedures are as follows:

- The cornerstone of the new measurement procedures is the emphasis on respondents' *use of personal income records* to assist income reporting. Accurate recall from memory of a four-month stream of income is usually very difficult, and often virtually impossible; the revised procedures explicitly recognize this fact. They attempt to take the reporting task out of respondents' heads entirely (and thus are not really "cognitive" at all) by insisting instead that respondents use their personal records to report their income, in order to preempt their use of overly simple response strategies, and to ensure accuracy. Interviewers are also responsible for training respondents how to interpret their records, and how to maintain them for the next interview. This includes giving respondents a file folder for storing records between interviews, and, for income not accompanied by any record, a sheet on which to record the relevant details concerning that income (date, amount, source and recipient).

- In the absence of income records, interviewers are trained to recognize unacceptable shortcut strategies, and to guide respondents to *use more realistic recall strategies*. In such circumstances, respondents are first asked to describe the "usual" pattern of payment dates and amounts; then to list factors that can conceivably affect payment dates or amounts; next to consider whether any of the possible "change" factors occurred during the reference period, and if so, when; and finally, from this complex mix of information, to reconstruct what actually happened during the reference period.

- To avoid "story-telling," to reinforce the message that accuracy is the primary goal of the survey, and to make records easier to use, the new procedures *collect individual, "to-the-penny" income payments*, not monthly totals. Regardless of how often respondents receive income from a particular source, interviewers collect dates and amounts for each individual payment. Monthly totals are produced by computer, not in respondents' heads. Even income sources for which data users may not need exact amounts are collected with the same level of precision, in order to ensure a consistent message to respondents that accuracy is vital, and that estimates are not acceptable.

- The new procedures use *unstandardized interviewing techniques in the collection of income information*. The alternative SIPP interview begins with a "free recall" section which attempts to clearly set out the goals of the section, and then allows respondents substantial control over the reporting of their income for the reference period. There is a structure to this part of the interview - the information goals are explicit, and the data capture mechanism gives clear guidance about the specific data needed. What is missing is the inviolable script, with pre-set questions in a pre-set order. There are many potential benefits of this format. By allowing respondents to report salient facts about their income with little delay, in the most natural order, without having to endure long strings of inapplicable or seemingly irrelevant questions, it enables them to become immediately involved in the interview and in the production of good information. It also equips the interview with the flexibility to handle the great complexity and diversity of people's income situations. Moore, Bogen, and Marquis (1992) provide a thorough description of this aspect of the revised SIPP procedures.

- The new procedures attempt to *simplify the reporting tasks* as much as possible, and *clearly explain to respondents the purposes and goals* of each section. Items have been re-ordered to make sections of the questionnaire more coherent. This change, as well as the "free recall" procedures described above, has eliminated the need for many complex skip patterns, allowing interviewers to concentrate on their essential task, which is no longer question-reading, but problem-solving. In some instances, the revised questionnaire opts to ask some items of a slightly larger-than-necessary universe of respondents, in order to eliminate preceding screener questions. Another change is the addition of short transition statements between the major sections of the questionnaire, to provide respondents with a guide for what to expect next.

- For the first interview, the new procedures insist on *self-response, preferably "family-style," in a non-distracting interview setting*. These components of the revised interview are intended to both reinforce the message that the survey seeks the highest quality information, and provide an environment that is most conducive to achieving high quality. In subsequent interviews, if the household has records available, the self-response and group interview rules can be relaxed; their purpose initially is to make sure all household members understand the goals and importance of the survey, to allow them to help each other recall income sources and details, and also to provide implicit approval for household members to share income information, thus paving the way for collecting high-quality proxy information (with records, of course) in future interviews.

- To directly attack the seam bias, in particular the overreporting of change at the seam, the revised procedures use *overlapping reference periods with reconciliation* of discrepant information, a technique adapted from Murray, *et al.* (1991). Unlike standard SIPP, each wave's reference period extends to the date of the interview, rather than ending on the last day of the full month preceding the interview. Since the reference period for the next interview starts at the beginning of the month in which the preceding interview took place, for interviews after the first there is an overlap period covered by both the current and the previous interview. At the second and subsequent interviews, the interviewer first collects income information independent of the previous interview, then reviews the information with respondents in light of previous information. There are two stages to this review. First, the interviewer resolves any discrepancies in income sources, checking all income sources reported in one interview but not the other for possible omissions. Following this, interviewers review both waves' data for discrepant income information in the overlap period, and resolve all discrepancies with respondents.

- An essential underpinning of the new procedures is a set of *revised interviewer evaluation criteria*, which are intended to encourage interviewers to attend to quality-oriented performance. The revised procedures no longer place primary and almost exclusive emphasis on high response rates and high efficiency, but add many indicators of the extent to which their performance is consistent with the primary quality goals, and raise those indicators to first-level importance. The main form of feedback is through monitoring a sample of tape-recorded interviews (all interviews are supposed to be taped) in such areas as obtaining group interviews and self-response, persuading respondents to use records, reconstructing income details in the absence of records using complex recall strategies, providing feedback to respondents, recognizing and solving respondents' difficulties, *etc.*

4. THE RESEARCH PLAN

The Census Bureau has designed a research program, currently in progress in the field, for evaluating and refining the revised "cognitive" procedures. This program includes two small pretests, a full-scale measurement quality evaluation study (now in the field), and an implementation research panel to address operational issues.

4.1 Pretest 1

The first pretest was conducted in Milwaukee, WI, from August through November 1991. Wave 1 interviews were conducted in August and September, with a standard four-month reference period; Wave 2 interviews, with a shortened two-month reference period[2], were conducted in October and November in households that had completed an initial interview two months before. The sample was 130 randomly selected addresses. The purpose of the first pretest was to assess the feasibility of, and refine as necessary, the new field procedures and instruments.

4.2 Pretest 2

The second pretest employed the same general design as the first: two months of Wave 1 interviews in December 1991 and January 1992, with a four-month reference period, and two months of Wave 2 interviews in February and March 1992, with a two-month reference period. The sample for Pretest 2 consisted of 130 individuals (and their associated household members), who resided in Milwaukee, and who were identified on official record systems as having received one of five income types - Aid to Families with Dependent Children (AFDC), Food Stamps, Unemployment Insurance, Supplemental Security Income (SSI), or earnings from a specific Milwaukee-area employer. The purpose of the second pretest was to test procedures for sampling from and matching to administrative and employer records; to develop data entry, database management, and data analysis strategies and programs; and to further test and refine the revised procedures and instruments.

4.3 The Evaluation Study

The Evaluation Study is currently underway, also in Milwaukee. When complete, it will include two waves of interviewing, each with a full four-month reference period. Wave 1 interviews began in September 1992, and will continue through January 1993; Wave 2 interviews will be conducted in February through May 1993. As in the second pretest, sample cases consist of individuals (and their associated household members) drawn from the record systems of one of five income sources. The goal is to complete approximately 350 Wave 2 interviews under each of two randomly assigned treatments: standard SIPP measurement procedures and the redesigned procedures.

The purpose of the Evaluation Study is to provide a direct comparison of measurement quality across the two treatments, using administrative and employer records as the primary criteria for assessing quality. Program (and employment) participation and amounts as reported by the respondents will be compared to the "true" information in the records. In addition, cost component comparisons (travel time, interview time, edit time, *etc.*) will be made across the two treatments to evaluate the costs of the new procedures, and to identify the causes of any cost differences. Lastly, in addition to a simple comparison of nonresponse rates, the record data will permit some comparisons of the characteristics of nonrespondents across treatments, which may provide an indication of nonresponse bias differences between the two treatments.

4.4 Implementation Research

If the Evaluation Study yields evidence of substantial quality improvements with the new procedures, with reasonable costs and reasonable nonresponse, further research will be conducted to address the many operational issues that will inevitably remain (for example, generalizability to other sites, respondent cooperation over

[2] In both pretests, the Wave 2 reference period was shortened in order to allow the research program to meet survey redesign schedule deadlines. This aspect of the design of the pretests may have affected key results, especially those having lo do with the apparent reduction in the seam bias (see Section 5.2).

multiple waves, use of computer assisted personal or telephone interviewing, differential effects on subgroups, costs and response quality effects of the individual components of the new procedures, *etc.*). The exact design and goals of this implementation (or operational) research have yet to be specified.

5. PRETEST RESULTS

The main purpose of the first pretest, and an important purpose of the second pretest as well, was to field test the new procedures and instruments, and to identify and correct the most obvious problems. While none of the basic features of the new procedures proved infeasible in the field (and several were surprisingly successful), throughout the pretests many refinements were made to the procedures and instruments as a result of situations encountered in the field and feedback from the interviewers. The second pretest was very informative about sampling from the various record systems, as a test of these procedures for the Evaluation Study. One important finding was the frequency with which the household roster for the address supplied by the agency/employer failed to include the target sample person[3]. To account for this attrition, more sample cases were selected for Wave 1 of the Evaluation Study. Another goal of the second pretest was to test data entry procedures. That exercise, too, was very informative, pointing to the need for some important modifications for the next research phase.

The remainder of this section summarizes pretest results in three areas: the successful implementation of the new quality-oriented field procedures, indicators of improved measurement quality with the new procedures, and areas in which there is clear need for improvement - nonresponse and costs.

5.1 Implementation of the Quality-Oriented Procedures

Tape Recording. Interviewers were reasonably successful in tape recording pretest interviews, although there is certainly room for improvement. In each test, about 75% of all completed interviews were tape recorded. It is worth noting that, according to interviewers' reports, only one or two of the taping failures was attributable to respondents[4]. In almost every case, the failure to tape was due to mechanical failure, operator error, or the interviewer's failure to make the request of the respondent (which often happened in refusal conversion cases). These results offer fairly convincing evidence that tape recording is not a major issue for respondents.

On the other hand, the interviewer performance evaluation and enhancement system as a whole, the reason for the tape recording, did not work very well in the pretests. Objective evidence is scant, but there appear to have been major problems in the process of converting monitoring results into effective interviewer performance feedback. One problem was turnaround time, which was often much too protracted. A more basic difficulty, and one for which operational solutions are not immediately apparent, is interviewers' negative reactions to the monitoring. Our intent was to provide a continuous, on-the-job training system that would assist interviewers in improving their performance; interviewers, however, tended to be blind to any positive side to monitoring. Some felt that the system failed to take account of all of the interview interactions not amenable to audio taping, and viewed it mainly as a way to tally and document their errors. Interviewers did acknowledge, however, that the monitoring forms conveyed very clearly the highest priority interviewing goals and the behaviors in which we were most interested.

Group Interviews and Self-Response. The group interview and self-response procedures appear to have been quite successfully implemented in Pretest 1 (data are not yet available for Pretest 2). Three-fourths of all interviewed adults who lived in multiple-adult households participated in a group interview, and 92% of all interviewed adults self-responded. Standard SIPP procedures typically yield about 65% self-response. (Standard SIPP only allows individual interviewing, so there is no comparable group interview figure available for comparison.)

[3] Observed match rates - the rate at which the target sample person was found in the roster of household members in Wave 1 interviewed households - ranged from a low of 68% for AFDC to 96% for the employer and Unemployment Compensation samples.

[4] The fact that Wave 2 nonresponse was higher in the SIPP-CR pretests than standard SIPP typically experiences (see Section 5.3) could suggest a negative response by Wave 1 respondents to the new procedures, including, possibly, the taping requirement. However, there is no explicit mention of a problem with tape recording in any of the Wave 2 noninterview reports.

5.2 Indicators of Improved Quality

The most definitive evidence of data quality derives, of course, from the matching of survey data with the administrative and employer records. A limited set of such matched survey/record results is currently available from Pretest 2. These results, as well as two other sets of analyses drawn from both pretests - respondents' use of records, and a reduced seam bias - suggest that the revised procedures do yield improved data quality.

Record Use. Respondents' use of records in the pretests far exceeded expectations. At the household level, 87% of all households (in both pretests combined) produced at least one record to assist income reporting, with very little difference between Wave 1 and Wave 2. Record use at the income source level was 72% - that is, for 72% of the income sources respondents reported, at least one record was used to substantiate the date and amount of a payment. Similarly, at the payment level, respondents used records to report 63% of their individual payments. The Wave 2 payment-level record use rate of 74%, versus 57% in Wave 1, again suggests that, although there is still substantial room for improvement, interviewers successfully trained respondents in record maintenance between interviews[5].

Standard SIPP procedures also encourage interviewers to ask respondents to use records. According to results summarized by Singh (1991, 1992), the record use rate at the income source level was about 20% in the initial waves of the 1991 SIPP panel. Regular SIPP's rather limited success in this regard may be in part attributable to interviewers' fears that asking for records will irritate respondents, causing breakoffs and subsequent nonresponse, and will also increase interview time, thus lowering their efficiency.

Seam Bias. More direct evidence of improved quality with the revised SIPP procedures is apparent in an analysis of the seam bias. Table 1 shows an overall "seam bias index" - the ratio of the average number of month-to-month changes on the seam to the average number off the seam - for each pretest, collapsed across all income types. An index value of 1.0 indicates no seam bias; that is, the index is 1.0 if the number of transitions measured at the seam is the same as the number of transitions in an average off-seam pair of months. For the first pretest, the overall seam bias index is .95; for Pretest 2 it is a slightly higher 1.55, still substantially lower than the results reported for standard SIPP by Burkhead and Coder (1985)[6].

Table 1: Seam Bias Results for the SIPP-CR Pretests and Standard SIPP.

SEAM BIAS INDEX:	
SIPP-CR Pretest 1:	0.95
SIPP-CR Pretest 2:	1.55
Representative Seam Bias Indices for Standard SIPP (Burkhead and Coder 1985):	
Unemployment Comp:	1.9
Earnings:	2.2
Food Stamps:	3.5
Social Security:	3.9
AFDC:	4.9
Private Pensions:	6.3

We can speculate about why the new SIPP procedures appear to result in a more even distribution of reported change. Marquis and Moore (1989) have shown that the seam bias is the net result of both an underreporting of changes within an interview (off the seam) and an overreporting of changes across interviews (on the seam). The focus on individual payments in the new procedures may encourage respondents to report income receipt in all (or at least more of) its messy detail. Regular SIPP procedures, because of their focus on monthly aggregates, push respondents away from details, and toward telling a plausible, summary story. At the next interview, respondents may tell a slightly different plausible story; by this process, change may be minimized within an interview, and forced to appear at the seam. Other procedural changes which may also have

[5] A simple t-test under the assumption of sample independence is significant. Taking the correlation of the Wave 1 and Wave 2 observations into account does not change the conclusion drawn from the original test. Since some persons appear in only one wave, we re-estimated the record use proportions only including people who were in both waves. The results are very similar, so we conclude that using all available cases does not distort the difference conclusion importantly.

[6] The SIPP-CR seam bias results are based on data from all households which completed both interview waves; 74 in Pretest 1 and 79 in Pretest 2. The data reported by Burkhead and Coder are from the first three interview waves of the 1984 SIPP Panel, comprising approximately 20,000 households.

contributed to the seam bias reduction are the use of overlapping reference periods, the resolution of income source discrepancies across interviews, and the resolution of differences in reported income receipt during the overlap period. It must also be noted that the seam bias reduction is to an unknown extent an artifact of the design of the pretests, which used a shortened, two-month Wave 2 reference period instead of the four-month reference period of standard SIPP.

Underreport Errors. As noted above, there is direct evidence of the measurement error effects of the SIPP-CR procedures in the administrative and employer record data available in Pretest 2. To date, the survey reports of known program participants have been evaluated against record-based "truth" for two programs, Food Stamps and Supplemental Security Income (SSI).

Table 2 presents monthly program participation underreporting error rates - that is, the proportion of "true yes" months of participation which respondents failed to report in the survey[7]. Because of the small sample size of Pretest 2 and the vast differences in design between the pretests and standard SIPP, we have not attempted any statistical tests, and thus make no claims concerning the statistical significance of the observed differences. Nevertheless, the limited evidence again suggests that the revised procedures are moving in the right direction with regard to making important improvements in the measurement of key SIPP statistics.

Table 2: Program Participation Underreporting for the SIPP-CR Pretests and Standard SIPP.

MONTHLY PROGRAM PARTICIPATION UNDERREPORTING:

% of "true yes" months reported as "no":

	SIPP-CR	Standard SIPP
Food Stamps	9.7%	23.7%
SSI	11.1%	23.2%

(Standard SIPP results from Marquis and Moore 1990)

5.3 Areas Needing Improvement - Household Nonresponse and Costs

The pretests were not designed to provide definitive operational comparisons to standard SIPP procedures. However, pretest data suggest that, as currently designed and implemented, the new procedures may fall well short of standard SIPP performance in two key areas - nonresponse and costs.

Nonresponse. Across both pretests combined, the Wave 1 household response rate (the number of interviewed households divided by the number of eligible households) was 73%; the rate for Wave 2 (based only on Wave 1 interviewed households) was 87%, yielding a longitudinal response rate of 63%. This rate indicates the proportion of Wave 1 eligible households that were interviewed in both waves. Regular SIPP achieves a Wave 1 response rate of about 92%, and a longitudinal rate at Wave 2 of about 88%. While SIPP's rates are not exactly comparable to the pretests' due to procedural differences (for example, regular SIPP follows movers, the SIPP-CR pretest procedures did not), it is quite clear that the pretest response rates were much lower from the outset at Wave 1, and that attrition is also likely to have been higher in Wave 2.

We reviewed interviewers' descriptions of the circumstances of each noninterview they encountered for evidence that the new procedures caused the higher nonresponse. With one or two possible exceptions, there is scant evidence in these reports that any noninterview was a direct result of the new procedures. Not-at-home noninterviews, which comprised about 20-25% of the noninterview cases, are unlikely to be due to any special survey procedures, certainly not in a first interview wave. The majority of noninterviews were refusals. In almost every case, Wave 1 refusals happened before the interviewer could even begin to explain the purpose of the survey and what was involved. Although Wave 2 refusals by definition occurred with knowledge of what the interview held in store, even those refusers, according to interviewers' reports, did not implicate any of the

[7] SIPP-CR Pretest 2 sample persons experienced 165 true months of Food Stamps participation, according to the administrative records, of which they reported 149 in the SIPP-CR interview; for SSI, the comparable numbers are 135 true participation months, of which 120 were reported. The data reported by Marquis and Moore are from the SIPP Record Check Study, which used a three-state subset of the first two interview waves of the 1984 SIPP Panel. Eligible SIPP sample persons during this time period experienced 1,451 true months of Food Stamps participation, according to administrative records, of which they reported 1,107 in the standard SIPP interview; for SSI, the comparable numbers are 919 true participation months, of which 706 were reported.

cognitive procedures in their refusal behavior. People did not refuse because they were asked to get records or because they were going to be tape recorded. The pretest nonresponse problems appear to have been much more administrative in nature; potential refusals and difficult-to-locate respondents were often not identified soon enough to take effective corrective action, or, if they were identified early, followup action was often not immediate.

Costs. While it is difficult to compare the SIPP-CR pretest costs directly to the costs for regular SIPP (due to much smaller assignments in SIPP-CR, for example, and a highly clustered sample design for regular SIPP), it is quite clear that the SIPP-CR pretests experienced substantially higher field costs than those associated with the standard administration of the survey, perhaps as much as 50% higher. An obvious hypothesis is that some of the features of the new procedures - maximum self-response, group interviews, insistence upon an appropriate interview setting, the use of records, *etc.* - were responsible for the cost increases, since they required many additional visits to the households that would have been avoided under standard SIPP procedures.

We reviewed interviewers' reports of their visits to Pretest 1 Wave 1 households, all of which were supposed to have been recorded, and subjectively judged whether each would have been necessary under standard SIPP procedures, or whether it was an "extra" contact, required only to carry out the new procedures. All first visits, for example, were classified as non-extra's; all visits to obtain missing income records were "extra." The Pretest 1 visit record data do not show an unreasonable number of "extra" household visits (data from Pretest 2 have yet to be analyzed). Although an exact count of the number of extra visits is impossible, an upper limit can be assessed; we estimate that at most 14% of all Wave 1 personal visits to interviewed households were extra. While these extra visits (and the many extra telephone calls that would not have been necessary under standard SIPP procedures) undoubtedly contributed to higher field costs, they do not seem sufficient in number to explain the full cost differential.

Another contributor to the higher costs of the new procedures is actual in-house interviewing time. For the second SIPP-CR pretest, a Wave 1 interview took an average of 71 minutes per household; for regular SIPP, the average is about 52 minutes per household. This difference may be attributable to interviewer inexperience (all SIPP-CR interviewers were new to the job), or it may be due to the procedures; in either case it is unlikely to have contributed greatly to the observed cost difference.

A much clearer major cause of the higher pretest field costs was the fact that the interviewers made many unproductive visits (Krasko 1992). There was a clear avoidance of interviewing in the evenings, so interviewers made repeated daytime visits that did not yield any contact with potential respondents. Since travel costs are a major component of field costs, these non-productive visits undoubtedly contributed to the higher direct interviewing costs. It is possible that the interviewers' inexperience as survey interviewers, the fact that they did not live in their assignment areas, and the lack of emphasis on costs and efficiency (in training, supervision, and feedback), all contributed to the interviewers' making so many non-productive visits.

6. CONCLUSIONS AND NEXT STEPS

Although work on the revised, "cognitive" SIPP procedures is still very much in progress, indications from small-scale pretests are that the new procedures have the potential to substantially reduce some of the survey's important measurement problems. At the same time, the operational difficulties encountered in the pretests - high nonresponse and high costs - clearly put at risk the notion that they are a viable option for national, production implementation.

The Evaluation Study currently underway - a side-by-side experimental comparison of the new procedures and standard SIPP procedures, using administrative record data as criterion measures - will yield solid evidence about the measurement error benefits of the revised procedures. Should the measurement error results prove sufficiently positive, the Census Bureau will conduct additional research to address the many operational issues that will remain, including, of course, how to bring nonresponse and costs under control, but also the generalizability of the results to other sites, respondent cooperation over multiple waves, how best to exploit computer assisted personal or telephone interviewing with the new procedures, the differential effects of the new approach to gathering income data on subgroups (especially high income subgroups), and many other issues.

Among these "other" issues, two deserve special mention. One has to do with interviewer behavior, which the new procedures clearly push in new directions. A reasonable interpretation of the pretest nonresponse results is that while respondents show little reluctance about cooperating with the new procedures, interviewers may well be signaling some reluctance to administer them. Our evaluation of the results of the Evaluation Study must be attentive to interviewers' perceptions: what, if anything, do they find particularly onerous about the new procedures, and why? We must also be prepared to accept the possibility that the new procedures may require further revision and refinement to make them truly doable, as well as new classroom training methods, and new approaches to nurturing and supervising interviewers in the field.

The second concerns the package of SIPP-CR procedures themselves. The creation of the current cluster of procedures was driven by SIPP's redesign schedule deadlines. We were not allowed the luxury of time to develop and refine the components individually, but instead had to take an unquestionably "kitchen sink" approach. Although we can speculate, we do not know where in the package the real quality gains occur, and where the gains are not sufficient to justify the added expense. If increased costs and nonresponse continue to accompany any gains in response quality, it will be essential that the next phase of the research address the discrete cost and quality effects of the individual components of what is now the SIPP-CR measurement package.

REFERENCES

Burkhead, D., and Coder, J. (1985). Gross changes in income recipiency from the survey of income and program participation. *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, D.C., 351-356.

Krasko, N. (1992). SIPP-CR pretest I type A analysis. Unpublished U.S. Bureau of the Census memorandum for Stephen Willette, February 1992.

Marquis, K. (1990). Report of the SIPP cognitive interviewing project. Unpublished report, U.S. Bureau of the Census, August 1990.

Marquis, K., and Moore, J. (1989). Some response errors in SIPP - with thoughts about their effects and remedies. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 381-386.

Marquis, K., and Moore, J. (1990). Measurement errors in SIPP program reports. *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 721-745. Also available as Report No. 9008 in the Census Bureau's SIPP Working Paper Series (June 1990).

Marquis, K., Moore, J., and Bogen, K. (1991). A cognitive approach to redesigning measurement in the survey of income and program participation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 413-418.

Moore, J., Bogen, K., and Marquis, K. (1992). Use of unstandardized interviewing techniques in a proposed redesign of the survey of income and program participation. Unpublished paper prepared for the Joint Meetings of the Census Advisory Committees of the American Marketing Association and the American Statistical Association, U.S. Bureau of the Census, Washington, D.C., October 22-23, 1992.

Moore, J., and Marquis, K. (1989). Using administrative record data to evaluate the quality of survey estimates. *Survey Methodology*, 15, 129-143.

Murray, T. S., Michaud, S., Egan, M., and Lemaître, G. (1991). Invisible seams? The experience with the Canadian labour market activity survey. *Proceedings of the Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C., 715-758.

Singh, R. (1991). SIPP 91: Wave 1 results of the record check study. Unpublished U.S. Bureau of the Census memorandum for the SIPP Research and Evaluation Steering Committee, December 19, 1991.

Singh, R. (1992). SIPP 91: Wave 2 results of the record check study. Unpublished U.S. Bureau of the Census memorandum for the SIPP Research and Evaluation Steering Committee, June 15, 1992.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

LOCATING NLSY RESPONDENTS AND MAINTAINING A 90% COMPLETION RATE IN 13 ANNUAL ROUNDS

A. Schoua-Glusberg and E. Hunt[1]

ABSTRACT

The National Longitudinal Survey of Labor Market Experience - Youth Cohort (NLSY) began in 1979. It is currently sponsored by the U.S. Bureau of Labor Statistics, and involves attempting annual face-to-face interviews with all live respondents interviewed in 1979 (about 9800), except for members of two of the oversamples. In Round 13, 98 percent of the respondents were located, and over 91 percent interviewed. This paper examines the strategies and procedures used to track and locate respondents of this longitudinal annual survey. The successful tracking of respondents in a panel survey is one of the ways in which panel attrition can be minimized.

KEY WORDS: Locating; NLSY; Longitudinal Surveys.

1. INTRODUCTION

1.1 The National Longitudinal Survey of Labor Market Experience

I will be talking today about the strategies and procedures we follow in the National Longitudinal Survey of Labor Market Experience/Youth Cohort (NLSY) to track and locate respondents. The study is sponsored by the Bureau of Labor Statistics at the United States Department of Labor. The Center for Human Resource Research (CHRR) at Ohio State University is the prime contractor, and the National Opinion Research Center (NORC) is the subcontractor in charge of data collection.

The NLSY is an annual survey of labor market experience in which we survey young adults. For the baseyear of the survey, 1979, a total of 12,686 respondents between the ages of 14 and 21 were selected to participate in a five-year study commissioned by the Department of Labor and the Department of Defense. The original five-year survey was extended and it has been conducted on an annual basis for thirteen years. Currently we are completing data collection, within the next couple of weeks, for the fourteenth round and we are gearing up for the fifteenth round of the survey.

1.2 NLSY Completion Rates

Over the years, completion rates for the NLSY have ranged from 90-96% of the live baseyear respondents. Maintaining such high completion rates has become increasingly difficult as the years go by, although we have not necessarily seen a continuous decrease. There have been some years in which the completion rate decreased and then went back up subsequently. There have been good reasons for that. For instance, after participating in the first five rounds, a substantial number of respondents felt that they had fulfilled their original five-year commitment and did not want to continue to be interviewed. That caused a substantial drop in completion rate after the fifth year. Or, another example: in Round 9 we collected the majority of the interviews by telephone. That was the only round in which we did that, and the completion rate was lower, but it increased again in Round 10 once we returned to face-to-face interviews. In any case, we have always maintained over 90% of the initial respondents.

[1] A. Schoua-Glusberg and E. Hunt, National Opinion Research Center, 1155 E. 60th St., Chicago, IL, U.S.A. 60637.

We are constantly exploring new approaches to persuade reluctant respondents to participate and, of course, we put into effect those which seem most workable and effective. Unquestioningly, one of the great successes of the NLSY in preventing panel attrition has been the ability to track respondents through the years.

2. TRACKING AND LOCATING IN THE NLSY

2.1 Types of Respondents Followed

What type of respondents do we follow? Thinking about the "following rules" that Graham Kalton discussed this morning, we essentially go after everybody except, of course, those who have died and one other category of respondents which we call 'hostile refusals' (the respondents who threaten to call the police or their lawyers, people who point shotguns at our interviewers, and the like). Those are still very few. We do follow people who go into institutions, as long as they are capable of responding to the questions. We interview, for instance, a number of respondents in jail every year. We also follow people who move abroad; some of them we interview by telephone and some of them in person. We follow those who are at inaccessible addresses, or at least we attempt to, and we follow as well non-respondents from prior rounds.

In general, NORC has had extensive experience with locating respondents who are hard to find. Particularly, we have accumulated a great deal of experience with locating poor urban youth, a group that happens to be included in many surveys. Our experience in locating generally suggests that a comprehensive and flexible approach, one that takes into consideration the characteristics of the various sample members, is essential for success in finding the hard-to-locate.

Let us now direct our attention to the unlocatables. Who are they? What do they look like? By "unlocatable" we mean those cases which, at the end of the field period, have not been located. (The figures I will be discussing correspond to the first twelve years of the survey, since that is the extent of the publicly released data at present.) First, let us consider the fact that, of the initial sample of 12,686 respondents, there are only 1190 who have ever been unlocatable. Of those, one fourth or 25 percent are "repeat offenders" and three fourths (75 percent) have only been unlocatable once, while 81 percent have been unlocatable once or twice.

Different sample sizes have been fielded in different years. For the first few years we fielded the entire initial sample of 12,686. Then, after Round 6, we dropped the military oversample. That reduced the sample to approximately 11,600. Two years ago we dropped the oversample of economically disadvantaged whites, the "poor white" oversample (and this happened to be a group that was particularly easy to locate and interview). As a result we are now fielding about 9,800 cases. Of the fielded cases, never more than 2.5% have been coded as unlocatable at the end of a given round. In absolute numbers, the highest number of unlocatables at the end of a round has been 293 respondents.

2.2 Demographic Characteristics of the Unlocatables

Let us examine some demographic characteristics of the unlocatables. As far as gender goes, 60 percent are males and 40 percent are females, while in the sample both genders are equally represented. In terms of race and ethnicity, Hispanics are hardest to locate, which is not unexpected. Whereas they constitute 15 percent of the sample, they are 30 percent of the unlocatables. In our Hispanic oversample, we observe a great deal of movement across the border with Mexico, with many respondents coming and going frequently. We sometimes end up finding them in Mexico and interviewing them there. Blacks are 28 percent of the unlocatables and 25 percent of the sample, while whites are 41 percent of the unlocatables and 59 percent of the sample.

2.3 Locating Tools

What type of locating information do we have? Every year, we conclude the interview by updating with the respondent all the locating information in the last section of the questionnaire. We ask again for a restatement of the full name, because that sometimes changes. We ask for the respondent's full address and phone number, place of employment, whether s/he would be willing to be contacted at work. We ask women for their maiden name. We ask for driver's license number, expected moves, names of friends and relatives who are likely to always know where the respondent is. What do we do with this information? Every year, as we prepare to field the survey, all the new locating information is added to a Locating Sheet that we give the interviewer for each case she has assigned to her. Through the years, we have been forced to reduce the size of the font in which we print the Locating Sheet, as we try to make fit in one sheet of paper all of the locating information that we have collected through the years. We do not just include the most recent or best information. We also keep from each and every round of the study every piece that we have collected in terms of references, that is relatives or friends, that the respondent has given us. A few years ago we decided to drop some old references from the Locating Sheet, as new information was included, in an attempt to produce a more legible sheet. But interviewers told us they felt that move had a truly negative effect because, sometimes, a very old reference person will be a lead when everything else proves fruitless.

2.4 Prefield Locating

In addition to collecting locating information which we update with respondents every year, we follow a number of other locating steps. From one year to the next, there is a constant refining and updating of steps and procedures as more information becomes available, particularly in the area of database lookup. Let me describe the basic steps which we follow at present. One of the things that we do every year in the Central Office before attempting any type of contact with the respondent is to check electronic databases. We check the National Change of Address databases and, even when those yield Post Office updates for old addresses, we keep the old ones in the Locating Sheet. Experience has shown us that in a fraction of cases, database updates or changes are not correct. In the past, we have been "burned" by taking addresses from the new, supposedly improved, information and dropping the old. So, we now keep the old as well. That initial prefield step is essentially all we do before contacting the respondent.

In March, we do an initial mailing to the respondents, which consists of an advance letter with a coupon in which they can correct or update their address. The returned coupons are the basis for further action in Central Office. For those cases for which we do not receive a coupon back, we assume -- and we realize this is just a big assumption -- that we have mailed to the respondent's correct address. No further action is taken with those cases before the field period. The returned coupons can be classified into different groups. We receive some coupons verifying that the mailing address was indeed correct. (Although we do not specifically ask respondents to return the coupon if the printed address is correct, some respondents are just very compliant and they return it saying our locating information is still correct). Then, we receive another group of coupons with a change of address or telephone. We also receive some coupons back from the Post Office with forwarding addresses; that is, the letter is forwarded to the respondent by the Post Office, which in turn lets us know of the new address. Yet another group comes back as undeliverable because of a poor address.

Let us look, for instance, at what happened in Round 12, a typical year for locating. Excluding the "hostile refusals," we made an advance mailing of over 10,500 letters. For approximately half of those, we received nothing back. Adding all the ones that were returned with corrections, either provided by the respondent or by the Post Office, and those returned as undeliverable, one third of the advance letters mailed required some change. So, there is a substantial number of respondents who, from one year to the next, change addresses in some way. Of course, some small fraction of the cases required a correction from the respondent due to the occasional data entry error which makes us miss something in the full address. At the end of that advance mailing process in Round 12 we ended up with 783 undeliverables that we had to find in some way. This means that we were able to locate approximately two thirds of the initial undeliverables. That ratio is very consistent across rounds.

2.5 Locating during Data Collection

Another way to examine the issue of unlocatables is to look at the week-by-week status of the cases during the field period, which currently runs from late May through the end of October. At peak we have somewhere between 5 and 7 percent of the sample as unlocatable. Of those, by the end of data collection we are typically able to we find two thirds. At the end of data collection for Round 12 we had been able to locate all but 246 respondents.

Once the results of the advance mailing are tabulated, and the addresses are updated, the cases are fielded. Interviewers have several locating steps that they follow with their assigned cases. As they begin attempting to contact respondents by telephone, they start running into locating problems and disconnected telephones. They go to the respondent's last known address and are unable to find the person, or they discover some new unlocatable that they did not expect from the initial mailing.

There are a number of standard resources that we have them utilize: directory assistance, telephone directories, vital statistics records, criss-cross and reverse directories for neighbors, changes of address at the local Post Office or verification of the old address, Department of Motor Vehicle checks with the drivers' license information that we collect every year, voter registration rolls, tax assessor's records. So, these are all standard approaches that interviewers take. They also have designated field supervisors that have access to computerized databases which they can check at that stage, such as credit bureau checks and telephones' databases, as someone mentioned this morning.

A small group of respondents cannot be located by the approaches I have described, and end up requiring some in-depth sleuthing in the area of last residence or following other leads we may have from previous years. The interviewer begins contacting and locating by referring to certain aids in the case and gathers as much information as possible in the geographical area of the old address. She will talk to neighbors who may have previously been given as references by the respondents. Or she will talk to people who work in stores in the area. We try to keep these efforts to a minimum, but on occasion that is all that can be done and the results are good.

To what do we attribute the success of our locating efforts? We feel that it has a lot to do with the persistence of the interviewers and their familiarity with this sample. To some extent, we still get great cooperation from the parents of the respondents. Since the youth were between 14 and 21 years old when we first started interviewing them, we have had extensive contact with the parents through the years. In general the interviewers feel very positive about the sample, and it is a matter of pride for them to find the NLS youth.

We hear anecdotes from the field about interviewers going from lettuce field to lettuce field in California looking for undocumented respondents and finding them. In the cities, we have stories such as the one in which an interviewer remembered from a prior year that the respondent liked jazz music. She went to all the music stores in the neighborhood until she found someone who could lead her to him. Or, that interviewer who found out her respondent had became homeless and was staying in a particular wooded area. The interviewer walked by a clearing in the woods calling out the respondent's name until she found him. So, there is a lot of persistence and ingenuity there.

3. CONCLUDING REMARKS

The cost of our locating efforts needs to be addressed. While we do not keep separate track of locating activities, as opposed to other interviewing activities, we can tell that the ratio of interview time to total hours per case in the NLSY is not substantially higher than in many other of the lower budgeted surveys that we do. Thus, we deduce that not much time is being spent in locating efforts. It is a very small number of cases that get the extensive locating treatment; when that is divided across such a large sample, the impact on cost per case is not substantial.

Finally, one should not give up on unlocatables. We find that at the end of Round 12, of the 246 unlocatables we had, only 60 had been unlocatable for the prior two rounds. When we look at how respondents come in and out of the unlocatable status, we realize that sometimes people are coded as unlocatable because they are trying to avoid us, not wanting to be found. Yet we also see that, when we do find them, the majority (about 75%) do not become another kind of nonrespondents; they become respondents. So, for us, the answer is to keep looking for them, because when we find them it usually means success.

SESSION 2

Sample Selection and Weighting

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

CO-ORDINATED SELECTION OF STRATIFIED SAMPLES

F. Cotton and C. Hesse[1]

SUMMARY

This paper describes a system of co-ordinated selection of stratified samples which has been used at INSEE for several years for annual business surveys. This system is based on assigning to units in the survey base variable random numbers, recalculated after each selection according to a simple formula which depends on the type of co-ordination desired. With this method, co-ordinated stratified selections can be easily made using different criteria, even if these are different but related units. An example of this latter type of application, dealing with enterprises and their establishments, is described in detail, and a few simulations are presented to give an idea of the co-ordination effects obtained.

KEY WORDS: Co-ordination; Stratified samples; Random numbers; Business surveys.

1. INTRODUCTION

The annual business survey conducted by the French business statistics system is in fact a combination of several surveys carried out jointly by various agencies:

- departmental statistics units (surveys of industry, agri-food industries, construction and transportation);
- two sections of INSEE (surveys of commerce and services, survey of industrial microenterprises).

Enterprises to be surveyed are selected by stratified simple random sampling (SSRS). Above a size threshold which varies with the industrial sector, surveys are exhaustive. The survey base is managed and the samples are selected at INSEE (Institut National de la Statistique et des Etudes Economiques), which has the central registry of enterprises and establishments (Sirène) from which the sampling frame is derived.

This sampling system was restructured in terms of data processing and methodology in 1989. At that time, new statistical techniques were adopted, in particular a technique for selecting and co-ordinating samples by assigning random numbers, which is the subject of this paper.

2. SELECTION WITH ASSIGNMENT OF RANDOM NUMBERS

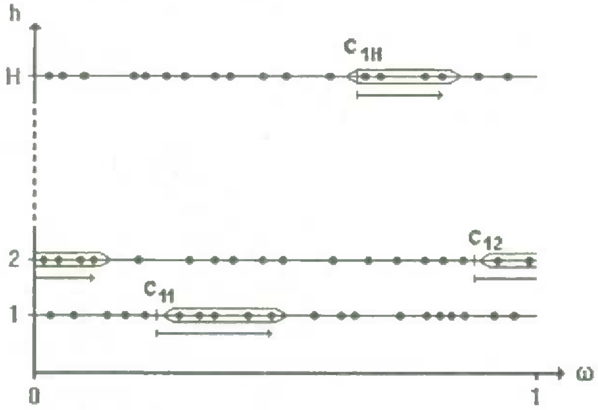
We consider a population U consisting of i individuals, $i=1,...,N$. This population is divided into $\{U_h\}_{h=1,...,H}$ strata according to a criterion known for all individuals. The objective is to take stratified simple random samples (SSRSs) from this population, defined as juxtapositions of independent simple random samples in each stratum.

To begin, we shall consider only the selection of two samples s_1 and s_2, without any intervening change in the population or the stratification. Note that n_{1h} (and n_{2h} respectively) represent the number of units we wish to select from stratum h for s_1 (and s_2 respectively).

[1] C. Hesse is head of the methodology cell in the Direction des Statistiques Économiques at INSEE, France. F. Cotton works in the Département des Projets at INSEE, 15, blvd. Gabriel Péri, F-92245 Malakoff Cedex.

We assign independently to each individual i a number ω_i drawn according to a uniform distribution of [0,1[. $(\omega_1,...,\omega_N)$ thus follows a uniform distribution of Ω = $[0,1[^N$. A simple way of selecting s_1 (Figure 1) is to arrange individuals in each stratum in increasing order of ω_i, determine for each stratum, independently of ω_i, an origin $c_{1h} \in [0,1[$, then select from stratum h the first n_{1h} individuals whose random number ω_i is greater than or equal to c_{1h}. Throughout this paper we shall reason in modulo 1. This means in the present case that if we reach the end of the stratum before selecting n_{1h} units, we shall return to the origin of the ω_i's and thus select as a complement the first individuals such that $\omega_i + 1 \geq c_{1h}$. We call such a sequence of units arranged by ω_i a selection window.

Figure 1: Stratified selection by assignment of random numbers.



Each horizontal line represents a stratum, and each point a unit. Units are arranged according to their random numbers. Enclosed units are selected when the technique described above is applied.

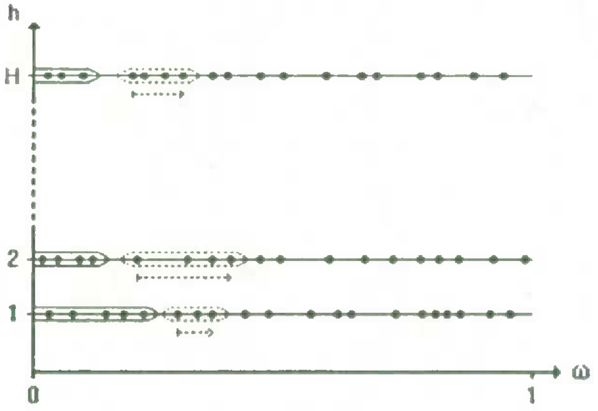
Obviously, it is possible to make our selection "towards the left" of c_{1h}, in other words in the direction of decreasing ω_i: we then take the last n_{1h} individuals such that $\omega_i < c_{1h}$ (still modulo 1). Unless otherwise stated, we shall assume here that samples are selected "towards the right".

We select s_2 as s_1, on the basis of a set of origins $\{c_{2h}\}_{h=1,...,H}$. We can obtain co-ordination effects between s_1 and s_2 by c_{1h} and c_{2h}, or the direction of selection (toward the left or toward the right). Let us suppose, to illustrate this point, that all the origins (c_{1h}) are zero and that s_1 is drawn toward the right. These assumptions do not cause any loss of generality.

We can co-ordinate s_1 and s_2 positively, in other words maximizing in s_2 the number of units already drawn by s_1 by choosing $c_{2h} = 0$ for all h. In this case, by the way, it is not necessary for the stratification of s_1 and of s_2 to be identical.

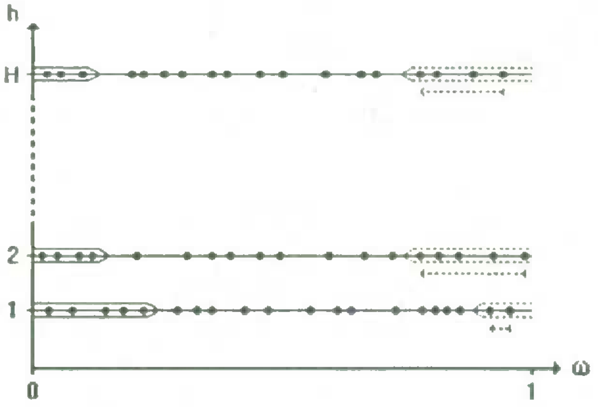
To obtain a negative co-ordination, we must move the selection windows after the selection of s_1 for s_2. Figures 2 and 3 describe two possible methods. We see that the method in Figure 2 does not stand up well to a change in stratification between s_1 and s_2, since it "desynchronizes" the origins of the selection windows. In the example in this figure, if the last unit selected for s_1 from stratum 1 moves to stratum 2 before s_2 is selected, it will be selected again for s_2: negative co-ordination will not be maximized. On the contrary, the choice in each stratum of a window consisting of the last n_{2h} units of the stratum leads to maximum negative co-ordination between the two samples. This situation is illustrated in Figure 3, where the second sample is selected with windows having the same origins as the first (zero, in this case), but in the direction of decreasing ω_i.

Figure 2: Negative co-ordination of two samples.



Since the first sample was selected with zero origins (solid lines), a second sample can be co-ordinated negatively using selection windows whose origin comes after the last unit drawn in the first selection (dotted lines).

Figure 3: Negative co-ordination of two samples.



Since the first sample was selected with zero origins (solid lines), a second sample can be co-ordinated negatively using selection windows located at the end of strata (dotted lines).

3. RENUMBERING UNITS

Now let us consider two SSRSs which can adopt different stratifications. When we seek negative co-ordination, we may, instead of changing the selection window between the two selections, keep the window but "move" the units by changing their random numbers ω_i.

We note $\alpha_{h1}(j), j \in \{1,...,N_h\}$, the jth random number in increasing order in stratum h of the first selection: this is random number ω_i for an individual i, where i may be specified as a function of j and h. This gives us, almost certainly, all different random numbers:

$$\alpha_{h1}(1) < ... < \alpha_{h1}(N_h).$$

We have selected for s_1 individuals such that $j = 1,...,n_{1h}$, and we are looking for a renumbering procedure:

$$\begin{aligned} v_{hl} : [0,1[^h &\rightarrow [0,1[^h \\ \alpha_{hl} &\rightarrow \beta_{hl} \end{aligned}$$

associating with individual i of number $\alpha_{hl}(j)$ a new number $\beta_{hl}(j)$, such that the classification according to this new number places individuals selected by s_1 at the end of the stratum. INSEE chose to use the transformation v^1_{hl}:

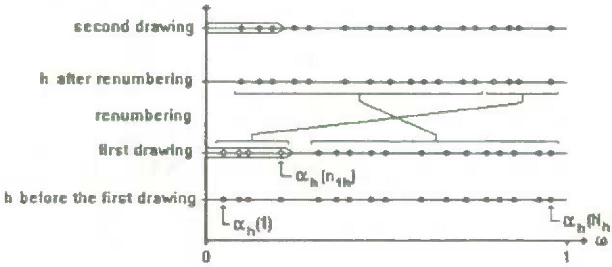
$$\begin{cases} \beta_{hl}(j) = \alpha_{hl}(j) + \alpha_{hl}(N_h) - \alpha_{hl}(n_{hl}) & j \leq n_{hl} \\ \beta_{hl}(j) = \alpha_{hl}(j) - \alpha_{hl}(n_{hl}) & j > n_{hl}\,. \end{cases} \tag{1}$$

Other specifications are possible, such as transformation v^p_{hl} by permutation of numbers:

$$\begin{cases} \beta_{hl}(j) = \alpha_{hl}(j + N_h - n_{hl}) & j \leq n_{hl} \\ \beta_{hl}(j) = \alpha_{hl}(j - n_{hl}) & j > n_{hl}\,. \end{cases} \tag{2}$$

The sample s_2 is obtained by selecting units whose new numbers are at the beginning of the strata of s_2. Figure 4 provides an illustration of this method of co-ordination with renumbering.

Figure 4: Renumbering of units.



This graph illustrates the evolution over time of the distribution of units in a stratum *h*. The units selected in the first sample (below) are relocated to the end of the stratum by renumbering, while those which were not selected are displaced toward the beginning of the stratum, and are thus selected as a priority when the second sample is taken (above).

Note that it is possible, by renumbering, to relocate to the end of the stratum only a portion of the individuals selected for the first sample. We have only to replace n_{1h} in formula (1) with a smaller integer. This yields partial co-ordination effects between samples.

Transformation v^1_{hl} possesses good properties[2]. In particular, it does not affect the law of probability on Ω. Moreover, the joint selection $p(s_1,s_2)$ by this method is identical to the one which consists of using for s_1, without changing the numbers, the windows located at the start of the strata (of s_1), and for s_2 the windows located at the end of the strata (of s_2). We thus obtain maximum negative co-ordination of the two SSRSs.

The selection method with renumbering can easily be extended in the event that the number of samples to be drawn is greater than 2. In particular, it allows maximum rotation of units in successive SSRSs, with changes of stratification. We have only to follow these sequences:

[2] See F. Cotton and C. Hesse, Tirages coordonnés d'échantillons, INSEE working paper No E9206.

- select (first time) or renumber ω_i;
- restratify the population;
- sample the units at the beginning of the strata.

In the event that the number of samples is greater than 2, renumbering has an advantage over the technique of selecting s_2 from the end of the stratum in figure 3: with the latter method, we do not know what windows to use for s_3 in order to ensure negative co-ordination with both s_1 and s_2.

4. APPLICATION: SELECTING UNITS FROM DIFFERENT LEVELS

A number of co-ordination effects between samples can be achieved by simply manipulating the random numbers associated with units, provided that we do not alter the uniform distribution of those numbers.

This point can be illustrated in another situation. Let us consider a sampling frame containing two related types of units, for example enterprises and their establishments. It may sometimes be useful to co-ordinate selections from these two types of units, with a view to distributing the respondent burden among survey subjects.

Suppose that a selection p_1 has been made from enterprises using the linear renumbering method. We now wish to proceed with a selection p_2 of establishments, but avoiding, if possible, the selection of establishments belonging to enterprises included in p_1. After renumbering, such enterprises have fairly high ω_i values. Their establishments must also have high random numbers so as not to be selected at the start of their stratum by p_2. The idea is thus to connect the enterprise number with the smallest of its establishment numbers.

If, on the other hand, p_1 covered establishments, we might wish, in a selection p_2 of enterprises, to avoid those whose establishments were selected by p_1. Consequently, when one of the establishment numbers is high, we want the enterprise number to be high also, which suggests a link between the enterprise number and the largest number of the enterprise's establishments.

This link must preserve the uniform distributions of enterprise numbers ω_i and of establishment numbers ω_{ij}. For this purpose, we use the property that if $F(x)$ is the distribution function of a random variable X, $F(X)$ follows the uniform distribution of [0,1[. Let n be the number of establishments of enterprise i (to be absolutely correct, n should be indexed by i, since the number of establishments may vary from enterprise to enterprise).

In the direction establishments → enterprise, to calculate enterprise numbers initially from establishment numbers or to reflect back on enterprises a renumbering of establishments, we may formulate:

$$\omega_i = [\max(\omega_{i1},\ldots,\omega_{in})]^n, \text{ or:} \tag{3}$$

$$\omega_i = 1 - (1 - \min(\omega_{i1},\ldots,\omega_{in}))^n . \tag{4}$$

We refer to these as max link and min link respectively.

The direction enterprise → establishments is used to reflect a renumbering of enterprises on establishments, for example. We can thus maintain the min link if we attribute to the establishment which initially has the smallest number the new value $1 - (1 - \omega_i)^{1/n}$, whose law is that of the minimum of real n selected according to a uniform distribution from [0,1[.

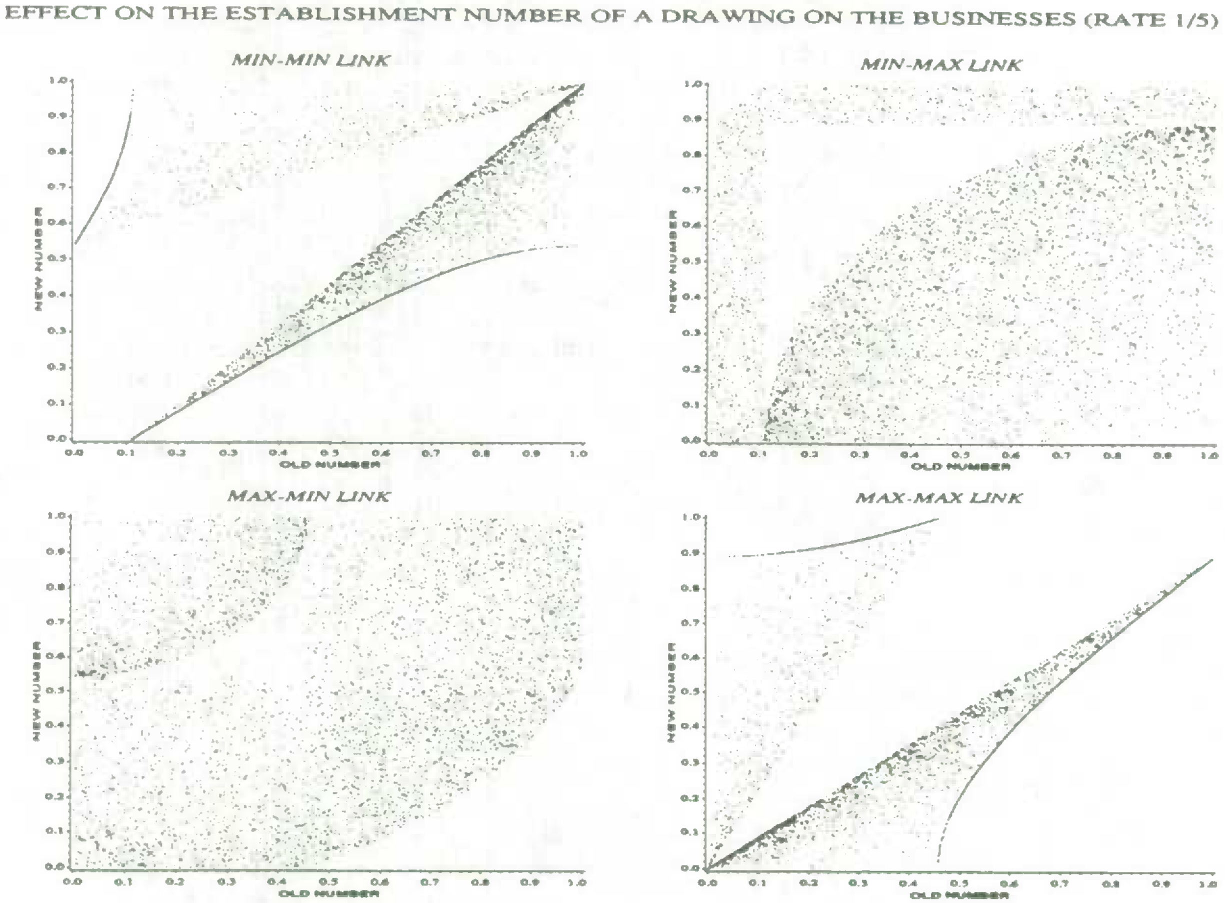
In order to assign the new numbers of the enterprise's other establishments, we use the property that, conditionally on the minimum of ω_{ij}, higher ω_{ij} must also follow a uniform distribution between this minimum and 1 in order for their marginal distribution to be uniform. We can thus either select these numbers according to a uniform distribution between this minimum and 1 or distribute them in this interval with spacing proportionate to that of their previous distribution (which is supposed to be uniform already).

The enterprise → establishments max link is established by assigning to the establishment which initially had the largest number the new number $\omega_i^{1/n}$, and proceeding in a manner similar to the min link to calculate the new numbers of the other establishments.

Figures 5 and 6 present simulations for simple cases which allow us to visualize the co-ordination effects obtained.

In Figure 5, we consider an unstratified population of 10,000 enterprises with two establishments. A number selected at random between 0 and 1 is assigned to the establishments. The random number of the enterprise is calculated on the basis of the establishment numbers by one of the above formulas. We then make a simple random survey of enterprises and then renumber those enterprises. The establishment numbers are then recalculated, again using a link by the min or by the max.

Figure 5: Simulations for 5,000 enterprises with two establishments.

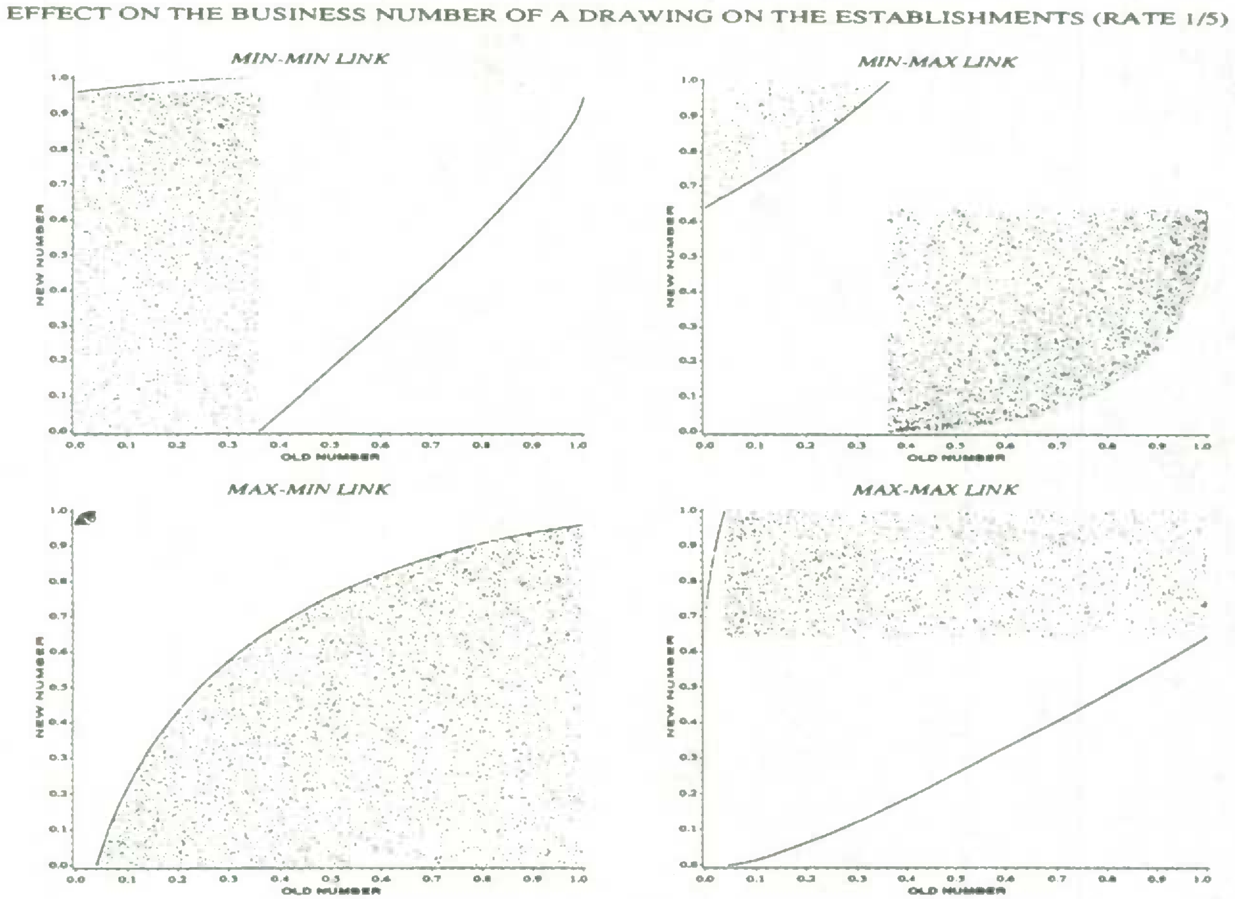


We have traced on the figure the correspondence between the former establishment numbers and the new numbers for each of the possible links (min-min, min-max, max-min and max-max). Each establishment is represented by an abscissa point for its former number and an ordinate point for its new number. The four figures are composed of two sets of points: the first, situated to the upper left, corresponds to establishments whose enterprises were selected; the second, to the lower right, corresponds to establishments whose enterprise was not selected. The distinction is especially sharp in the case of min-min and max-max links. We observe that establishments whose enterprise was selected have their numbers increase, in other words their probability of selection in the next sample of establishments decrease. The effect is reversed for establishments whose enterprises were not selected.

Which link to choose depends on the co-ordination effect we want. In the case of the min-min link, all establishments whose enterprises were selected have a high new number (it is easy to calculate that this is greater than $1 - \sqrt{\tau}$, where τ is the sampling rate, 1/5 in our example). This ensures good negative co-ordination with the next sample selection. On the other hand, co-ordination with any previous sample of establishments is not so good, since some high abscissa establishments have their enterprise selected. Conversely, with the max-max link, negative co-ordination is good with the preceding selection, but not as good with the selection following, since some establishments whose enterprises are selected have a new number close to the origin. The max-min link ensures good co-ordination with both the preceding and the following selection, but presents other disadvantages, in particular making greater changes to the numbers of establishments whose enterprises are not selected.

In Figure 6, for the same population as previously, we simulate a sampling of establishments, which thus changes the numbers of establishments by renumbering. Each point represents an enterprise: its abscissa is the enterprise's initial number, calculated from the numbers of its establishments before the selection, and its ordinate is the enterprise's number recalculated from the establishment numbers after the selection. Each image consists of three areas, which are alternately curves and clouds of points, corresponding to enterprises for which both establishments were selected (upper left), those for which only one establishment was selected, and lastly those for which no establishments were selected (lower right).

Figure 6: Simulations for 5,000 enterprises with two establishments.



We observe from the figure that enterprises for which both establishments were selected show a marked increase in their number, and thus the probability that they will be selected subsequently is largely reduced. The same effect, although less marked, can be observed among enterprises for which only one establishment was selected. The other enterprises decrease in number. Here again, we can compare the advantages and disadvantages of the various possible links. This time, the best compromise seems to be the min-max link.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

SELECTION AND MAINTENANCE OF A HIGHLY STRATIFIED PANEL SAMPLE

J.L. Czajka and A.L. Schirm[1]

ABSTRACT

From its 1987 sample of individual tax returns the U.S. Internal Revenue Service designated 90,000 returns to initiate an annual panel. The sample is highly stratified by income, with sampling rates ranging from .02 percent to 100 percent. Evidence from the first three years indicates that interstratum movement has produced significant changes in sample composition. This paper explores issues arising from the dynamic behavior of a highly stratified panel sample, including implications for sample selection, the development of weights for cross-sectional and longitudinal estimation, strategies for sample supplementation, and effects on the precision of sample estimates over time.

KEY WORDS: Income; Post-stratification; Weighting.

1. INTRODUCTION

This paper addresses a problem in panel sample design and analysis that has received very little attention in the literature. The problem is caused by the use of stratification in selecting a panel sample-more specifically, the combination of:

- stratification on characteristics, such as income, that are not fixed over time;
- widely varying selection probabilities.

Under these circumstances, the composition of a panel is not fixed with respect to the variables defining the stratification. The dynamic behavior of the panel has implications for the quality of estimates derived from the panel data and introduces a dimension of complexity into panel design that is not present for cross-sectional samples.

We encountered this problem in working with an administrative record panel of U.S. individual (as opposed to corporate) tax returns, developed by the Statistics of Income (SOI) Division of the U.S. Internal Revenue Service (IRS). This sample of 90,000 filing units employs a very high level of stratification, with sampling rates covering a range of four orders of magnitude. Using this sample for illustration, this paper explores some of the consequences of compositional change and considers the implications for sample design.

This paper is organized as follows. Section 2 discusses uses of panel tax data and provides a description of the SOI individual panel. Section 3 describes changes in the composition and characteristics of the panel over time. Section 4 demonstrates the implications of these changes for the precision of sample estimates. Section 5 discusses two general strategies-reweighting and sample supplementation-to address the effects of compositional change. Section 6 discusses approaches to panel sample design that reflect a longitudinal perspective on sample selection. Finally, Section 7 presents some concluding observations.

[1] J.L. Czajka and A.L. Schirm, Mathematica Policy Research, Inc., 600 Maryland Avenue, SW, Suite 550, Washington, DC, U.S.A. 20024-2512.

2. THE SOI INDIVIDUAL PANEL

To appreciate the potential implications of panel composition change for both analysis and design of longitudinal data, it is important to understand not only how panel data are collected but how they are intended to be used. As background for the rest of the paper we discuss possible uses of panel tax data and then describe the design of the SOI individual panel.

2.1 Uses of Panel Tax Data

Panel data drawn from tax returns have many possible applications where cross-sectional data will not suffice. Examples of questions that may be addressed with panel tax data are listed below.

- How does family dissolution affect the income and filing characteristics of former members?
- What is the shape of the income profile over different parts of the life cycle? And how does this vary across broad income classes?
- How did the relative incomes of different population segments change over a period of time?
- What is the serial pattern of capital gains realization?
- How soon are taxpayers able to use losses carried over to subsequent tax years?
- How do charitable contributions respond to a change in income or marginal tax rates?
- How do taxpayers use the tax credit associated with a net operating loss, which can be applied to earlier or later years?
- What conditions account for taxpayers' use of the "married filing separately" status?

Some applications involve the investigation of inherently longitudinal phenomena-for example, life cycle changes in income. Other applications involve the study of behavior that can be investigated with cross-sectional data but for which longitudinal data are really more appropriate-for example, responses to changes in tax rates. Still other applications involve the study of cross-sectional events-for example, the use of a particular tax credit-for which key explanatory variables may be found in prior years.

Of course, panel data may have cross-sectional applications as well-particularly when important data items are collected only from a panel sample. While this is not true of the SOI panel at present, as there is a comparably sized annual cross-sectional sample, other tax data panels have been supplemented with data not collected in the cross-sectional sample, and there are plans to supplement the SOI panel as well.

2.2 Design of the SOI Panel Sample

Each year the SOI Division draws a sample of individual tax returns from the population of returns processed during that calendar year. The stratified random sample is quite large, frequently exceeding 100,000 returns. With the 1987 tax year the SOI Division initiated a panel of nearly 90,000 filing units. The base year sample was selected from the 1987 cross-sectional sample and is representative of nondependent tax returns processed by IRS in 1988. The returns filed by panel members have been captured in each year following the panel's inception.

Mirroring the cross-sectional sample from which it was selected, the base year panel sample is characterized by sharply differential selection probabilities by stratum. The sample is stratified by type of return and income or receipts. Returns are classified into seven types, defined hierarchically-that is, a return is assigned to the first return type for which it qualifies. Return types and their corresponding stratum numbers are listed below:

28	High income, nontaxable;
38	High business net profit or loss;
80-84	Foreign earned income (Form 2555);
90-94	Foreign tax credit (Form 1116);
60-68	Sole proprietorship, business (Schedule C);
50-58	Sole proprietorship, farm (Schedule F);
40-48	Nonbusiness, nonfarm (all other returns).

The first two return types are sampled at 100 percent. Sampling rates for the remainder are a function of income or (for business and farm returns) total receipts. There are five income classes for "foreign" returns and nine classes for the others. For example, among nonbusiness, nonfarm returns stratum 40 includes returns with incomes under $25,000; these returns were sampled at a rate of .04 percent. Stratum 48 includes returns with incomes of $5 million or more; these returns were selected with certainty. For a complete description of the panel strata see Schirm and Czajka (1992).

3. CHANGE IN SAMPLE COMPOSITION OVER TIME

3.1 Net Change

We now consider net change in the sample composition with respect to the variables that defined the selection strata. Specifically, if we were to select another sample one or two years after the panel base year and employ the same stratum definitions, but with current year income, where would the panel members fall with respect to these later year strata?

Table 1 reports the distribution of panel returns in the base year and two subsequent years (1988 and 1989) by current year sample stratum for panel members selected in the base year as nondependents. The final two columns indicate for 1988 and 1989 the percentage change from the base year.

In the upper portion of the table we observe large net changes in the frequency of certain types of returns. The number of high income nontaxable returns (stratum 28) is down 70 percent after one year and 75 percent after two years. Returns with high net profit or loss (stratum 38) are down 41 percent after one year and 56 percent after two. On the other hand, the number of returns with foreign tax credit (strata 90-94) is boosted substantially after two years, from 54 to 442 percent depending on the income class.

In the lower part of the table we see large changes at the highest income levels. Returns with incomes of $1 million or more (codes ending in 6, 7 or 8) are generally down 40 percent after two years while those with incomes between $500 thousand and $1 million (codes ending in 5) are up between 50 and 60 percent.

3.2 Gross Change

The net changes seen in Table 1 are the result of substantially greater gross changes. Table 2 indicates for each base year stratum the percentage of panel members located in another stratum in 1988 and 1989. Overall, 38 percent of the returns in 1988 and 49 percent in 1989 reside in a different stratum than the one from which their filers were first selected. Of the filers selected in 1987 with high income nontaxable returns, 80 percent have left that status by 1988 and 87 percent by 1989. Similarly, 60 percent of the panel members with high business profit or loss in 1987 no longer meet that definition in 1989. Among the lower three groupings of strata we find that the likelihood of moving to another stratum is strongly related to the level of income in 1987. For example, more than 60 percent of the returns originating in strata 43-48 have moved to another stratum by 1989 compared to only 14 percent in stratum 40.

Table 1: Panel Returns by Current Year Sample Stratum, 1987-1989: Panel Members Selected as Nondependents.

Current sample stratum	Number of returns			Change from 1987	
	1987	1988	1989	1988	1989
Total	89,755	86,353	87,318	-3.8%	-2.7%
28	873	259	220	-70.3	-74.8
38	9,590	5,635	4,202	-41.2	-56.2
80	29	37	48	27.6	65.5
81	7	29	42	314.3	500.0
82	120	158	158	31.7	31.7
83	167	101	79	-39.5	-52.7
84	39	29	16	-25.6	-59.0
90	50	96	141	92.0	182.0
91	55	171	298	210.9	441.8
92	531	992	1,345	86.8	153.3
93	957	1,222	1,656	27.7	73.0
94	747	1,083	1,149	45.0	53.8
60	3,089	2,901	3,651	-6.1	18.2
61	3,527	3,557	3,959	.9	12.2
62	3,763	3,779	3,707	.4	-1.5
63	2,291	2,517	2,595	9.9	13.3
64	1,896	3,208	3,529	69.2	86.1
65	1,078	1,678	1,734	55.7	60.9
66	1,732	1,514	1,483	-12.6	-14.4
67	1,684	1,350	981	-19.8	-41.7
68	985	812	609	-17.6	-38.2
50	259	301	495	16.2	91.1
51	493	517	619	4.9	25.6
52	374	375	396	.3	5.9
53	177	227	315	28.2	78.0
54	337	363	373	7.7	10.7
55	198	288	320	45.5	61.6
56	554	331	331	-40.3	-40.3
57	626	370	267	-40.9	-57.3
58	176	121	104	-31.3	-40.9
40	19,548	17,523	19,321	-10.4	-1.2
41	10,757	11,635	12,395	8.2	15.2
42	7,909	8,728	7,669	10.4	-3.0
43	2,559	3,176	3,059	24.1	19.5
44	3,176	3,417	3,299	7.6	3.9
45	1,415	2,141	2,187	51.3	54.7
46	3,343	2,211	2,087	-33.9	-37.6
47	2,939	2,135	1,487	-27.4	-49.4
48	1,705	1,366	990	-19.9	-41.9

Table 2: Percent of Panel Members in Different Cross-Sectional Stratum Than the Base Year.

Base year stratum	1988	1989
Total	37.7%	48.9%
28	79.5	87.3
38	43.4	59.7
80	29.6	60.0
81	57.1	57.1
82	32.7	57.4
83	48.4	66.0
84	48.6	78.4
90	45.5	63.3
91	69.1	66.7
92	52.4	61.6
93	55.4	62.8
94	50.0	53.8
60	35.0	40.3
61	41.6	54.6
62	40.8	57.1
63	48.6	63.2
64	49.2	64.0
65	64.5	76.6
66	69.4	78.6
67	63.5	79.6
68	57.2	70.8
50	29.2	28.1
51	34.5	41.4
52	44.5	62.8
53	52.4	69.1
54	50.5	69.9
55	60.2	77.2
56	66.7	79.0
57	62.5	75.9
58	71.4	81.4
40	13.5	14.4
41	26.3	35.8
42	26.7	45.8
43	46.1	63.1
44	46.8	60.9
45	59.9	71.9
46	65.7	74.5
47	62.2	76.7
48	53.4	65.5

Table 3 provides a cross-tabulation of the 1987 by 1989 cross-sectional stratum for panel members who were selected from nonbusiness, nonfarm strata and have remained within this return type. Because of the sharply differential sampling rates by income class, downward movement is much better represented than upward movement. Thus, while we observe panel members who were selected from stratum 48 (income of $5 million or more) distributed among all seven lower income strata in 1989, we observe essentially no movement into that stratum from a distance of more than two income classes. In general, we rarely see in the panel sample any increases in income sufficient to cause a movement of more than two classes.

Table 3: Interstratum Mobility between 1987 and 1989: Strata 40-48.

1987 stratum	1989 Cross-sectional stratum								
	40	41	42	43	44	45	46	47	48
40	14,833	1,381	135	11	*	*	*	*	*
41	2,361	6,940	622	18	5	*	*	*	*
42	396	2,292	4,362	210	14	*	*	*	*
43	79	159	881	951	153	10	*	*	*
44	54	93	280	803	1,242	175	30	5	*
45	26	26	69	101	387	396	120	17	4
46	43	39	116	192	399	674	848	247	31
47	44	41	85	101	235	293	562	692	198
48	23	12	23	36	113	89	119	250	594

* Cell count is three or fewer.

4. LOSS OF PRECISION

Leaving aside issues of population coverage, the SOI panel will provide unbiased estimates of means and totals when panel records are weighted by the inverses of their base year selection probabilities. However, increases in within-stratum variances due to changing characteristics of panel members contribute to a loss of precision relative to the base year (or to a current year cross-sectional sample of equal size). This loss of precision will occur for any characteristic that is related to the original stratification.

Table 4 reports coefficients of variation (CVs) for estimates of aggregate amounts of selected income and tax items for the panel sample and the SOI cross-sectional sample for 1988 and 1989. The CVs for each cross-sectional sample have been adjusted to reflect a sample size and composition by stratum that is comparable to the base year panel sample.

Comparing the cross-sectional CVs with CVs for the panel sample weighted according to the 1987 stratum assignments alone (the design-based weights), we find that for five items in each year the panel CV ranges from nearly double the cross-sectional CV to three or four times its size. Of particular note, in 1988 the panel CV for adjusted gross income (AGI), which is closely related to the income concept used in the SOI sample design, is more than double the CV for the cross-sectional sample. In 1989 the panel CV is three times the size of the cross-sectional CV. Comparable relationships hold for total tax liability.

5. STRATEGIES TO COMPENSATE FOR CHANGING PANEL COMPOSITION

Two general strategies are employed to compensate for change in the size and composition of a panel sample after the fact. One strategy involves reweighting the panel observations. The second involves supplementing the sample with additional observations.

Table 4: Coefficients of Variation (Percent) for Estimates of Aggregate Amounts of Selected Income and Tax Items, 1988 and 1989.

Item	1988 Cross-sectional sample	1988 Panel sample by stratification		1989 Cross-sectional sample	1989 Panel sample by stratification	
		1987 strata alone	Post-stratified by 1988		1987 strata alone	Post-stratified by 1989
AGI or deficit						
Income	.15	.31	.15	.15	.44	.21
Deficit	2.28	2.95	2.60	2.47	3.17	2.75
Salaries and wages	.27	.29	.24	.28	.37	.25
Taxable interest	1.19	1.00	.97	.99	.99	.95
Dividends	1.72	2.81	1.76	1.55	1.72	1.52
Pensions/annuities in AGI	1.73	1.47	1.41	1.39	3.20	1.44
Business net profit or loss						
Profit	1.36	1.46	1.24	1.29	1.58	1.28
Loss	3.24	3.47	3.17	3.16	3.77	3.46
Net capital gain or loss						
Gain	1.08	3.18	1.98	1.20	4.93	3.61
Loss	2.79	2.16	2.24	2.40	2.15	2.28
Supplemental gain or loss						
Gain	4.93	6.67	5.98	5.29	6.01	5.67
Loss	7.56	9.24	9.11	7.75	8.99	8.35
Net schedule E income or loss						
Income	1.42	2.72	1.83	1.53	3.62	1.79
Loss	1.63	1.78	1.65	1.74	2.67	2.02
Total itemized deductions	.60	.53	.51	.55	.87	.55
Total tax liability	.24	.56	.26	.27	.73	.37

5.1 Reweighting

Post-stratification provides one means of adjusting panel sample weights to compensate for changes in the panel's composition. For the SOI panel we have seen evidence of significant movement between the income categories from which panel returns were originally selected and the income categories to which they would have been assigned for cross-sectional sample selection in subsequent years. Post-stratifying on current year cross-sectional stratum membership, using readily available population totals, could improve the precision of cross-sectional estimates.

The second of each pair of panel CVs in Table 4 reflects the results of post-stratification to current year population totals (the same totals used in weighting the cross-sectional sample estimates). Post-stratification produces an appreciable reduction in the estimated CV for every item where we observed a large difference between the cross-sectional and panel CVs. For AGI, dividends, and total tax liability in 1988 and for pensions and annuities in 1989, post-stratification reduces the panel CV to a level comparable to that of the cross-sectional sample. For net schedule E income in both years the results are nearly as striking.

In short, even the very simple post-stratification tested here substantially compensates for the adverse effects of panel composition change on the precision of cross-sectional estimates. Where large differences remain, post-stratifying to additional population totals may yield further improvements in precision. If this is insufficient, we would suggest consideration of more fully model-based methods of estimation.

5.2 Sample Supplementation

Panel designs sometimes add observations to compensate for unrepresented growth in their target populations. A more common feature is outright replacement of one panel with a new panel. The Survey of Income and Program Participation, for example, features a 2-1/2 year panel life, with new panels introduced annually. Respondent burden and attrition are probably more important considerations than population coverage. With an administrative record panel, however, respondent burden is generally not an issue, and attrition is limited mainly to exits from the target population. With the potentially longer life of an administrative panel, strategies to deal with coverage become more important, and we assert that strategies to deal with changing composition become important as well.

A general strategy for dealing with changing panel composition might include supplementing the panel at certain intervals to compensate for the loss of observations with particular characteristics. In the SOI panel, for example, the loss of high income returns could be addressed by adding new high income observations every year or two. In earlier work we questioned the value of sample supplementation as a strategy to address emerging deficiencies in the population coverage of a panel sample, partly on the grounds that the missing segments are very difficult to sample efficiently (Czajka and Schirm 1992). However, if the goal of sample supplementation is to add observations to well-defined segments, such as sample strata, replacement *can* be accomplished efficiently (in this case by simply replicating the original selection procedures with current year characteristics).

Collecting prior year data on these new observations will benefit the panel in two ways. First, it will enable the use of these observations to improve sample representation of upward transitions. Second, it will reduce problems related to the analysis of a panel sample with differential panel lifetimes.

6. PANEL SAMPLE DESIGN CONSIDERATIONS

The strategies discussed above are used to address the complications of panel composition change after such change has occurred. How might we design a panel sample that incorporates needed stratification yet is less exposed to the negative consequences of changing composition?

6.1 Mid-Life Selection

One solution that is suggested by our discussion is to select a panel to achieve a desired composition at the panel's *mid-life* as opposed to the base year. There are alternative ways to accomplish this. One is to select the panel in the middle year of its expected life. Having done so, one could then collect data back to a desired beginning retrospectively and collect future data prospectively. This approach would enable study of transitions into and out of strata of interest. The Treasury Department produced such a panel from tax return data (Hubbard, Nunns and Randolph 1992). Another approach is to backcast from the desired mid-life composition, using estimated transition probabilities, to define a base year sample that will age into the desired mid-life composition. The viability of this approach is contingent upon: (1) knowledge of the relevant transition probabilities and (2) the elimination of small probabilities by conditioning. For example, with the SOI panel the attainment of a target mid-life sample for, say, stratum 28 would involve transition probabilities that were very small, implying enormous base year samples. To achieve the desired sample size for stratum 28, we would need to identify substrata with much higher transition probabilities, so that we could oversample these. All too often we may not know the relevant conditions or may not be able to replicate them with the data available for stratification.

6.2 Panel Sample Design from a Longitudinal Perspective

The approach just described still takes an essentially cross-sectional perspective. Alternative approaches are suggested by considering the problem of panel sample design from a longitudinal perspective-in essence, adding a time dimension to the population that we wish to sample. If in drawing a panel sample we had access to longitudinal information on prospective sample units, how would we design the sample?

Considering some of the uses of longitudinal tax data discussed above suggests alternative stratifying variables for an *ideal* longitudinal sample of tax returns. Three such variables are cumulative income, transitions between income classes (or between "statuses" generally), and occurrences of "events" (for example, the use of tax credits or married couples filing separate returns). With a retrospective survey we could draw a sample on such characteristics, but the screening costs would be prohibitive, and the limitations of retrospective survey data are well-known. For an administrative sample the necessary linked data may not exist or may be impossible to create. What other options do we have? Let us consider some possibilities for each of the three types of longitudinal variables.

6.2.1 Sampling Cumulative Income

One approach to selecting a sample on the basis of cumulative income is to attempt to predict cumulative income from cross-sectional income. This might be done by calculating predicted cumulative income as a weighted sum of multiple income components, where the more "stable" components are weighted more heavily than the less stable components. There is an analog to this approach in the "permanent income" concept employed by economists, which is often operationalized as a regression-based prediction of income from several variables reflecting human capital and other variables.

Using cumulative income as the stratifier has both advantages and disadvantages. One advantage is the downweighting of volatile income components. It is possible that volatile income components such as capital gains or partnership income may account for the base year sample selection of high income filing units that subsequently dropped to much lower income levels. Another advantage is the utilization of stratifying variables as predictors of cumulative income that themselves predict or condition change.

One disadvantage is that the ability to anticipate persons acquiring new income sources or experiencing large changes (particularly positive ones) may be weak. A second disadvantage is that good predictors of cumulative income may not be available for stratification. For the SOI panel, being restricted to predictors reported on the tax return is obviously very limiting; direct measures of human capital are not present.

6.2.2 Sampling Transitions

We have noted that the stratification of the SOI panel greatly overrepresents downward transitions in income class relative to upward transitions-the more so the larger the transition. The stratification of a panel should reflect the relative importance of different types of transitions to the researchers who will use the data. The broader the interest in transitions, the less overall stratification is indicated.

Efficiently oversampling particular types of transitions generally requires knowledge of the factors that condition these transitions. For example, which low income taxpayers will account for most of those who experience large growth in income? To oversample low to high income transitions requires an ability to differentiate among filing units with respect to their probability of realizing large increases in income. If we had such ability we would probably become millionaries ourselves.

6.2.3 Sampling Events

We have shown that the stratification of the SOI panel provides substantially fewer high income nontaxable returns and high net profit/loss returns in the out years than in the base year. We may use these return types to illustrate the problems associated with the panel representation of "events" over time. Research needs may require relatively equal representation of such events over time. Achieving this objective requires an ability to identify the population segments from which such rare groups of taxpayers in future years might originate, so that they can be oversampled. Again we come back to our (lack of) knowledge of variables that predict or condition change and their (in)availability for use as potential stratifiers.

7. CONCLUSION

If the stratification of a panel sample includes variables with nonfixed values, then change in the characteristics of the sample may lower the precision of estimates over time and substantially reduce the observations available for the study of particular behaviors of interest to the analyst. We have shown that post-stratification on "current" values of original stratifiers can improve appreciably the precision of panel-based estimates of cross-sectional characteristics in years after the base year. We have suggested that sample supplementation to increase the representation of certain current characteristics may be needed to satisfy sample size targets in the out years. We have argued as well that it is important to think about panel sample design from a longitudinal perspective and have provided some examples of how this might be approached. To incorporate a longitudinal perspective into the design of a panel sample requires a prioritization of the analytical objectives of the analysts who will use the data. Depending on these priorities, various alternative design strategies may be applicable.

By way of additional discussion, we suggest that the material we have presented carries some implications for the duration of a panel sample. Even if sample attrition is not a significant issue, the inability to address some of the adverse consequences of composition change (or to achieve the ideal design) will limit the useful life of a panel for some purposes. Yet this need not be true of all segments of the panel. Consequently, we recommend the consideration of multi-duration designs, where some portions of the original panel are retained for longer periods of time while other segments are retained for as few as three periods of observation. The longer segments of such a panel would serve certain research objectives while successive short segments serve other objectives. For example, the study of life cycle phenomena would require panel observations over a relatively long period of time while the study of, say, filing status transitions might require (or be served most efficiently) by only a relatively short duration.

ACKNOWLEDGMENTS

This research was funded by the SOI Division of IRS, and for this support we are grateful. We would also like to thank William Randolph and Robert Gillette of the Office of Tax Analysis for sharing their thoughts on prospective uses of panel tax data; Roderick J. A. Little of UCLA and Datametrics for helpful comments and suggestions on the topics of precision and post-stratification; and Daisy Ewell of Mathematica Policy Research for skillfully executing many tough programming assignments.

REFERENCES

Czajka, J.L., and Schirm, A.L. (1992). Enhancing the representativeness of a longitudinal sample of individual tax returns: Weighting and sample supplementation. In *Proceedings of the Eighth Annual Census Bureau Research Conference.*

Hubbard, R.G., Nunns, J.R., and Randolph, W.C. (1992). Household income mobility during the 1980s: A statistical assessment based on tax return data. U.S. Department of the Treasury.

Schirm, A.L., and Czajka, J.L. (1992). Weighting a panel of individual tax returns for cross-sectional estimation. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

WEIGHTING FOR THE SURVEY OF LABOUR AND INCOME DYNAMICS

P. Lavallée and L. Hunter[1]

ABSTRACT

In 1994, Statistics Canada will introduce a large scale household panel survey. The Survey of Labour and Income Dynamics (SLID) will follow individuals and households, tracking their labour market activities and changes in income and family circumstances. In addition to providing longitudinal information, cross-sectional estimates will also be produced. In this paper, the sampling scheme for SLID is first described. Second, we present the determination of the basic weights which correspond, for most individuals, to the inverse selection probabilities. Third, we discuss non-response adjustments and the post-stratification to be used for SLID. We finally give a brief conclusion together with future plans.

KEY WORDS: Longitudinal survey; Cross-sectional estimation; Selection probabilities; Post-stratification.

1. INTRODUCTION

In 1994, Statistics Canada will introduce a large scale household panel survey. The Survey of Labour and Income Dynamics (SLID) will follow individuals and households, tracking their labour market activities, and changes in income and family circumstances. The main goal of the survey is to provide longitudinal information. Annual estimates (often referred to as cross-sectional estimates) will also be provided.

The present paper addresses the problem of making the SLID longitudinal sample representative for cross-sectional estimation. In order to attain this goal, an appropriate weighting scheme needs to be developed. The objective is to achieve the best weighting approach in order to satisfy criteria such as unbiasedness and efficiency, as well operational feasibility. Despite the fact that cross-sectional estimation is mainly considered here, it should be kept in mind that longitudinal studies are generally cross-sectionally representative of the selection year of their longitudinal samples. Hence, longitudinal and cross-sectional weighting will both, in some sense, be treated within the present paper.

Cross-sectional weighting is often discussed in the survey sampling literature since at lot of surveys are cross-sectional. In the present case however, the situation is slightly complicated by the longitudinal aspect of the survey. The sample is not reselected independently at each wave of interviewing but is one where most units have been selected on previous occasions with no major update to the sample. Some special considerations must therefore be taken into account because of the longitudinal aspect of the sample.

In the present document, the sampling scheme for SLID will first be described. Second, the determination of the selection probabilities will be discussed. Third, non-response adjustments and post-stratification will be described. Fourth, a method to improve the estimates will be presented based on an approach suggested by Gouriéroux and Roy (1978). Finally, a brief conclusion will be provided, including future plans.

[1] P. Lavallée and L. Hunter, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2. SAMPLING SCHEME

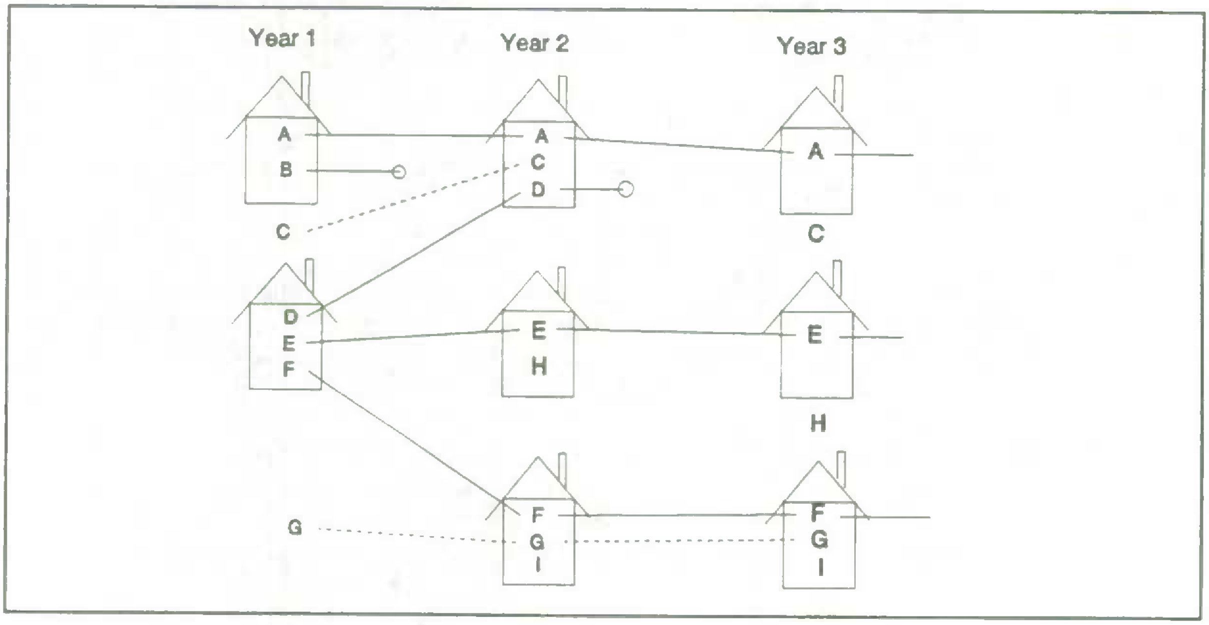
The SLID target population is defined as all persons, regardless of age, living in the provinces of Canada, excluding the persons living in the Territories, institutions, Indian reserves, and military barracks.

For 1993, the SLID sample will be a subsample of the Canadian Labour Force Survey (LFS). The LFS produces monthly estimates on total employment, self-employment and total unemployment. It uses a multi-stage stratified sample design based on an area frame with dwellings as ultimate sampling units. All persons belonging to the households in the selected dwellings are included in the LFS sample. The sample uses a rotating group design where each month one of six rotation groups is replaced after staying six months in the sample. Each rotation group contains approximately 10,000 households, which represents about 20,000 individuals. For more details about the LFS sample design, see Singh *et al.* (1990).

For 1993, the SLID sample will be selected from two groups that are rotating out of the LFS. The SLID sample will therefore be a subsample of the LFS. This sample, or panel, will contain about 20,000 households. In 1996, a second panel of 20,000 households will be selected to form a total sample of 40,000 households.

SLID will follow individuals through time, but household characteristics are also of interest. All persons belonging to the households in the selected dwelling will be initially selected for the SLID sample. In later interviews, people living with the individuals selected in the initial sample will also be interviewed in order to obtain data for the complete household. These people will be either new entrants or cohabitants. *New entrants* are individuals joining the target population and selected in the SLID sample. They are represented by H and I in Figure 1. For example, a new entrant might be a newly born person in 1994, or a person who came from a foreign country. *Cohabitants* are individuals joining the SLID sample but are not new entrants in the target population. A cohabitant is in fact an individual who was part of the initial year's target population but was not selected then. For example, a cohabitant might be a person who got married to someone in the SLID sample after the initial sample was selected. In Figure 1, C and G represents cohabitants. Individuals who exit the target population, such as B and D in Figure 1, are called *leavers*.

Figure 1: Interview of individuals within households.



New entrants and leavers will need to be considered to maintain the cross-sectional representativity. Note that the longitudinal component of SLID will not be updated after its selection. All new individuals (added, for example, by virtue of living with longitudinal individuals) will only be considered for cross-sectional purposes.

One difficulty with new entrants is how they will be added to the sample. Some new entrants will be selected by including them in the sample as they join a *longitudinal household*. A household is considered as being longitudinal if it contains at least one individual who is part of the longitudinal sample. This method however will not include households with only new entrants. It is planned to get households with new entrants through a sample of dwellings obtained from the ongoing LFS. Note that by being a sub-sample from the LFS, the SLID sample was initially selected from a sample of dwellings. The dwellings used for the selection of new entrants could be the original dwellings used in the selection of the initial sample or a new set of dwellings selected independently each year from groups rotating out of the LFS. Reselecting dwellings or not does not affect the statistical characteristics of the sample but makes a difference with respect to operating costs. Using the sample of original dwellings is costly because it requires visiting each dwelling every year to find new entrants; if dwellings are obtained from those rotating out of the LFS, the identification of new entrants can be done directly during LFS interviews. However, to simplify the discussion, it will be assumed that *originally selected dwellings* will be used to select new entrants.

3. BASIC WEIGHTING

One issue in weighting of the SLID sample is the determination of the *basic weights*. For most individuals, these correspond to inverse selection probabilities. The *basic weights* are in fact the weights to be used in the estimation process before any adjustment or post-stratification is applied. The problem of determining the basic weights is complicated by the fact that cohabitants and new entrants can be part of the sample at any wave of interviewing by joining a longitudinal household.

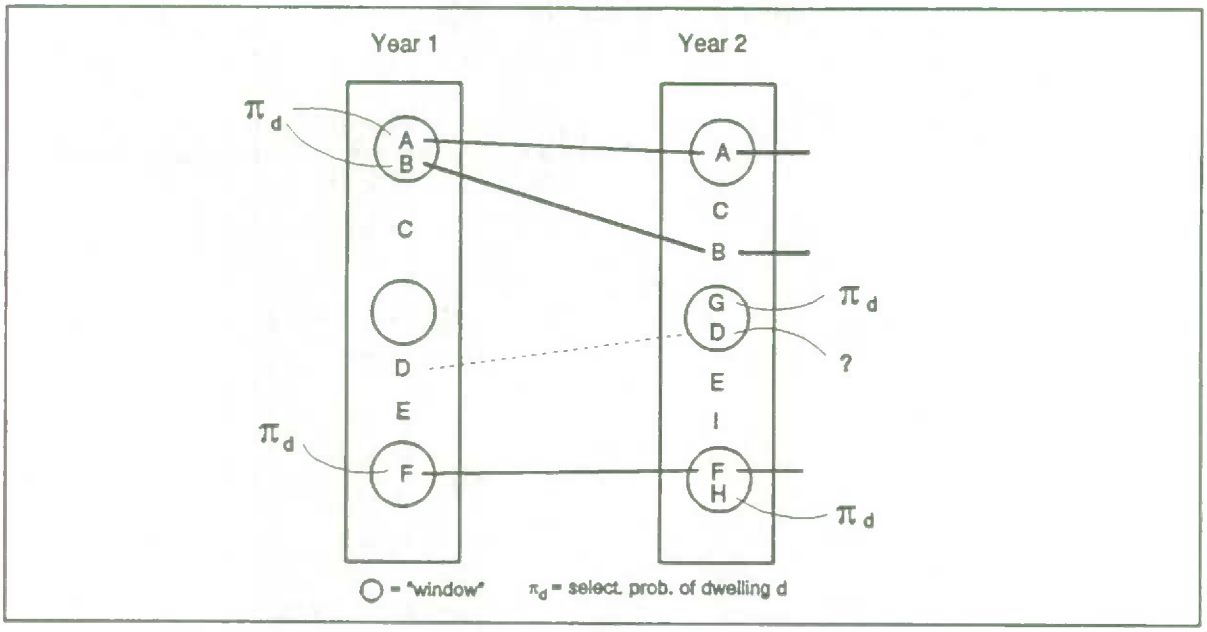
In the first year, individuals are selected through a sample of dwellings selected from the area frame of LFS. Then every individual j of the selected dwelling d receives the selection probability π_d. Their basic weights are the inverse selection probabilities. In the following years, individuals originally selected keep the selection probabilities of the first year. For each individual j, we therefore have the selection probability $\pi_j^{(O)}$ set to π_d.

For cross-sectional purposes, the sample will be updated by new entrants and cohabitants. New entrants entering into the sample through an original dwelling d will receive the selection probability π_d of the dwelling. This is exactly as if the new entrants had been selected at the time of the initial selection. Note that if a new entrant is selected by joining a longitudinal household which is living in an original dwelling d, he will also receive the selection probability π_d of the dwelling. One can in fact see the original dwellings as "windows" to enter the sample, as shown in Figure 2. The individuals selected in the sample through these selected "windows" receive the probability of selection π_d of the dwelling d and can thus be considered to be legitimately selected in the sample. These individuals will be referred to as *legitimate individuals* (originally selected individuals plus new entrants living in an original dwelling).

At some year after the initial selection, consider an original dwelling in which a longitudinal household is now living. This household might be cohabitants and/or new entrants in addition to the longitudinal individuals. By being selected the initial year, the longitudinal individuals certainly have selection probabilities. (See letters A and F in Figure 2.) The new entrants receive the selection probability of the dwelling since the household is living in an original dwelling. (See letters G and H in Figure 2.) On the other hand, cohabitants do not have the selection probability of the dwelling because they were not selected in the original sample, even though they are now living in an original dwelling. (See letter D in Figure 2.) Anyone who is a member of the target population can enter the sample through a "window" only at the time the initial sample is selected; only new entrants to the target population can enter the sample through a "window" in subsequent waves. For the case of a selected household living in an original dwelling, only longitudinal individuals and new entrants are therefore legitimate individuals.

Consider next a longitudinal household which is not living in an original dwelling. As before, this household might contain cohabitants or new entrants along with the longitudinal individuals. Because the household is not living in an original dwelling, the new entrants do not receive the selection probability of the dwelling. Cohabitants again do not have the selection probability of the dwelling because they were not selected in the original sample. For the case of a longitudinal household not living in an original dwelling, only longitudinal individuals are therefore legitimate individuals.

Figure 2: Selection of individuals through "windows".



The selection probabilities for non-legitimate individuals are not clearly identifiable. With new entrants joining a longitudinal household which is not living in an original dwelling, it is not clear which selection probabilities should be assigned to these new entrants. For cohabitants, there is also a problem in assigning the selection probabilities since the only reason they are added to the sample is because they have joined a longitudinal household, even if the household is living in an original dwelling.

By being in the target population the starting year, each cohabitant has, in theory, a selection probability. However, it is usually unknown. We can illustrate this by the following example: In a multi-stage design, the cohabitants could have been part of a primary sampling unit (PSU) which was not part of the sample. Since the unselected PSUs are generally not visited, the selection probabilities of the individuals contained in these PSUs remain unknown. It should be noted that even if the selection probabilities of cohabitants could be determined precisely, there would be a need to update the selection probabilities of all the sampled individuals to reflect the fact that new individuals are now part of the sample. This process would be difficult and costly. Therefore, cohabitants are always considered as non-legitimate individuals.

In order to solve the problem of assigning selection probabilities (or basic weights) to new entrants not living in an original dwelling and also to cohabitants, two approaches have been proposed. The first approach, which we will call the Share Approach, is the one partially described by Ernst (1989). The weighting for the Survey of Income and Program Participation (SIPP) is inspired from the Share Approach. The second is the one proposed by J.N.K. Rao which makes use of composite estimation. These two approaches are presented below.

3.1 The Share Approach

Basically, this method inspired by Ernst (1989) assigns to each selected non-legitimate individual a basic weight obtained from an average of weights computed within each household. An *initial weight* corresponding to the inverse selection probability is first obtained for each legitimate individual. Second, for each non-legitimate individual, an *initial weight* of zero is assigned. The basic weight is then computed by taking the average of the initial weights at the household level. Finally, the basic weight is redistributed to all individuals within the household. It should be noted that having all individuals with the same basic weight has the desirable property of providing consistency in the estimates of individuals and households.

The Share Approach can be presented formally by distinguishing between two cases.

Case 1: Households occupying an originally selected dwelling *d*

Suppose the selected household *i* of year 2 occupies an original dwelling *d*. Let household *i* contains $M_i^{(O)}$ individuals selected at year 1 (*i.e.* original longitudinal individuals), $M_i^{(C)}$ cohabitants, and $M_i^{(N)}$ new entrants. As we described earlier, new entrants are assigned the selection probability of the dwelling *d*. For each individual *j* of household *i* in the original dwelling *d*, we assign the initial weight

$$w_{ij}' = \begin{cases} 1/\pi_j^{(O)} & \text{for } j=1,\ldots,M_i^{(O)} \\ 0 & \text{for } j=(M_i^{(O)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}) \\ 1/\pi_d & \text{for } j=(M_i^{(O)}+M_i^{(C)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}+M_i^{(N)}) . \end{cases} \tag{1}$$

Then the basic weight w_i of household *i* is obtained by

$$w_i = \frac{1}{M_i}\sum_{j=1}^{M_i} w_{ij}' , \tag{2}$$

where $M_i = M_i^{(O)} + M_i^{(C)} + M_i^{(N)}$. The basic weight w_i is finally assigned to each individual of the household.

Case 2: Households not occupying an original dwelling

Let us consider the selected household *i* which does not occupy an original dwelling. Then, for each individual *j* of household *i*, we assign the initial weight

$$w_{ij}' = \begin{cases} 1/\pi_j^{(O)} & \text{for } j=1,\ldots,M_i^{(O)} \\ 0 & \text{for } j=(M_i^{(O)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}) \\ 0 & \text{for } j=(M_i^{(O)}+M_i^{(C)}+1),\ldots,(M_i^{(O)}+M_i^{(C)}+M_i^{(N)}) . \end{cases} \tag{3}$$

The basic weight $w_{i'}$ of household *i* is obtained according to equation (2). Again, the basic weight w_i is assigned to each individual of the household.

The estimate $\hat{Y}$ of the population total *Y* is finally given by

$$\hat{Y} = \sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij} y_{ij} , \tag{4}$$

where *n* is the total number of selected households and $w_{ij}=w_i$ for all individuals *j* in the *i*th selected household. By proceeding as Ernst (1989), this estimate can be shown to be unbiased.

3.2 The Composite Estimation Approach

The Composite Estimation Approach proposed by J.N.K. Rao focuses more on the estimator formulation than on the weighting of each of the selected individuals. It in fact uses a model-based approach to construct an estimator for non-legitimate individuals. Basically, selected individuals are divided into two sets D_L and D_{NL}, depending on whether they are legitimate or not. An estimator of the population total *Y* is then produced from each of the two sets. The estimate from the set D_{NL} of non-legitimate individuals is obtained by assuming that these are selected according to stratified simple random sampling. A composite estimator is finally formed with the two estimates, weighted by the relative number of individuals of each set.

The idea behind the composite estimator is the following: Normally, only the legitimate individuals would enter into the estimation of cross-sectional quantities such as the total *Y*. However, because the complete households are interviewed to obtain household estimates, leaving out the non-legitimate individuals would correspond to ignoring some information. The idea then is to take into account the non-legitimate individuals, but only in proportion to their relative importance.

Set D_L: Legitimate individuals

Each individual j of the set D_L is weighted according to the inverse of its selection probability π_j. An estimate $\hat{Y}_L$ of the population total Y is then obtained as follows:

$$\hat{Y}_L = \sum_{j \in D_L} \frac{y_j}{\pi_j} \, . \tag{5}$$

This simply corresponds to the Horvitz-Thompson estimator.

Set D_{NL}: Non-legitimate individuals

It is supposed that the sample of D_{NL} corresponds to a stratified simple random sample selected from a superpopulation which could be described by the following model:

$$Y_j = \mu_h + \epsilon_j \ \ if \ j \in h \, , \tag{6}$$

where $E(\epsilon_j)=0$, $\mathrm{Var}(\epsilon_j)=\sigma_h^2$ and $\mathrm{Cov}(\epsilon_j,\epsilon_{j'})=0$ for $j \neq j'$. The strata h are assumed to be at an aggregate level, such as the provinces.

A second estimate $\hat{Y}_{NL}$ of the population total Y is then computed as follows:

$$\hat{Y}_{NL} = \sum_{h=1}^{H} \frac{M_h}{m_{NL,h}} \sum_{j=1}^{m_{NL,h}} y_{hj} \, , \tag{7}$$

where M_h is the population number of individuals in stratum h, $m_{NL,h}$ is the number of sampled non-legitimate individuals in stratum h, and H in the total number of strata.

The final estimate of the total Y is finally obtain from a composite estimator of the form

$$\hat{Y}_c = \frac{1}{(m_L + m_{NL})} (m_L \hat{Y}_L + m_{NL} \hat{Y}_{NL}) \, , \tag{8}$$

where m_L and m_{NL} are the number of individuals in D_L and D_{NL}, respectively.

The composite estimator $\hat{Y}_c$ provides an initial weight to each individual j of household i as follows:

$$w'_{cij} = \begin{cases} \dfrac{m_L}{(m_L + m_{NL})} \dfrac{1}{\pi_j} & \text{for } j \in D_L \\[2ex] \dfrac{m_{NL}}{(m_L + m_{NL})} \dfrac{M_h}{m_{NL,h}} & \text{for } j \in D_{NL} \text{ and } j \in h \, . \end{cases} \tag{9}$$

These weights could be directly used for estimation of different characteristics y. However, for providing consistency in the estimates of individuals, it is desirable to obtain a single weight within each household. It is therefore suggested to average the initial weights w'_{cij} within each household to obtain basic weights.

$$w_{ci} = \frac{1}{M_i} \sum_{j=1}^{M_i} w'_{cij} \, . \tag{10}$$

This weight is finally assigned to each individual, legitimate or not, within household i.

The composite estimator $\hat{Y}_c$ turns out to be design biased. Its bias comes from $\hat{Y}_{NL}$ mainly because the sample from D_{NL} is not a stratified simple random sample. One can also argue that the second estimate $\hat{Y}_{NL}$ might not be really representative of the current population Y because non-legitimate individuals constitute a population of "sociable" persons by being in the set D_{NL} only because of joining a non-empty household. That is, "non-sociable" persons who will never live with another person have a zero probability of being part of non-legitimate individuals. If they are acting differently from the legitimate individuals, then a bias might be introduced in the estimation. Note that the bias of $\hat{Y}_{NL}$ is likely to be small because of the relatively small sample size m_{NL}, but will increase over time as m_{NL} increases.

3.3 Simulation Results

In order to verify which of the two approaches, the Share Approach or the Composite Estimation Approach, is better, simulations had to be done. Fortunately, the sample design and labour questions for SLID are similar to an existing survey called the Labour Market Activity Survey (LMAS). The data from the LMAS was therefore used to perform the simulations.

The LMAS was conducted on two panels, each time using 5 rotation groups from the LFS as its sample. The first panel was interviewed for two years (1986-87); the second panel was interviewed for three years (1988-89-90). For the three-year panel, information was also collected from new entrants and from cohabitants, as is being proposed for SLID. Unlike SLID, however, the LMAS was designed to have separate but overlapping samples for its cross-sectional component and its longitudinal component. For the longitudinal sample, individuals sampled in the first year were traced and re-interviewed in the second (and third) year. For the cross-sectional sample, the originally selected dwellings were re-visited in the second (and third) year, and any in-scope individuals living there were interviewed. Therefore, originally selected individuals who did not move contributed to both the longitudinal and the cross-sectional data; originally selected individuals who did move contributed to the longitudinal data only; and people who were not selected in the first year, but who moved into an original dwelling contributed to the cross-sectional data only. The ideas behind the weighting adjustments for SLID are that instead of having a separate cross-sectional sample, the same degree of representativity can be achieved using the longitudinal individuals plus their cohabitants. The second panel of LMAS data was used to test this hypothesis.

To start, the first two years of data (1988-89) were used. Each person was classified as an original longitudinal person, a new entrant into the population, or a cohabitant. Each household was classified as living in an originally selected dwelling, or living in a non-sampled, or non-original, dwelling as a result of a longitudinal person moving. On the 1989 LMAS file, 49,874 individuals were both in the longitudinal and cross-sectional samples, 8,016 were longitudinal respondents only, 14,874 were cross-sectional respondents only, and 2,355 individuals were cohabitants with the longitudinal respondents only. From the 60,245 individuals in the longitudinal sample, 82.8% were also in the cross-sectional sample.

A starting weight was assigned to each longitudinal person, based on the selection probability associated with his/her dwelling in the first year. A ratio adjustment was then done within stratum-components to adjust for non-response to the second year interview. These non-response adjusted weights were used as the initial weights going into the different weighting approaches described in 3.1 and 3.2. For the Composite Estimation Approach, the stratification for the non-legitimate individuals was assumed to be at the province level. A final post-stratification to province-age-sex group totals was done for each approach. Final weights were calculated under the proposed approaches for longitudinal persons and their cohabitants who responded in 1989. Estimates were finally calculated using these weights and were compared with the actual LMAS estimates from the 1989 cross-sectional data.

Estimates of population proportions were calculated by marital status, number of jobs held in 1989, number of weeks employed/unemployed/out-of-labour-force during 1989, as well as average weeks employed/unemployed/out-of-labour-force. The results are presented in Table 1. In general, the Composite Estimation Approach performed as well as the Share Approach in that the estimates were close to the actual cross-sectional estimates. The two versions of the Composite Estimation Approach (using different and same weights for different household members) showed very little difference. Because of its unbiasedness property, the Share Approach might then be more suitable than the Composite Estimation Approach.

Table 1: Comparison of National Estimates, ages 25-64, LMAS 88/89 (post-stratified by province-age-sex).

	Cross-Sectional File (1989)		Share Approach Longit. File Estim.	Composite Est. Approach Longitudinal File Estim.	
	estimate	c.v. (%)		unaveraged weights	averaged weights
average wks employed	38.0	0.5	38.5	38.4	38.3
average wks unempl1	2.0	2.7	1.9	2.0	2.0
average wks out of LF	12.7	1.3	12.2	12.3	12.4
weeks empl = 0 (%)	19.5	1.4	18.5	18.5	18.4
weeks empl = 1-26	8.2	2.7	8.1	8.2	8.2
weeks empl = 27-48	10.8	2.2	10.8	10.9	10.9
weeks empl = 49+	61.5	0.7	62.6	62.3	62.2
weeks unempl1 = 0 (%)	88.0	0.3	88.5	88.3	88.3
weeks unempl1 = 1-26	9.1	1.9	8.8	9.0	8.9
weeks unempl1 = 27-48	2.5	4.4	2.4	2.4	2.4
weeks unempl1 = 49+	0.4	15.7	0.3	0.3	0.3
weeks out of LF = 0 (%)	64.4	0.1	65.2	65.0	64.9
weeks out of LF = 1-26	12.5	6.7	12.6	12.7	12.6
weeks out of LF = 27-48	5.0	17.1	4.8	4.9	4.9
weeks out of LF = 49+	18.1	10.7	17.4	17.4	17.6
number of jobs = 0 (%)	19.5	1.4	18.5	18.5	18.7
number of jobs = 1	66.5	0.5	69.7	69.5	69.4
number of jobs = 2	11.4	2.1	9.8	9.9	9.9
number of jobs = 3+	2.5	4.4	2.0	2.0	2.0
MS=married (%)	74.6	0.5	76.8	76.6	77.5
MS=single	15.7	2.3	13.6	13.7	12.9
MS=widowed	2.0	4.9	2.1	2.1	2.1
MS=separated/divorced	7.7	3.0	7.5	7.6	7.6

LF=Labour Force, MS=Marital Status

4. NON-RESPONSE ADJUSTMENTS AND POST-STRATIFICATION

The present section deals with adjustments that will be done to the basic weights to improve estimation. These adjustments are of the form of a multiplying factor referred to as g-weights (see Särndal *et al.* (1992)). That is, for each individual j of household i, the basic weight w_{ij} is multiplied by some adjustment factor to obtain the final weight $w_{ij}^F = g_{ij|s} w_{ij}$. The g-weight $g_{ij|s}$ usually depends on the weighting procedure and the actual samples. Unfortunately, simulations are not yet available.

4.1 Non-response adjustments

As for all surveys, SLID will be faced with non-response. We can expect to have both item non-response and unit non-response, as well as wave non-response and hard core non-response. Furthermore, data can be missing for individuals within a household or for complete households. Corrective measures will therefore be needed to deal with these complex varieties of non-response.

Whenever possible, imputation will be used for non-response. The main requirement for imputation is that at least one member of the current household responded. As a consequence, no imputation is likely to be performed for complete household non-response. As an example, a new household formed by an individual leaving a selected household will not be imputed if that individual did not respond, because we have no knowledge of that individual's current household composition. For these cases where imputation is not feasible, a household non-response adjustment will be done to the basic weights.

One method of constructing a weight adjustment for non-response is to use response (or non-response) modelling. With response modelling, a set of assumptions is made about the true unknown response mechanism of the survey. In particular, logistic regression is often a convenient method (see Little (1986) and Hunter, Michaud and Torrance (1992)). The multiple logistic response function is given by

$$E(R_i|\underline{x}_i)=[1+\exp(-\underline{\beta}'\underline{x}_i)]^{-1}\,, \tag{11}$$

where R_i is the dependent variable, $\underline{\beta}$ is column vector of regression parameters and $\underline{x}_i$ is a vector of independent variables available for all the households. The dependent variable R_i equals 1 if household i is a respondent, and 0 otherwise. Thus, $E(R_i|\underline{x}_i)$ can be seen as the response probability $\theta_{i|s}$, which can depend on the sample s.

After estimating $\underline{\beta}$ by maximum likelihood, we obtain an estimated response probability for an household with value $\underline{x}_i$:

$$\hat{\theta}_{i|s}=[1+\exp(-\hat{\underline{\beta}}'\underline{x}_i)]^{-1}\,. \tag{12}$$

After correcting the basic weight w_i of household i, the adjusted basic weight w_i^A is given by

$$w_i^A = w_i \times \frac{1}{\hat{\theta}_{i|s}}\,. \tag{13}$$

Note that the estimated response probabilities can be used to form response homogeneity groups (RHG) in which all sample elements are assumed to have the same probability of response, which is then simply estimated by the response rate within each RHG (see Särndal *et al.* (1992)).

One problem with the use of response modelling for SLID is the availability of auxiliary variables to be used as dependent variables $\underline{x}_i$ with the logistic response function (11). Because households for which a non-response adjustment will be those with no response from current household members, it is likely that almost no information will be available on the current household. For example, using regions and/or household size as auxiliary variables would be useless for those households which were not traced. Therefore, household non-response adjustment through response modelling does not seem to be practically feasible for SLID. It is hoped in fact that most household non-response will be corrected through post-stratification.

4.2 Post-stratification

In sample surveys, post-stratification is done for two main reasons. First, it is used to correct for under-representation in the sample for certain sub-populations. This under-representation may be due, for example, to non-response. In particular, if the response probabilities $\theta_{i|s}$ obtained from (11) depend on the same variables used to define the post-strata, post-stratification will explicitly correct for non-response (see Särndal *et al.* (1992)). Second, if the variables of interest are found to be homogeneous in certain classes (or post-strata), then post-stratifying with respect to those classes will help to improve the precision of the estimates.

SLID being a subsample from LFS, it is convenient to consider at least the same post-stratification variables. This set of post-stratification variables might then be augmented with other appropriate variables. LFS estimates are actually post-stratified based on regions and age-sex groups (see Singh *et al.*). This post-stratification is done at the individual level, but using an integrated approach which produces an equal weight for all members within a household (see Lemaître and Dufour (1987)).

The post-stratification variables under consideration for SLID are: regions, age, sex, revenue and interprovincial mobility. These variables will not be totally crossed because the population totals are not available, and also because this would produce a large number of cells including many cells with no sample. Two sets of population totals are being tested: one by cross-classifying regions, age-sex groups and revenue classes, and the other by cross-classifying province, age-sex groups and interprovincial mobility.

Post-stratifying by revenue is mainly to correct for underrepresentation of low and high revenues. On the one hand, it is found that households with low income tend to have a higher mobility than households with medium or high revenue, which makes them more difficult to trace. Once traced, they also have a higher tendency to not respond, especially to sensitive topics such as income. On the other hand, households with high income are often non-respondents because they are unwilling or unable (*e.g.* their finances are handled by an accountant) to provide the information. Post-stratification by revenue will require revenue classes which will be determined through simulations using LMAS data with the addition of an income variable corresponding to the "Income subject to tax" variable of the Survey on Consumer Finances. The population totals to be used for stratification will be obtained from the tax data of Revenue Canada / Taxation. The counts obtained from tax data will be adjusted to take into account coverage problems brought by the fact that not all individuals fill an income tax return.

By post-stratifying according to interprovincial mobility, we make the assumption that the movers behave differently from the non-movers. This post-stratification is mainly done to improve the precision of the estimates. Note that it might also be correct for household non-response caused by tracing problems. Post-stratification by mobility might be difficult to achieve if the post-strata are defined by crossing the complete set of provinces of origin and provinces of destination. We might then consider constructing the post-strata according to the province of destination only. For post-stratifying the province of Manitoba, for example, we would then only distinguish the individuals according to whether they just moved into the province (*i.e.* movers) or not (*i.e.* non-movers). This is a much simpler post-stratification than crossing the complete set of provinces of origin and provinces of destination but it might not be as efficient if the individuals behave quite differently based on their province of origin. The population totals to post-stratify by mobility are available from the Demography Division of Statistics Canada.

Once the post-stratification variables have been determined, the implementation will be done using either the integrated approach of Lemaître and Dufour (1987), which is actually used by LFS, or the calibration estimators developed by Deville and Särndal (1992). These last estimators cover the integrated approach of Lemaître and Dufour (1987) but also allow some control on the weights to avoid negative values. Calibration estimators have been studied for LFS by Stukel and Boyer (1992).

5. CONCLUSION AND FUTURE PLANS

In this paper, we examined two methods for weighting the SLID sample in order to achieve the best cross-sectional estimates. From the results obtained, it seems that the Share Approach is the most appropriate to use because of simplicity in the assumptions used, and because of its unbiasedness property.

We also described the use of post-stratification variables to correct for under-representation and to improve the precision of the estimates. Studies are in progress to determine the usefulness of revenue and/or interprovincial mobility as post-stratification variables.

ACKNOWLEDGEMENT

This research has been made possible through financing by the Research Block Fund of Statistics Canada. The authors would like to thank Hussain G. Choudhry and J.N.K. Rao for useful comments.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques, Third Edition*, John Wiley and Sons, New York.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. *Panel Surveys*, John Wiley and Sons, New York.

Gouriéroux, C., and Roy, G. (1978). Enquête en deux vagues: renouvellement de l'échantillon. *Annales de l'INSEE*, 29, 115-135.

Hunter, L., Michaud, S., and Torrance, V. (1992). Modelling non-response in a longitudinal survey. Paper presented at the 1992 Conference of the American Statistical Association.

Lavallée, P. (1992). Sample representativity for the survey of labour and income dynamics. Statistics Canada internal report.

Lemaître, G. (1989). Variance estimation for surveys using the LFS frame weighting system, user documentation.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families, *Survey Methodology*, 13, 2, 199-207.

Little, R.J.A. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 2, 139-157.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey, 1984-1990.* Minister of Supply and Services Canada, Ottawa, Statistics Canada, Catalogue 71-526.

Stukel, D.M., and Boyer, R. (1992). Calibration estimation: An application to the Canadian Labour Force survey. Statistics Canada working paper SSMD-009E.

SESSION 3

Nonresponse and Attrition

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

ONTARIO CHILD HEALTH STUDY FOLLOW-UP EVALUATION OF SAMPLE LOSS PART II

M.H. Boyle, B. Wheaton, D.R. Offord, Y.A. Racine and G. Catlin[1]

ABSTRACT

This article uses information from a 4-year follow-up of children aged 4 to 12 who participated in the Ontario Child Health Study in 1983 to present a multivariable procedure for evaluating the effects of sample loss. The procedure involves (1) statistically modelling sample loss using wave one data; (2) generating predicted probabilities of attrition for everyone in the sample; (3) evaluating the need to construct unit weights to represent lost subjects; and (4) assessing the impact of weights by comparing statistical estimates for the likelihood of poor outcome (psychiatric disorder) based on unweighed and weighted analyses. The results suggest that the effects of sample loss are variable dependent. Of the eight variables included in the analyses, the use of weights had an impact on two reducing their strength of association (relative odds) with poor outcome from 6.87 to 4.35 and from 5.05 to 3.55.

KEY WORDS: Sample loss; Follow-up surveys; Childhood disorder.

1. INTRODUCTION

1.1 Description of the Problem

Longitudinal studies are vulnerable to the adverse effects of sample loss or attrition. Subject losses over time may occur because of death, migration, failure to locate and refusal. Nonparticipation for these reasons invariably accumulates with each data collection period, weakens the precision of statistical estimates and may lead to systematic distortions or bias because of sample unrepresentativeness. The potential for subject losses to invalidate the findings of longitudinal studies focused on children should not be overlooked. For longitudinal studies of childhood psychiatric disorders, it is common for 20 to 30% of subjects to be lost from the first follow-up. This occurred in the Isle of Wight follow-up (Schachar, Rutter and Smith 1981), the midtown Manhattan follow-up (Gersten, Langner, Eisenberg, Simcha-Fagan and McCarthy 1976) and the Ontario Child Health Study follow-up (Offord *et al.* 1992). In studies of adolescent substance use, subject losses at follow-up from 20 to 55% are not unusual (Johnston, O'Malley and Eveland 1978; Donovan, Jessor and Jessor 1983; Kandel and Logan 1984; O'Malley, Bachman and Johnston 1984; Brook, Whiteman, Gordon and Cohen 1986; Newcomb and Bentler 1988).

Evaluating the effects of subject losses on the accuracy of statistical estimates and validity of inferences from longitudinal studies of children has received little empirical attention. Typically, the issue is ignored (*e.g.*, Schachar *et al.* 1981) or addressed by comparing the features of participants and nonparticipants and arguing that no important differences exist in variables examined one at a time (*e.g.*, Gersten *et al.* 1976; Johnson *et al.* 1978; Kandel and Logan 1984; Brook *et al.* 1986) or in various combinations (*e.g.*, Newcomb *et al.* 1986; Macera, Jackson, Farach and Pate 1988). The adequacy of these methods for detecting bias attributable to sample loss is questionable. Drawing from the clinical trial literature, it has been shown that even with randomization, statistically insignificant differences in group characteristics can lead to biased estimates of treatment effects (Altman 1985).

[1] M. Boyle, D. Offord and Y. Racine, Department of Psychiatry, Hamilton, Ontario, Canada. B. Wheaton, Department of Sociology, University of Toronto, Toronto, Ontario, Canada. G. Catlin, Special Surveys Division, Statistics Canada, 24-Q, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada.

1.2 Objectives of the Paper

In 1987, a follow-up study to the original Ontario Child Health Study (OCHS) was done to assess the outcome of selected childhood disorders, to identify variables that predict persistence of disorder (prognosis), and to evaluate risk for disorder over time among children classified originally as free of psychopathology (Offord *et al.* 1992). Subject losses at follow-up in the OCHS provided the motivation to describe the nature of this loss and to evaluate its impact on a variety of statistical estimates to quantify outcome, prognosis and risk (Boyle, Offord, Racine and Catlin 1991).

This paper extends the earlier analyses of sample loss in several important ways. First of all, a broad array of child, parent and family data are used to investigate differences between participants and nonparticipants at follow-up. The earlier analyses focused only on childhood disorder, family income and family dysfunction. Second, multivariable approaches are used to model participation and to develop weighting adjustments to compensate for differences between participants and nonparticipants. The earlier analyses relied on bivariate methods and stratification to develop weighting adjustments. This paper begins with a brief description of the original OCHS and follow-up then focuses on sample loss.

2. RESEARCH METHODS USED IN THE OCHS

2.1 Original OCHS

The methodology of the original OCHS is described in detail elsewhere (Boyle *et al.* 1987). Briefly, the target population included all children born from January 1, 1966, through January 1, 1979, whose usual place of residence was in a household dwelling in Ontario. The sampling unit consisted of all household dwellings listed in the 1981 Census; the sampling frame was the 1981 Census; and the sampling selection was done by stratified, clustered, and random sampling from the Census file of household dwellings. The sampling frame excluded only three groups of children representing 3.3% of the population of children, aged 4 to 16 years: those children living on Indian reserves; those living in collective dwellings, such as institutions; and those residing in dwellings constructed after June 1, 1981 (Census day). Data collection took place between January and March 1983. There was 91.1% participation among eligible households, and only 3.9% refused. Other reasons such as illness and no contact, accounted for the remaining nonparticipants.

2.2 OCHS Follow-up

All children and families who participated in the original OCHS were eligible for the follow-up study; they were located in October and November 1986. The Special Surveys Division of Statistics Canada collected data on follow-up participants in April and May 1987. The methods for measuring psychiatric disorder were the same for 8- to 16-year olds in 1987 as they had been for 4- to 12-year olds in 1983. Very briefly, the measurement of each psychiatric disorder included in the analyses here (conduct disorder, hyperactivity, and emotional disorder) was based on the diagnostic criteria in DSM-III. The Child Behaviour Checklist (CBCL) (Achenbach and Edelbrock 1981) furnished the basic pool of items for measuring the diagnostic criteria, and additional items were generated when the items from the CBCL were felt to describe a particular criterion inadequately. Self-completed problem checklist ratings were obtained from parents and teachers to assess disorder in children under age 12 and from parents and youths themselves to assess disorder in adolescents, aged 12 and over. Ratings of problem behaviours were summed to obtain scale scores. The thresholds for classifying each disorder as present or absent were the scale scores that best discriminated diagnoses made independently by child psychiatrists in a stratified random sample of children (N = 194) participating in the original OCHS. More details on the classification of disorder are available elsewhere (Boyle *et al.* 1987). A new instrument was developed to assess psychiatric disorder in 17- to 20-year olds in 1987. This group will be considered in a separate report.

2.3 Sample Loss at Follow-up in the OCHS

2.3.1 Extent of Sample Loss

In this paper, evaluation of sample loss at follow-up is restricted to children aged 4 to 12 in the original OCHS who were aged 8 to 16 at follow-up. Teacher assessment data are excluded from consideration in this paper, yielding a larger sample for analyses than available in the 1983 (original) and 1987 (follow-up) studies. Ninety-one percent of households agreed to participate in the original OCHS, but missing information on questionnaires in 1983 reduced the sample of individual children for analysis from 2,279 to 1,843. At follow-up, there were 1,402 children with complete data in both 1983 and 1987 and 441 children with complete data in 1983 but missing data in 1987. The analyses in this paper is restricted to sample losses occurring at follow-up.

2.3.2 Characteristics at Follow-up of Participants and Nonparticipants

Participants and nonparticipants in 1987 were compared on three sets of variables assessed in 1983:

(1) measures of childhood psychiatric disorder - the primary focus of the OCHS follow-up;

(2) sociodemographic characteristics of children and families that might distinguish participants from nonparticipants; and

(3) variables shown to be correlates of childhood psychiatric disorder in cross-sectional analyses, which could be examined prospectively to assess their risk potential.

The three sets of variables, accompanied by brief definitions, are grouped in Table 1 under child and parent/family features.

Prior to testing the statistical importance of differences between participants and nonparticipants on the variables defined in Table 1, the patterns of association between ordinal and interval level measures (*i.e.*, sibship size, family dysfunction, family income, number of confidants and negative affect score) and participant status were examined for nonlinear effects. Depending on the observed pattern of association, variables were recoded to maximize their discrimination of participant status.

Table 2 shows the results of comparing participants and nonparticipants on the variables defined in Table 1. The chi-square statistic (X^2) is used to test for statistically significant differences between the groups. Among the twenty comparisons, there are ten variables that distinguished between the groups at a p-value less than .05. The magnitude of the differences exceeds 10.0 percent for three comparisons: aged 9-to 12-years (46.6 *vs.* 60.3) urban residence (59.3 *vs.* 69.6) and a family dysfunction score of 12 to 20 (50.5 *vs.* 37.9). Statistically significant differences between participants and nonparticipants on the other seven variables are no greater in magnitude than 7.5 percent (Table 2). Using statistical significance and magnitude of difference as criteria, there is no evidence that childhood psychiatric disorder distinguishes between participants and nonparticipants (Table 2).

3. METHODS OF EVALUATING SAMPLE LOSS

3.1 Background

The conditions for selective loss have been described elsewhere (Greenland 1977; Criqui 1979; Kleinbaum, Morgenstern and Kupper 1981; Boyle *et al.* 1991). Briefly, selective loss accompanied by bias occurs when subject losses within exposure categories are unevenly distributed among the outcome categories. The direction and magnitude of bias depends on the distribution of subject losses in the cells that define the cross classification of exposure and outcome. To quantify accurately the extent of bias attributable to selective sample loss requires outcome information on nonparticipants obtained in a separate study. Because such studies are difficult and expensive to undertake, outcome data on nonrespondents are rarely available to investigators.

Table 1: Definition of Variables.

Child

Conduct Disorder: Characterized by physical violence against persons or property and/or a severe violation of social norms.

Hyperactivity: Characterized by inattention, impulsivity, and motor activity.

Emotional Disorder: Characterized primarily by feelings of anxiety and depression.

One or more disorders: One or more of conduct disorder, attention-deficit disorder and emotional disorder.

Problems getting along: A child was classified as having problems getting along if the parent checked on a five point scale that the child was getting along "not too well, frequent problems" response option 4 or "not well at all, constant problems" response option 5 in one or more of three different circumstances: with other kids such as friends or classmates, with teachers at school or with the family.

Needs professional help: A child was classified as needing professional help if the parent responded positively to two questions: "During the past six months, do you think that he/she has had any emotional or behavioral problems?" and "Do you think that he/she needs any professional help with these problems?"

School failure ever: Parent reports that the child failed or repeated a grade at some time during his/her school career.

Functional limitations: Parent reports that the child has had a limitation in physical activity, mobility, self-care or role performance for at least six months.

Parent/Family

One parent family: Only one parent in the home.

Urban residence: Residence located in an urban population of at least 1,000 people with a population density of 400 persons or more per square kilometre (Statistics Canada 1982).

Large sibship: Four or more siblings aged 4 to 16 in the family. For comparing participants and nonparticipants, the number of siblings was recoded to two levels: (1) 3 or more siblings and (2) 0, 1 or 2 siblings.

Over-crowded home: A ratio of individuals to rooms ≥ 1.0.

Mobile family: Change in residence two or more times during the past two years. For comparing participants and nonparticipants, mobility was recoded into two levels: (1) 3 or more moves; and (2) 0, 1 or 2 moves.

Family dysfunction: A parent reported score of 27 to 48 (range 12 to 48) on the 12-item general functioning scale derived from the McMaster Family Functioning Assessment Device (Byles, Byrne, Boyle and Offord 1988). Family functioning is assessed on six dimensions: problem solving, communication, roles, affective responsiveness, affective involvement, and behaviour control. For comparing participants and nonparticipants, family dysfunction was recoded into three levels: (1) scores from 12 to 20; (2) scores from 21 to 25; and (3) scores over 25.

Parent(s) treated for nerves: Parent reports that self or partner was treated at some time for nerves or a nervous condition.

Family income < $10,000: Total family income in the year preceding the survey (1982) was < $10,000. For comparing participants and nonparticipants, family income was recoded into three levels: (1) less than $10,000, (2) $10,000 to $39,999 and (3) $40,000 or more.

No confidants: Parent reports having no one to confide in about personal problems or difficulties. For comparing participants and nonparticipants, number of confidants was recoded into two levels: (1) 1 to 5 confidants and (2) 0, 6 or 7 confidants.

Negative affect score: A parent reported score of 5 or more (range 0 to 10) on the 5 item negative affect scale developed by Bradburn.

Table 2: Percent Distribution of 1983 Characteristics by Follow-up Status in 1987.

	Follow-up Status 1987			
	Participants	Nonparticipants		
1983 Characteristics	N = 1,402	N = 441	X^2(df)	p-value
Child				
Male	51.1	50.6	0.02(1)	NS
Aged 9 to 12 years	46.6	60.3	24.79(1)	0.00
One or more disorders	6.9	5.7	0.66(1)	NS
Conduct disorder	0.9	1.4	0.27(1)	NS
Hyperactivity	1.9	2.0	0.00(1)	NS
Emotional disorder	5.5	4.8	0.23(1)	NS
Problems getting along	2.4	5.0	6.64(1)	0.01
Needs professional help	2.4	4.1	2.78(1)	NS
School failure ever	7.3	10.7	4.49(1)	0.04
Functional limitations	4.0	5.9	2.42(1)	NS
Parent/Family				
One parent family	8.9	8.6	0.01(1)	NS
Urban residence	59.3	69.6	14.76(1)	0.00
3 or more siblings	29.3	35.8	6.37(1)	0.02
Over-crowded home	13.7	17.7	3.96(1)	0.05
3 or more residential moves	1.7	3.6	4.93(1)	0.03
Family dysfunction				
(1) Score of 12 to 20	50.5	37.9	26.65(2)	0.00
(2) Score of 21 to 25	37.2	42.4		
(3) Score > 25	12.3	19.7		
Parent(s) treated for nerves	21.5	19.3	0.91(1)	NS
Family income				
(1) <$10,000	5.7	9.3	10.70(2)	0.01
(2) $10,000-$39,999	66.3	68.3		
(3) $40,000 or more	28.0	22.4		
1 to 5 confidants	11.8	18.8	13.40(1)	0.00
Negative affect score	13.1	17.9	6.11	0.02

Although outcome data are not available on nonparticipants in follow-up studies, all information collected in previous assessments can be used to distinguish the special features of nonparticipants that could lead to sample unrepresentativeness. This evaluation, done one variable at a time, was shown in Table 2 and discussed in the previous section. Given that a number of the variables in Table 2 that distinguish participants from nonparticipants at follow-up may also be risk factors for childhood disorder, then grounds exist for suspecting that selective sample loss and bias have occurred.

3.2 Weighting Adjustments

One method recommended for testing whether or not selective sample loss and bias have occurred is to develop weighting adjustments intended to compensate for disproportionately high sample losses among respondents thought to be at greater risk for poor outcomes. The development of weighting adjustments based on bivariate analyses was presented in an earlier paper (Boyle *et al.* 1991). The method used here is based on multivariable procedures and borrows from the work of Aneshensel, Becerra, Fielder and Schuler (1989). The following steps are involved:

(1) the estimation of a logistic regression equation based on 1983 assessment data to predict sample attrition in 1987 as a binary variable;

(2) the use of the equation to generate predicted probabilities of attrition for everyone in the follow-up sample;

(3) an evaluation of the need to stratify respondents according to the probability of follow-up loss and to construct weights that reflect the size of sample loss in each stratum; and

(4) the construction of weights for respondents defined as the ratio of the number of respondents in 1983 over the number of respondents in 1987 for each stratum defined by the probability of follow-up loss in 1987.

3.2.1 Steps 1 and 2

Forward, stepwise logistic regression analyses derived from SPSS software was used to build an equation to predict sample attrition in 1987 from 1983 assessment data. Candidate variables were those shown in Table 2 having an X^2 value greater than 1.50. In addition to the variables selected for evaluation, three interactions were specified: family dysfunction by income, overcrowded home by 3 or more residential moves and 1 to 5 confidants by negative affect score. The statistical criteria used in building the model included a probability for variable entry set at 0.10 and a probability for variable removal set at 0.15. Liberal statistical criteria were invoked to maximize the predictive accuracy of the model. The final model included the following main effects: describing the child: aged 9-to 12-years and problems getting along; describing the parent/family: urban residence, 3 or more siblings, 3 or more residential moves and family dysfunction; and two interactions: family dysfunction by income, and 1 to 5 confidants by negative affect score. Using the regression coefficients and observed values of the 1983 characteristics, the probability of attrition can be estimated for every respondent in 1983.

3.2.2 Step 3

Step 3 involves an evaluation of the need to stratify respondents according to their probability of follow-up loss and to construct weights that reflect the size of sample loss in each stratum. The pertinent evidence in this evaluation includes:

(1) the extent to which the probability of follow-up loss differentiates between participants and nonparticipants in 1987; and

(2) the extent to which estimates of risk for poor outcome in 1987 among participants varies according to probability of attrition.

Both of these conditions must exist to some extent for any weighting system based on probabilities of sample loss to have an impact.

To examine the extent to which the probability of follow-up loss differentiates between participants and nonparticipants in 1987, a t-test was done on the probabilities of attrition estimated from the logistic regression in Step 2. To determine the extent to which estimates of risk for poor outcome in 1987 are modified by probability of attrition, follow-up respondents were divided into two strata based on their probability of attrition: high probability defined as $> 25\%$ chance of attrition, and low probability defined as $\leq 25\%$ chance of attrition. Next, we computed strata specific estimates for the strength of association (relative odds) between potential risk factors assessed in 1983 and one or more psychiatric disorders assessed in 1987. Whether or not the strata specific estimates were significantly different from one another was assessed using the test for homogeneity of the X^2 with one degree of freedom (see Schesselman 1982).

As expected, nonparticipants in 1987 had a higher probability of attrition (M = 0.281) than did participants (M = 0.226), t(1,841) = 10.37,p <.000. In addition, evidence exists that risk for poor outcome is modified by the probability of attrition in some instances.

Table 3: Relative Odds Between Potential Risk Factors Assessed in 1983 and One or More Disorders Assessed in 1987 by Probability of Attrition in 1987.

1983 Potential Risk Factors	Probability of attrition 25% or less N = 963	 >25% N = 439	Homogenity X^2(1df)	p-value
Child				
One or more disorders	9.03***	8.30***	0.03	NS
Problems getting along	19.57***	3.12	5.00	.05
Needs professional help	4.80*	4.75**	0.00	NS
School failure ever	1.22	1.47	0.07	NS
Functional limitations	6.02***	0.72	5.80	.02
Parent/Family				
Family dysfunction	2.55	2.67*	0.01	NS
Parent(s) treated for nerves	1.73	1.55	0.05	NS
Negative affect score	3.20*	1.48	2.31	NS

Table 3 shows the strata specific relative odds, between potential risk factors assessed in 1983 and one or more disorders assessed in 1987. For two of the variables - problems getting along and functional limitation - the relative odds for poor outcome vary depending on the probability of attrition. For example, problems getting along assessed in 1983 is a predictor of one or more disorders assessed in 1987. Among those with a low probability of attrition the relative odds is 19.57; among those with high probability of attrition, the relative odds is 3.12. In a similar manner, functional limitation assessed in 1983 is a strong predictor of one or more disorders in 1983 among those with a low probability of attrition (relative odds 6.02) but not among those with a high probability of attrition (relative odds 0.72).

3.2.3 Step 4

Step 4 involves the construction of weights for respondents at follow-up. To generate these weights, all respondents and nonrespondents at follow-up are stratified according to their probability of attrition - an estimate generated earlier by the logistic regression model. Respondents and nonrespondents within each strata are summed and this sum is divided by the number of respondents to give a unit weight within each strata. This unit weight is then divided by a constant so that the sum of the weights equals the sample size at follow-up: 1,402.

Table 4 shows what happens to the strength of association between the potential risk factors assessed in 1983 and one or more disorders assessed in 1987 when the weights are applied to the follow-up sample. Differences between the unweighted and weighted estimates for relative odds go from -0.19 (school failure ever, negative affect score) to 2.52 (problems getting along). According to the data in Table 4, the predominant direction of the bias attributable to sample loss is away from the null and quite substantial for two of the variables examined (*i.e.*, problems getting along and needs professional help).

Table 4: Unweighted *vs.* Weighted Relative Odds Between 1983 Potential Risk Factors and One or More Disorders in 1987.

	Relative Odds		▲
1983 Potential Risk Factors	Unweighted (1)	Weighted (2)	(1)-(2)
Child			
One or more disorders	8.96***	8.35***	0.61
Problems getting along	6.87***	4.35***	2.52
Needs professional help	5.05***	3.55**	1.50
School failure ever	1.44	1.63	-0.19
Functional limitations	2.64*	2.49*	0.15
Parent/Family			
Family dysfunction	2.82*	2.72***	0.10
Parent(s) treated for nerves	1.71*	1.65*	0.06
Negative affect score	2.27**	2.46***	-0.19

* p < .05, **p < .01, ***p < .001

▲ unweighted minus weighted relative odds

4. DISCUSSION

Although longitudinal research has the potential to extend our knowledge about childhood psychiatric disorder, methodological considerations, such as sample loss, are important determinants of data usefulness. This paper has focused on a multi-variable procedure for evaluating the effects of sample loss at follow-up to determine whether or not selective loss and bias has occurred. The procedure involves (1) statistically modelling sample loss using existing data; (2) generating predicted probabilities of attrition for everyone in the sample; (3) evaluating the need to construct unit weights; and (4) assessing the impact of weights (if needed) by comparing unweighted versus weighted analyses. In the present study, the evidence suggested that the construction of weights was justified. Assessment of the impact of the weights indicated that it was not uniform. For two of the child variables studied - problems getting along and needs professional help - the magnitude of the differences in relative odds were 2.52 and 1.50, respectively. The direction of the bias was away from the null. For the other variables studied, the use of weighting adjustments had only a small impact on the statistical estimates.

The procedure described in this paper for evaluating sample loss in follow-up studies is relatively simple to use yet provides a far more complete analysis of sample loss than simple comparisons of respondents and nonrespondents on relevant variables. It is important to note, however, that the procedure works best when the statistical model predicting sample loss is well specified. If respondents and nonrespondents are indistinguishable in all of the relevant variables then no basis will exist for predicting their probability of attrition. Furthermore, it is assumed that respondents and nonrespondents with the same probability of attrition will have the same follow-up experiences.

As we look increasingly to longitudinal research for answers about the nature of childhood psychopathology, methods for evaluating the effects of and appropriate adjustments for sample loss at follow-up should take on added importance. Weighting as a procedure for evaluating the effects of sample loss offers some promise as a means for determining whether or not sample loss has introduced bias into an analyses. Further work must be done to identify the conditions and circumstances in which the use of weighting adjustments effectively offsets the distortions introduced by selective sample loss.

ACKNOWLEDGEMENTS

This work was supported by funds from the National Health Research and Development Program (Grant Number 6606-3760-42) and the Ontario Ministry of Community and Social Services, and was carried out by the Child Epidemiology Unit, Department of Psychiatry, McMaster University and the Child and Family Centre, Chedoke-McMaster Hospitals, Hamilton, Ontario. Dr. Boyle is supported by a William T. Grant Foundation Faculty Scholar Award and Dr. Offord by a National Health Scientist Award, Health and Welfare Canada.

REFERENCES

Achenbach, T., and Edelbrock, C. (1981). Behavioral problems and competences reported by parents of normal and disturbed children aged 4 through 16. *Monographs of the Society for Research in Child Development*, 46, 188.

Altman, D.G. (1985). Comparability of randomized groups. *The Statistician*, 34, 125-136.

Aneshensel, C.S., Becerra, R.M., Fielder, E.P., and Schuler, R.H. (1989). Participation of Mexican American female adolescents in a longitudinal panel survey. *Public Opinion Quarterly*, 53, 548-562.

Boyle, M.H., Offord, D.R., Hofmann, H.F., Catlin, G.P., Byles, J.A., Cadman, D.T., Crawford, J.W., Links, P.S., Rae-Grant, N.I., and Szatmari, P. (1987). Ontario Child Health Study: I. Methodology. *Archives of General Psychiatry*, 44, 826-831.

Boyle, M.H., Offord, D.R., Racine, Y.A., and Catlin, G.P. (1991). Ontario Child Health Study Follow-up: evaluation of sample loss. *Journal of the American Academy of Child and Adolescent Psychiatry*, 30, 449-456.

Brook, J.S., Whiteman, M., Gordon, A.S., and Cohen, P. (1986). Dynamics of childhood and adolescent personality traits and adolescent drug use. *Developmental Psychology*, 22, 403-414.

Byles, J.A., Byrne, C., Boyle, M.H., and Offord D.R. (1988). Ontario Child Health Study: Reliability and validity of the general functioning subscale of the McMaster Family Assessment Device. *Family Process*, 27, 97-104.

Criqui, M.H. (1979). Response bias and risk ratios in epidemiologic studies. *American Journal of Epidemiology*, 109, 344-399.

Donovan, J.E., Jessor, R., and Jessor, L. (1983). Problem drinking in adolescence and young adulthood: A follow-up study. *Journal of Studies of Alcohol*, 44, 109-137.

Gersten, J.C., Langner, T.S., Eisenberg, J.G., Simcha-Fagan, O., and McCarthy, E.D. (1976). Stability and change in types of behavioral disturbance of children and adolescents. *Journal of Abnormal Child Psychology*, 4, 111-127.

Greenland, S. (1977). Response and follow-up bias in cohort studies. *American Journal of Epidemiology*, 106, 183-187.

Johnston, L.D., O'Malley, P.M., and Eveland, L.K. (1978). Drugs and delinquency: A search for causal connections. In D.B. Kandel (Ed.). *Longitudinal research on drug use: empirical findings and methodology issues*. Washington, DC: Hemisphere-Wiley, 132-156.

Kandel, D.B., and Logan, J.A. (1984). Patterns of drug use from adolescence to young adulthood, I: Periods of risk for initiation, continued use and discontinuation. *American Journal of Public Health*, 74, 660-666.

Kleinbaum, D.G., Morgenstern, H., and Kupper, L.L. (1981). Selection bias in epidemiologic studies. *American Journal of Epidemiology*, 113, 452-463.

Macera, C.A., Jackson, K.L., Farach, C., and Pate, R.R. (1988). The use of proportional hazards regression in investigating dropout rates in a longitudinal study. *Journal of Clinical Epidemiology*, 41, 1175-1180.

Newcomb, M.D., and Bentler, P.M. (1988). Impact of adolescent drug use and social support on problems of young adults: A longitudinal study. *Journal of Abnormal Psychology*, 97, 64-75.

Newcomb, M.D., Maddahian, E., and Bentler, P.M. (1986). Risk factors for drug use among adolescents: concurrent and longitudinal analyses. *American Journal of Public Health*, 76, 525-531.

Offord, D.R., Boyle, M.H., Racine, Y.A., Fleming, J.A., Cadman, D.T., Munroe Blum, H., Byrne, C., Links, P.S., Lipman, E.L., MacMillan, H.C., Rae-Grant, N.I., Sanford, M.N., Szatmari, P., Thomas, H., and Wordward, C.A. (1992). Outcome, prognosis and risk in a longitudinal follow-up study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 31, 916-923.

O'Malley, P.M., Bachman, J.G., and Johnston, L.D. (1984). Period, age and cohort effects on substance use among American youth, 1976-82. *American Journal of Public Health*, 74, 682-688.

Schachar, R., Rutter, M., and Smith, A. (1981). The characteristics of situationally and pervasively hyperactive children: Implications for syndrome definition. *Journal of Child Psychology and Psychiatry*, 22, 375-392.

Schlesselman, J.J. (1982). *Case-Control Studies: Design, Conduct, Snalysis*. Oxford: Oxford University Press.

Statistics Canada (1982). 1981 Census Dictionary (Cat. No. 99-901). Ottawa. Ministry of Supply and Services.

Proceeding of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

STRATEGY FOR MINIMIZING THE IMPACT OF NON-RESPONSE FOR THE SURVEY OF LABOUR AND INCOME DYNAMICS

S. Michaud and L. Hunter[1]

SUMMARY

Statistics Canada will launch a major panel survey of households in 1994. The Survey of Labour and Income Dynamics (SLID) will interview people for six years,conducting two interviews per year. It is felt that non-response and attrition of the sample are key issues to examine in the survey design. This paper will present approaches that are evaluated to compensate for non-response in SLID; a modelling approach to compensate for the non-response through a weighting adjustment, and imputation of the missing waves of data.

KEY WORDS: Non-response; Longitudinal survey; Modelling; Imputation.

1. INTRODUCTION

Statistics Canada will launch a major panel survey of households in 1994 called the Survey of Labour and Income Dynamics (SLID). The survey will follow individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. SLID has a strong base within Statistics Canada. Its origins are in several surveys, including the Labour Force Survey (LFS), the Survey of Consumer Finances (SCF) and the Labour Market Activity Survey (LMAS). Both LFS and SCF are cross-sectional surveys. As cross-sectional surveys, they offer a series of snapshots and are useful and efficient tools for monitoring trends at aggregate levels. The LMAS served both as a longitudinal and as a cross-sectional survey. Two panels have been conducted to date, a two year panel (1986-1987) and a three year panel (1988-1990). For each longitudinal panel, people that participated in the first wave were interviewed and traced. All people living with them in the following waves were also interviewed (but not traced).

Because SLID wants to interview people for six years, conducting two interviews per year, it is felt that non-response rates and attrition of the sample are key issues to examine in the survey design. Studies are currently being conducted on different aspects of non-response to the LMAS in hopes of finding approaches that will minimize the impact of non-response on the SLID data. This paper will discuss some of the studies that are underway to try to minimize the impact of non-response; the possibility of fitting a model or models to (a) adjust for non-response in weighting and (b) imputation procedures that are tested to try to compensate for a missing wave of information. Section 2 of the paper will first give more details on the SLID survey design, section 3 will concentrate on the modelling that is being tested to adjust for the weighting of the longitudinal files, section 4 will present the imputation strategy and section 5 will outline other plans to help to handle non-response and future work.

2. SLID SURVEY DESIGN

A number of longitudinal surveys have been carried in the United States and in other countries. Lepkwoski mentioned factors that could influence the strategy to handle non-response in a longitudinal context: the form of the analysis (whether the data is collected to perform mainly longitudinal comparisons, cross-sectional

[1] S. Michaud and L. Hunter, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

estimates or to look at longitudinal accumulation of data); the type of data collected (continuous, categorical or conditional); and the wave non-response pattern (attritors *vs* non-attritors). SLID has some extra features that will influence the design of the strategy: data for SLID will be collected in deffered interviews; SLID has two units of analysis, data has to be derived at the individual level as well as at the household level; and finally some of the collected information overlaps slightly at the seam.

2.1 SLID data collection and the deffered interviews

As mentioned earlier, SLID is a longitudinal survey. A sample of households will be selected in January 1993 from the Labour Force Survey sample. Starting in January 1994, the people selected from the 1993 Labour Force sample will be interviewed twice a year for six years. There will be an interview in January, to collect information on activity for the previous year (also called the reference year). The interview will get data such as start and end dates of jobs, detailed information on up to three employers, information on the non-working spells and the absences from work, as well as some questions on disability. The May interview will collect income and some wealth information, for the people who were in the January sample. The income information will be collected for the same reference period as the labour information (that is the previous calendar year). It is felt that income can be reported more easily around May when people fill in their income tax return. The May interview can be seen as a deferred interview from the January sample, since the information is collected on the same people, for the same reference period. A file will be released each year, with the labour and the income information combined. The principle of deferred interview, with each interview collecting different information, adds another dimension to the definition of non-response. Usually complete non-response is defined when a person does not respond to an interview. With the deffered interview, even if one interview is missing, one interview represents only a portion of the data that is released annually. So a missing interview could be viewed as partial non-response.

2.2 SLID units of analysis

SLID will interview persons, and a lot of information is collected and will be analyzed at the person level. However, measures of poverty and certain income studies require information at the household level. Since households are not a stable unit in time, it was decided that SLID would follow individuals through time (and not try to follow households). Derived households and economic family characteristics will be calculated each January and they will be attached to individuals as characteristics of the people (people live in a family with three persons, with a household income of "x"). This implies that non-response may be treated differently, if at least one person responded in the household as opposed to if the whole household did not respond.

2.3 SLID overlapping collection periods

For the labour component of SLID, activity is recorded from January first of the reference year, up to the interview date (which may be as late as February of the next year). This means there will be overlap of approximately one month between consecutive waves of labour interviews. This may give more tools for imputation.

2.4 Non-response

In the context of a longitudinal sample, Kalton divided non-response into two categories: the non-response due to attrition *vs* non-response due to non-attrition. Attrition occurs when a person who responded to one or more early waves, does not respond to any subsequent waves. For example, in a three year panel, people not responding to both the second and the third waves would be called attritors. Non-attrition happens when people who did not respond to one or more waves get to respond again.

Based on the Labour Market Activity three year panel, 16% of non-response was due to attrition and 4% were non-attritors. Experience with other surveys shows that the ratio of attritors to non-attritors decreases over time (a longer panel has a higher ratio of non-attritors over attritors compared to a short panel).

When the LMAS file was evaluated, non-response was quite different among certain groups:

- movers had a non-response rate (including non traceable) of close to 20% while non-response for non-movers was about 2%. The variable move/non-move was the characteristic that presented by far the most differences,
- based on characteristics from the wave 1, more non-response was found among the group that were unemployed in wave 1,
- based again on wave 1 information, non-response was higher for the group of people who were not married in wave 1,
- people living in urban area in wave 1 also had a higher non-response rate after three years,
- non-attritors seem to be different than respondents in the kind of jobs they hold.

The differences in characteristics between respondents and non-respondents suggested a couple of possibilities. First, the current non-response adjustment which uses sample design information only (basically some geographical information) may not be the best way to compensate for non-response. A technique which could take into account other characteristics based on previous years of data may provide better results. Secondly, a lot more information is available when non-attrition non-response occurs. Because there seems to be a lot of differences in the kind of jobs that are hold by respondents and non-respondents, imputation may be a better tool to adjust for non-response, especially when it is due to non-attrition. When data is collected for a long period, there may be more gains to impute for non-attrition by opposition to dropping the record and reweighting it to compensate for the non-response, especially if the number of missing waves is small. However, it has been pointed by other surveys that longitudinal imputation can be an difficult, extensive and expensive process. To be able to determine the best strategy to deal with non-response in SLID, some evaluations studies have started.

3. MODELLING

Two possible approaches were considered for the non-response adjustments in weighting: ratio adjustments within population subgroups, and regression model-fitting. The model-fitting approach was chosen because it was felt that this work could be used to serve other purposes as well. In particular, there are two possible uses of a non-response model in the context of our longitudinal survey: the prediction of non-response and a non-response weighting adjustment. While the same model could probably not be used for the two purposes, it was hoped that a base set of variables could be identified which would be common to the models. A small set of additional variables would be unique to the different purposes of the models. For example, the characteristic most correlated with non-response is whether or not the person moved since the time of the last interview (non-response being the result of the inability to trace in this case); clearly this information could be used in the model for weighting, but would not be available at the time of the previous interview (since the event had not yet taken place). At best, we could hope to find a variable or set of variables that is correlated with subsequent moves to use in the prediction model.

3.1 The Model

Logistic regression was used to create the model. This type of a model was chosen because non-response is a binary dependent variable. Logistic regression was preferred over discriminant analysis since logistic regression has fewer assumptions and is essentially as efficient as discriminant analysis (Harrell 1983).

The multiple logistic response function is

$$E\{Y \mid X\} = [1 + \exp(-\beta^T X)]^{-1}, \tag{1}$$

where Y is the dependent variable,
β is the column vector of regression parameters,
X is the $n \times (p\text{-}1)$ matrix of independent variables.

Equation (1) expands to

$$E\{Y \mid X\} = [1 + \exp(-\beta_0 - \beta_1 X_1 - \cdots - \beta_{p-1} X_{p-1})]^{-1}. \tag{2}$$

The dependent variable, Y_i, in this analysis indicated if the i^{th} respondent to the 1986 survey had become a non-respondent to the 1987 survey. Therefore, for the i^{th} individual

Y_i = 1 if the i^{th} individual did not respond in 1987,
Y_i = 0 if the i^{th} individual did respond in 1987.

The multiple logistic regression model states that Y_i are independent Bernoulli random variables with

$$E\{Y_i \mid X_i\} = [1 + \exp(-\beta^T X_i)]^{-1} \tag{3}$$

and X_i is the vector of p-1 independent variables associated with the i^{th} individual.

Denoting $P(Y=1 \mid X)$ as $\pi(X)$, the logit transformation is defined as

$$\begin{aligned} g(X) &= \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \quad . \end{aligned}$$

3.2 The Data

The 1986/87 panel of LMAS was used to fit and evaluate the non-response models. The dataset consisted of 66,817 individuals, of which 3,385 (5%) were non-respondents to the 1987 interview. Demographic variables that were likely to be related to non-response were chosen from the 1986 LMAS master file as possible independent variables for the model. One additional variable was collected in 1987 for all individuals - whether or not the person changed address since the 1986 interview.

3.3 Variables

The variables examined for inclusion in the non-response model were:

Geographical information: Province and Urban/Rural area at 1986 interview;

Household/dwelling information: Household size, Type of dwelling (house; other), Status of dwelling (owned; rented) at 1986 interview;

Demographic information: Sex, Age, Marital status, School attendance (full time; part time; none), Highest level of education at 1986 interview;

Labour characteristics: Any employment, Any unemployment, Any out-of-labour-force, Number of jobs, Any short tenure jobs (< 2 years), Any long tenure jobs (2 years or more), Any absences from work, Industry of job(s), Average weekly income (over all jobs), Received any unemployment insurance, Received any welfare in 1986;

Move status: Moved (changed address between 1986 interview and 1987 interview).

Categorical variables were analyzed and manipulated so that the final representation of this information was by groups of binomial (0-1) variables. The differences between respondents and non-respondents with respect to the independent variables were analyzed. The correlations between all pairs of these variables were examined to find any potential multicollinearity.

3.4 The Sample File

PROC LOGIST in SAS was used to fit the logistic regression model. Because of the size of the dataset, the procedure required a large amount of computer resources. Therefore, it was decided to select a sample of households from the original file to be used for the model-building stage. The sample file consisted of all households that contained a non-respondent plus a random selection of an equal number of households containing respondents only. This was preferable to a simple random sample since the variables associated with non-response could be more easily identified by using all the non-response information that was available. The parameters of the regression model were estimated using the full dataset.

3.5 Regression Procedures

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. This reduction in the choice of variables resulted in fewer variables to be entered into the logistic procedures saving considerable computer resources.

The variables given in the STEPWISE procedure were entered into the SAS procedure PROC LOGIST with the BACKWARD and FAST options. These options allowed LOGIST to use an approximate backward elimination method to eliminate nonsignificant variables. Different logistic regression models were fitted to the full dataset using combinations of the most significant variables identified from the sample file. A consideration in choosing the model was the number of variables. It was desired to have a model with a small number of variables so that utilizing the model would be simple.

The use for a non-response model in a longitudinal survey is to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total non-response, and the independent variables were characteristics observed the previous year (1986) plus the current year's information (1987) on whether or not the person moved.

The BACKWARD option of PROC LOGIST was used with the sample file to identify eight variables related to non-response.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)
Any employment	(ANYEMP)
Highest education=secondary	(EDUCSEC)
Moved since 1986 interview	(MOVED)
Household size	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity in the logit. Plots of the variables showed that neither appeared linear. Non-response was high for ages 16-24, low for ages 25-54, and rose slightly for ages 55+. Some transformations were attempted, but without success. It was decided instead to create age groups, and replace the continuous age variable with two binomial variables for age (AGE1, AGE2). Because few people in the sample came from very large households, it was decided to group together households of size 8 or more and assign a value of 8 to the recoded variable. A plot of non-response versus the recoded household size showed essentially a V-shaped distribution. The transformation ABS(HHS - 4.5) was used to linearize the variable. The transformed household size variable was called HHSTRANS.

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except RENT; (3) using all except EDUCSEC; (4) using all except EDUCSEC and AGE. Although all eight variables were significant using the sample file, when the models were fitted to the full data file, certain ones no longer appeared important. However, it was decided to retain them in the models anyway. The statistics for evaluating the fit of the models indicated few differences between the four models. The Pearson residuals were plotted against the fitted values and the residual plots were examined. Model (3) residuals indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. A summary of the fitted values for this model is given below. Note that the age and single variables as well as their interactions are not statistically significant. Nevertheless, when a model was fitted with these variables removed, it was found that there were more extreme values in the residuals.

Table 1: Parameter estimates for weighting final model.

Variable	$\hat{\beta}$	s.e.	χ^2
INTERCEPT	-3.81	0.14	702.59
HHSTRANS	0.13	0.06	4.97
MALE	0.25	0.04	41.98
RENT	.23	0.04	29.14
SINGLE	0.11	0.16	0.43
MOVED	2.31	0.04	3,065.95
AGE1	-0.15	0.17	0.75
AGE2	-0.19	0.15	1.65
AGE1*HHSTRANS	0.02	0.07	0.07
AGE2*HHSTRANS	0.05	0.06	0.55
AGE1*SINGLE	0.13	0.18	0.52
AGE2*SINGLE	0.11	0.17	0.40

Using the parameter estimates from the final model, predicted probabilities of non-response were calculated for all respondents to the 1987 interview. The non-response weighting adjustment was done by dividing the 1986 starting weight by (1-predicted probability). This gave a non-response adjusted 1987 weight. A post-stratification was then done to adjust the weights to population control totals. This was done by ratio adjusting the weights within categories of province-sex-age group, to produce a 1987 final weight.

3.6 Evaluation of the Weights

Because the LMAS is a longitudinal survey, the same people are present in the sample in both years. The only difference between the people on the 1986 file and the people on the 1987 file is that some are missing from the 1987 file because of non-response. If the non-response weighting adjustment is adequate, there should be no difference in estimates obtained from the 1986 respondents and estimates obtained from the 1987 respondents when tabulating on 1986 characteristics. A number of demographic and labour-related characteristics were evaluated. Estimates were calculated using the 1986 weights, the 1987 model-adjusted weight, and the 1987 regular weights (doing a ratio-adjustment at low geographic levels for non-response adjustment). For each characteristic a 95% confidence interval was calculated for the estimate based on the 1986 weights. The two 1987 estimates were compared for differences to the 1986 estimates as well as differences to each other. The table below shows some of the results.

Table 2: Comparison of estimates based on final weights, tabulated on 1986 characteristics.

	1986 estimate	95% c.i. for 1986 estimate	1987 model-based estimate	1987 regular estimate
Marital Status				
Married	64.6%	(64.1,65.1)	64.8%	65.1%
Single	26.7%	(26.3,27.0)	26.6%	26.4%
Widowed	3.1%	(2.9,3.3)	3.0%	3.0%
Divorced	5.7%	(5.4,6.0)	5.5%	5.4%
Highest Education				
Grade 0-8	14.7%	(14.2,15.2)	14.7%	14.6%
Secondary	50.3%	(49.7,50.9)	50.0%	50.1%
Some Post-Secondary	10.1%	(9.8,10.4)	10.2%	10.2%
Post-Sec. Cert./Dip.	12.9%	(12.5,13.3)	13.0%	13.0%
University Degree	12.0%	(11.6,12.4)	12.1%	12.1%
Labour Force Participation				
Any Employment	77.2%	(76.8,77.6)	77.3%	77.6%
Any Unemployment	17.3%	(16.9,17.7)	17.1%	16.9%
Any Out-of-labour-force	40.5%	(40.0,41.0)	40.3%	40.1%
Weeks Employed in 1986				
0 weeks	22.8%	(22.4,23.2)	22.7%	22.4%
1-26 weeks	12.0%	(11.7,12.3)	11.8%	11.8%
27-48 weeks	12.2%	(11.9,12.5)	12.1%	12.1%
49-52 weeks	53.0%	(52.4,53.6)	53.4%	53.7%

Of all the characteristics compared, only one 1987 estimate was outside the 1986 confidence interval: weeks employed=49-52 using the regular weighting. One pattern was clear, however. The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting. The two 1987 estimates were also compared using the sub-weights (before doing the post-stratification adjustment) and at provincial as well as national levels. In general, differences between the 1987 estimates were greater using sub-weights than they were using final weights. Differences were also greater for labour-related characteristics than for demographic characteristics; differences were greater for variables included in the non-response model; differences were greater at provincial level than at national level. Although the size of the differences are small, the indications are that the model-based approach is performing better. It is expected that when the non-response is extended over more years, the gains will be greater.

4. IMPUTATION

Imputation is often the alternative to weighting adjustments, to compensate for non-response. Even though SLID is first of all a longitudinal survey, cross-sectional estimates will also be produced. Those two different needs impact on the imputation strategy.

With the current plans, a cross-sectional file will be produced annually. Cross-sectional files will not be linked, so there is no requirement to ensure longitudinal consistency with the previous year. An imputation procedure just needs to ensure internal consistency. The cross-sectional file will mainly be composed of longitudinal respondents, but it may also include new people that joined the household. Whether a non-respondent is a joiner or a longitudinal person will impact on the information available to do the imputation.

A longitudinal file will be released at the end of the cycle (except may be for the first panel where SLID will likely release yearly longitudinal files until a full panel has been completed). The longitudinal file will produce a record that linked information on the longitudinal respondents for the whole panel. A lot of cleaning up was done for respondents to the LMAS, to ensure some consistency of the longitudinal information. The requirement for consistency impact on the imputation procedures. While the longitudinal information may give a more robust imputation, it also makes it a lot more difficult. Since a longitudinal survey often focus on measures of change, the imputation procedure should be done in a way that minimizes artificial changes.

4.1 Analysis of longitudinal non-respondents

Some testing has been done using the Labour Market Activity longitudinal files. Both the two years and the three year panels served to evaluate the feasibility of doing imputation with a longitudinal survey. A summary of the results can be presented as follow:

- most of the non-respondents were from a complete household non-response in the first year. The second year also had a relatively high non-response rate, mainly due to people that could not be traced,

- the variables that overlapped at the seam (end dates of some of the jobs), gave powerful predictors for the overall activity during the year,

- comparisons between the respondents to the three year cycle and non-attritors (people who responded to the first year and the third year, but not to the second year) showed differences on the type of jobs that were held by the two groups. Table 3 shows an example of the results obtained from the comparison of the type of jobs held in the first and the third year of the cycle. A longitudinal job is a job that was held both in year 1 and in year 3 (with the same employer and the same occupation),

- among respondents, for the longitudinal jobs, the job characteristics (such as unionisation, class of worker...) had a very high correlation between two years. Variables such as wages and salary or the hours worked are subject to more response error, and so they were not as correlated between two years as what was expected. Correlation between the presence of work interruption(s), lay-offs, absences, were much lower during two consecutive years,

- comparisons between respondents to the three year cycle and non-attritors on non-labour questions (such as disability and health questions) indicated that those characteristics seemed fairly stable.

Table 3: Comparisons of the general activity, for respondents to the three year cycle on LMAS *vs* the non-attritors.

	Respondents	Non-attritors
No jobs in year 1 and year 3	18%	16%
Jobs in year 1 but not in year 3	6%	12%
Jobs in year 3 but not in year 1	4%	4%
Jobs in year 1 and year 3, but no longitudinal job	27%	49%
Jobs in year 1 and in year 3, at least one longitudinal job	45%	19%

4.2 Imputation strategy

Based on those results, the imputation strategy tested on LMAS was as follows:

- exclude from the imputation the records that were complete non-respondents in LMAS. This decreased by more than 50% the amount of imputation that had to be done,

- divide the records that require imputation in two groups: the records with no longitudinal information, and the records with longitudinal information. For the records with no longitudinal information, do a hot-deck imputation for all the labour information. The imputation classes were formed using the person's demographic information plus some basic household information,

- the imputation for the records with longitudinal information was more complex; a balance had to be made between internal consistency and longitudinal consistency. A compromise strategy was tested; variables that required imputation were divided in three categories: (1) person's characteristics (*e.g.*, disability and health), (2) labour information that required longitudinal consistency and (3) other labour information,

- the person's characteristics (1) were imputed based on the responses from the previous year (or from both year for the non-attritors),

- for the labour information, a hot-deck imputation was done. The variables that were used to form the imputation classes combined some variables required to ensure consistency of the longitudinal information (*e.g.*, whether a job was to be held at the beginning of the year, and some global characteristics of that job); and other predictor variables such as age group, sex, marital status, number of jobs held during the previous year ...,

- the donor's information was used to create the labour component for the missing year of information (2) and (3); this allowed internal consistency within the record,

- the labour information that required longitudinal consistency (2) was re-imputed using the longitudinal data (*e.g.*, the employer name, the detailed industry and occupation that were imputed based on the donor's values were replaced by last year's values for that longitudinal job). This minimized the number of artificial changes that would have been introduced by carrying donor's values in fields that were expected to be longitudinally consistent.

Tests are still on-going. The imputation has been a long process to elaborate. Even when imputation classes are fairly large, a fair number of collapsing has to be done to be able to impute for all the non-respondents. Information for non-attritors provides lots of tools for the imputation and initial analysis seem to indicate promising results. However, the strategy developed does not work well for the few cases that have a very complex work history. A different imputation strategy may have to be looked at, for those limited number of cases.

5. OVERALL STRATEGY AND FUTURE WORK

The dilemma between weighting and imputation for handling non-response has been an issue for a number of surveys. Based on the studies that are currently under way, handling non-response in SLID will be a mix of weighting and imputation. Non-response to all waves will be weighted. Non-response for a deferred interview (responding to the labour interview but not to the income or vice-versa) will imply imputation. So far, it is proposed to handle non-attritors through imputation. As for the attritors, the choice between weighting and imputation will probably depend on (1) whether the attritor is part of a household that is completely non-respondent or not; (2) the number of years of data that have been obtained so far and (3) the non-response rates.

Longitudinal imputation is not a straight forward task. The analyses done so far have indicated that there is not a single "best" way to do the imputation of all the variables at the same time. Wave imputation requires the use of different models for different sets of variables. Up to now, imputation has been restricted to labour data. This is a simplification of the process that will have to take place in SLID, since "item" non-response will happen when only one of the two interviews (labour or income) is missing. Since the data in SLID will be collected through computer assisted interviewing, it is hoped that if a labour interview is missing, some minimum labour information can be asked during the income interview to give robustness to the imputation. If income is missing, SLID is looking at the possibility of matching to administrative sources such as the tax file, to be able to help with the imputation process.

Rao proposed a jackknife variance estimator under hot-deck imputation. This will be used to see the impact of imputation on the estimates. The imputation technique will also be assessed by doing more studies comparing transition rates for the non-attritors.

The weighting models show a great deal of promise. Although the differences realized with the model-based approach were small when tested over a one-year interval of non-response, it is expected that the gains will be greater over a longer period. Future plans include testing of the stability of the model for the three year panel of LMAS. Some operational questions have to be addressed if compensation for the attrition non-response is handled through a weighting adjustment. For instances, when the third year of data is added, some of the non-respondents will have information from the first year only, while some non-respondents will have information on both the first and the second year. Exactly how those complexities will be handled in the model has yet to be decided. This issue should be tested on the three year panel. A jackknife estimator for the evaluation of the model is also being developed, to be able to make proper evaluations of the model-based estimates.

In this paper, actions to handle non-response are reactive. The SLID team is also conducting some research on ways to reduce the response burden and get co-operation from respondents. SLID is also trying to implement a tracing strategy to minimize the number of respondents that are lost.

REFERENCES

Harrell, F.E. (1986). *The LOGIST Procedure, SUGI Supplemental Library Guide*, Version 5 Edition, Cary, NC: SAS Institute Inc.

Hosmer, D.W. Jr., and Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley & Sons.

Hunter, L., Michaud, S., and Torrance, V. Modelling for non-response in a longitudinal survey.

Kalton, G. (1986). Handling wave non-response in panel surveys. *Journal of Official Statistics*, 2, 3, 303-314.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. *Panel Surveys*, John Wiley & Sons, 348-374.

Rao, J.N.K. (1992). Jacknife variance estimators under hot-deck imputation. Working paper.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistics Canada publication, Catalogue 71-256.

The Labour Market Activity Survey. 1986-87 Longitudinal File, Microdata User's Guide, Special Surveys Group, Statistics Canada.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

IMPUTATION FOR WAVE NONRESPONSE IN THE SIPP

J. M. Lepkowski, D. P. Miller, G. Kalton and R. Singh[1]

ABSTRACT

Compensation for nonresponse when subjects in a longitudinal survey fail to respond to a wave can be made through imputation from the same or from other subjects. Wave nonresponse was simulated among individuals responding to all seven waves of the SIPP 1987 panel. Simulated missing waves of data were replaced using simple and modified carry-over methods and longitudinal hot-deck methods. Direct comparison of imputed and actual values are made, and the effect of imputation on spell length estimation examined.

KEY WORDS: Hot-deck; Carry-over; Simulated missing data; Spell length estimation.

1. INTRODUCTION

Wave nonresponse is an extended form of item nonresponse in which a sample unit fails to respond to a single or multiple waves of the same panel. Three basic approaches may be applied to compensating for wave nonresponse: weighting, imputation and combinations of the two (Kalton 1986; Lepkowski 1989). Weighting assigns compensatory wave nonresponse adjustments to sample units that do not have missing waves. A single weight may be assigned to sample units responding for all waves of the survey (as in the SIPP) to compensate for any sample unit that fails to respond for one or more waves. Alternatively, multiple weights may be assigned to compensate for specific patterns of wave nonresponse. For example, weights may be assigned to compensate for attrition nonresponse at each wave. Thus, a seven wave panel survey would have seven weights, one for the initial wave nonresponse, and six additional weights compensating for losses at each successive wave. Even more elaborate schemes may be devised, but use of a single weight is favored because it simplifies data manipulation for the analyst.

Imputation replaces an entire missing wave with data either from the same sample unit or from another sample unit. Substantial numbers of waves may be imputed for some wave nonresponse patterns, resulting in the majority of the data being imputed for a given subject. Combinations of wave imputation and weighting are possible, compensating for much but not all missing waves through an imputation, and compensating for the remainder using weights. Such an approach avoids the need for multiple weights and imputation of a large amount of data for a given subject.

The SIPP is a continuing panel survey of the U.S. civilian noninstitutionalized population (Short 1985). A new panel starts each year, and members of original sample households are followed for seven or eight interviews conducted every four months. Bureau of the Census interviewers collect data at each wave on a substantial number of income sources. The 1987 panel examined in this investigation covered seven waves of data collection providing 28 consecutive monthly reports on income and program participation for sample members remaining in panel all seven waves and lesser amounts of data for persons with wave nonresponse. At the conclusion of each panel, a longitudinal file linking individual interviews across waves is created, and served as the data source for this study.

The focus of this research is on the estimation of the duration of spells of receipt of AFDC income over the 28 month period covered by the 1987 panel. More specifically, the paper examines the nature of patterns of wave

[1] J. M. Lepkowski and D. P. Miller, Institute for Social Research, P.O. Box 1248, Ann Arbor, MI 48106, U.S.A. G. Kalton, Westat, Inc., Rockville, MD, U.S.A. R. Singh, U.S. Bureau of the Census, Washington, DC, U.S.A.

nonresponse by full and partial panel respondents and methods for compensating for wave nonresponse (Kalton 1990; Lepkowski 1989). This paper reports on work in progress. It examines the relative accuracy of several simple imputation techniques for wave nonresponse. It does not deal with issues of nonignorable nonresponse (Fay 1986 and Fay 1989), examining only compensation methods suitable for ignorable nonresponse.

The SIPP currently offers a longitudinal data analysis strategy which uses a single longitudinal weight. All sample units with one or more waves of nonresponse are present in the data set but assigned a weight of zero. The remaining full panel respondents are assigned a non-zero longitudinal weight that compensates for the excluded sample units. This approach has been criticized because it excludes a substantial amount of collected data from persons with partial response. For example, sample units with a single wave of interim nonresponse are excluded, despite the presence of six of seven waves of data.

In the next section, patterns of wave nonresponse are reviewed. A method to simulate those patterns among full panel respondents is described in Section 3, and the imputation methods used in the investigation are described in Section 4. Section 5 presents results for the imputation of AFDC recipiency, including the impact of wave imputation on estimation of number of spells and spell length. Section 6 describes briefly further research that is being conducted on imputation of amounts received.

2. WAVE NONRESPONSE IN THE 1987 SIPP

Table 1 presents patterns and frequencies of wave nonresponse (and response) in the 1987 SIPP. Nearly 80% of the sample persons (the unit of analysis in this investigation) responded on all seven waves, or on all waves for which they were eligible members of the population of inference. A small percentage (2.6%) of the panel members became ineligible at a particular wave of the panel, and remained ineligible for all subsequent waves. The rate of ineligibility is relatively constant across waves, with nearly equal losses at each wave due to death, entering the armed forces or an institution, or moving abroad. Cases with any wave nonresponse or ineligible waves are discarded in the SIPP longitudinal weighting.

Among the wave nonresponse patterns, the most frequent are attrition patterns, accounting for three-fifths of all wave nonresponse. Most attrition occurs at the second or third wave. Slightly more than one-quarter of the wave nonresponse patterns are interim nonresponse, where a wave without a response is preceded and followed by a wave with a response. The majority of the interim nonresponse patterns are single wave, although there are a few cases with patterns with as many as four missing interim waves.

Type *Z* nonresponse occurs when data cannot be obtained from a sample person within a household for which other sample persons have responded. The subsequent analysis is limited to sample persons who responded at the first wave of the 1987 panel. However, there are a number of sample persons who were Type *Z* nonresponse at the first wave. We have included them in the table even though, for the purposes of simplicity, we have excluded them from the simulation and imputation exercise. Finally, there are a small number of sample persons who have wave nonresponse prior to becoming ineligible. For the sake of completeness, the investigation includes the wave nonresponse patterns for these sample persons in the simulation and imputation exercise.

3. SIMULATING WAVE NONRESPONSE

3.1 Sample Selection Model for Full Panel Respondents

Kalton and Miller (1986) simulated wave nonresponse patterns among full panel respondents in the first three waves of the 1984 SIPP panel. They employed a SEARCH algorithm which, through a binary splitting algorithm (Sonquist, Baker and Morgan 1973), identified subgroups of sample persons across which there was a substantial variation in patterns of wave nonresponse. They subsequently subsampled full panel respondents within those subgroups. Their procedure allowed a minimum percentage (61.6%) of full panel respondents to be chosen in each group. Full panel respondents were chosen for the simulation exercise in a manner which replicated the population distribution from which they were chosen.

Table 1: Patterns and frequency of wave nonresponse in the 1987 SIPP.

Nonresponse pattern	Frequency	%
Total sample	30,769	100.0
Panel members	24,448	79.5
Responded all seven waves (1111111)[a]	23,653	76.9
Nonresponse due to ineligibility	795	2.6
Died	357	1.2
Entered institution	167	0.5
Entered armed forces	62	0.2
Moved abroad	206	0.7
Reason not listed	3	0.0
Non-panel members	6,321	20.5
Attrition nonresponse	3,887	12.6
1222222	1,453	4.7
1122222	713	2.3
1112222	571	1.9
1111222	480	1.6
1111122	338	1.1
1111112	332	1.1
Interim nonresponse	1,714	5.6
Single wave (*e.g.*, 1211111, 112111)	1,323	4.3
Two waves (*e.g.*, 1221111, 1122111)	298	1.0
Three or more waves	93	0.3
Interim and attrition nonresponse	373	1.2
Type Z non-interview at wave 1	271	0.9
Both nonresponse and ineligible waves	76	0.2

[a] 1 = wave respondent, 2 = wave nonrespondent.

Replicating this approach in an investigation involving seven waves of data would be difficult. Further, the Kalton and Miller approach uses a sample of all wave nonrespondents, decreasing the precision of subsequent analyses. We developed an alternative procedure that employed the existing SIPP wave 1 cross sectional weights and the longitudinal weights to "select" full panel respondents. Figure 1 presents a hypothetical sample to illustrate the simulation selection. Suppose a sample of $n = 24$ subjects has been selected with rate $f = 1/100$, with a population and sample distribution that is one-half female. Each sample person has a base weight W_1 of 100. Only eight of 12 male sample persons respond, while 10 of 12 females do (represented by the response indicator $R_i = 1$). In particular, $Pr\{R_i = 1 \mid X_i = M\} = 0.67$, while $Pr\{R_i = 1 \mid X_i = F\} = 0.83$. To retrieve the original population (and sample) distribution, compensatory nonresponse adjusted weights W_2 are computed, with values of 150 for males and 120 for females.

Wave nonresponse may be handled in a similar manner under a missing at random assumption. Let $F_i = 1$ denote full panel response. Among responding sample persons, $Pr\{F_i = 1 \mid R_i = 1\} = 13/18$. Again, the full panel response rate differs among males and females, and compensatory weights, W_3, of 240 and 150 are computed for males and females, respectively. Under a missing at random assumption, nonresponse at wave 1 and wave nonresponse are compensated by weights varying by subgroups across which the rates vary.

The cross-sectional wave 1 weight W_2 and full panel weight W_3 can be used to derive a weighting simulation alternative to the sampling simulation strategy employed by Kalton and Miller. A model for full panel nonresponse, conditional on a set of characteristics X, is given by

$$Pr\{F_i = 1 | R_i = 1, X\} = \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}\} \,/\, \left[1 + \exp\{\beta_0 + \sum \beta_j X_{ji}\}\right].$$

This conditional probability can be estimated for each full panel respondent, and its inverse used to compensate for full panel nonresponse among wave 1 sample persons.

Figure 1: Hypothetical sample illustrating unit and full panel nonresponse.

i	X_i	R_i	F_i	Base weight W_1	Cross-sectional weight W_2	Full panel weight W_3
1	M	1	1	100	150	240
2	M	1	1	100	150	240
3	M	1	1	100	150	240
4	M	1	1	100	150	240
5	M	1	1	100	150	240
6	M	1	0	100	150	0
7	M	1	0	100	150	0
8	M	1	0	100	150	0
9	M	0	...[a]	100	0	0
10	M	0	...	100	0	0
11	M	0	...	100	0	0
12	M	0	...	100	0	0
13	F	1	1	100	120	150
14	F	1	1	100	120	150
15	F	1	1	100	120	150
16	F	1	1	100	120	150
17	F	1	1	100	120	150
18	F	1	1	100	120	150
19	F	1	1	100	120	150
20	F	1	1	100	120	150
21	F	1	0	100	120	0
22	F	1	0	100	120	0
23	F	0	...	100	0	0
24	F	0	...	100	0	0
Total	--	18	13	2,400	2,400	2,400

[a] Not applicable

Unfortunately, the present investigation did not have the resources to undertake the development of such a model. Instead, we took advantage of the investigation of full panel nonresponse previously conducted by the U.S. Bureau of the Census to develop the full panel weights. We hypothesize that the relationship between the cross-sectional weight W_2 and the full panel weight W_3 is the conditional probability of full panel response that is needed for appropriate weighting. Specifically, we assert that for $W_2 = Pr\{R_i = 1 \mid X\}^{-1}$ and $W_3 = Pr\{F_i = 1 \mid R_i = 1, X\}^{-1}$ the conditional probability of full panel response is W_2 / W_3. Thus a compensatory weight for full panel response which conditions on the (in our case) unknown X is W_3 / W_2.

We use the ratio W_3 / W_2 as a weight for full panel respondents in the simulation exercise. Wave nonresponse patterns are assigned to full panel respondents at random according to the distribution observed in the data. In order to retrieve the sample distribution of characteristics X used in the "selection" of full panel respondents, and other characteristics Z not used in that selection, the results are weighted by $\{1 - W_3 / W_2\}^{-1}$. This method is not subject to losses in precision due to sample selection (as in Kalton and Miller 1986), although some loss in precision due to weighting is experienced.

3.2 Assigning Wave Nonresponse Patterns

Each full panel respondent was assigned a wave nonresponse pattern at random. The assignment could not be made completely at random because panel members with periods of ineligibility were also assigned patterns. Many wave nonresponse patterns cannot be assigned to panel members with periods of ineligibility. Let ϕ_{ij} = Pr{sample person with true *i*-th panel response pattern has the *j*-th wave nonresponse pattern}. Values ϕ_{ij} could be computed from a model of independence between full panel response pattern and wave nonresponse pattern, except for the presence of "structural zeroes," combinations of response patterns and wave nonresponse patterns that are impossible. Figure 2 illustrates the problem. Let 0 represent ineligibility at a given wave. Wave nonresponse pattern corresponding to attrition at wave 2, 1222222, can be assigned to any panel response pattern since during a period of wave nonresponse a sample person could become ineligible without our knowledge. On the other hand, wave nonresponse pattern 1122222 cannot be assigned to panel response pattern 1000000, since we cannot assign receipt of an interview at wave 2 to a person who was known to be ineligible at that wave.

Figure 2: Assignment of wave nonresponse pattern to full panel response patterns.

Wave non-response pattern (j)	n_j	ϕ_j	Panel response pattern (i)				
			1111111	1000000	1100000	. . .	1111110
1222222	1,453	0.047	ϕ_{11}	ϕ_{21}	ϕ_{31}	. . .	ϕ_{71}
1122222	713	0.023	ϕ_{12}	0	ϕ_{32}	. . .	ϕ_{72}
1112222	571	0.019	ϕ_{13}	0	0	. . .	ϕ_{73}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
Total	6,321	1.000	$\phi_{1\cdot}$	$\phi_{2\cdot}$	$\phi_{3\cdot}$	. . .	$\phi_{7\cdot}$

A model of quasi-independence (Agresti 1989) allows the computation of estimates of the joint probabilities for each combination of wave nonresponse and panel response patterns. Once the joint probabilities ϕ_{ij} were computed from the quasi-independence model, a random assignment of wave nonresponse patterns was made to the full panel respondents and other panel members. The weighted distribution of simulated wave nonresponse patterns matched the observed for the total sample and approximately 20 subgroups (*e.g.*, persons ages 18-24 females).

4. IMPUTATION METHODS

4.1 Carry-over Methods

In a simple carry-over method, values in nonmissing months are carried forward within the same record to impute for missing months later in the 28 months period. Since all cases with missing data on wave 1 have been excluded from this analysis (including Type *Z* nonresponse at wave 1), the simple carry-over imputation begins with wave 2. If wave 2 is missing, the value from month 4 (*i.e.* the last month in Wave 1) is carried forward for all four months of wave 2. If wave 3 is missing, the value from month 8 (*i.e.* the last month of wave 2) is carried forward to all four months of wave 3. If the wave 2 data used to impute wave 3 had been imputed from month 4, then wave 3 data consist of the value of month 4 for each month. The process is repeated sequentially through the remaining waves. The simple carry-over imputation does not introduce any change in recipiency status. Months in all nonresponse waves receive the same value, that carried forward from the last month in the last nonmissing wave.

Another carry-over procedure was also implemented that, for interim nonresponse, carried data forward as well as backward. A random number was chosen from one to the total number of months to be imputed. The random selection was not uniform, but gave higher probabilities (based on an empirical distribution of the months at which change occurred) to months at the end of a wave. This effectively represented the "seam" problem (see, for example, Singh, Weidman and Shapir 1989; Coder and Ruggles 1988; or Murray, Michaud, Egan and Lemaitre 1991) in the imputed data. All months following the randomly selected month received a "carry-back" imputation from the first month of the next nonmissing wave, while all months up to and including the selected month received a "carry-forward" imputation from the last month of the last nonmissing wave.

4.2 Longitudinal Methods

A basic longitudinal hot-deck imputation procedure was used. The data were sorted by seven variables: gender, age (five categories), race (four categories), Hispanic origin status, family income to poverty level value at the first wave (five categories), sampling stratum, and half-sample code. An upper triangular matrix of "hot" values was created (see Figure 3) in which up to a total of 24 months could be imputed for a sample person. Depending on the wave at which wave nonresponse started (*e.g.*, wave 2, wave 3, *etc.*), the hot values were obtained from the corresponding wave of the matrix.

Two values on the recipient record were examined: the first month recipiency status (for example Food Stamps in month 1) and the recipiency in the last month preceding the wave nonresponse. The combination of these two items for the donor were matched to the values of the recipient from the appropriate wave to be imputed. Thus, donations were made from sample persons similar to the imputation recipient on the sort variables, and matching exactly on the month of recall (*i.e.* wave), first month recipiency status, and last month recipiency status. In order to provide initial or "cold" values for the matrix prior to the imputation, the data were passed through the imputation procedure in reverse order without donations being made, filling in the cells in the matrix prior to imputation.

Figure 3: Match characteristics of program-specific longitudinal hot-deck procedure and hypothetical "hot" values.

Match characteristics			Hot-deck values to be imputed for each month			
	Food Stamps					
Wave	Month 1	Month t-1	5 6 7 8	9 10 11 12	. . .	25 26 27 28
2	On	On	X X X X	O O O O	. . .	O O O O
	On	Off	O O O O	O O O O	. . .	O O O O
	Off	On	X X X X	X X X X	. . .	O O O O
	Off	Off	O O O O	X X X X	. . .	X X X X
3	On	On	(Not applicable)	X X X X	. . .	X X X X
	On	Off		O O O O	. . .	X X O O
	Off	On		X X O O	. . .	O O O O
	Off	Off		O O O X	. . .	X O O O
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
7	On	On	(Not applicable)	(Not applicable)	. . .	X X X X
	On	Off			. . .	O O O O
	Off	On			. . .	O O O O
	Off	Off			. . .	X X X X

Two other hot-deck approaches were also used. In one, the recipiency status in the first and last month for one program was used to establish an exact match in the matrix, but the values imputed were for another program. For example, donors and recipients were matched on Food Stamps recipiency in the first and last months, but

AFDC status in missing months was imputed from donor to recipient. This alternative allowed examination of the quality of matches for the type of joint or simultaneous imputation that is typically performed using a hot deck where multiple variables are imputed simultaneously using a single sort and match procedure. In the second, the recipiency match was altered to use recipiency for two programs simultaneously. The last month's recipiency for each program was used to obtain an exact match in the matrix. This alternative is referred to as a joint hot-deck because it depends on the joint distribution of recipiency for the two programs being imputed simultaneously.

Thus, for AFDC recipiency imputations were made using three longitudinal hot-deck methods. The AFDC hot-deck imputed from a match on first and most recent month of AFDC recipiency. The Food Stamp hot-deck imputed from a match on first and last month Food Stamp recipiency. Finally, the joint hot-deck used a match on the last month of AFDC and Food Stamp recipiency.

5. IMPUTING PROGRAM RECIPIENCY

5.1 Imputing Months

Figure 4 defines notation for comparing actual and imputed values for each month. For example, "a" denotes the number of months where the imputed value agrees with the actual when the actual value is "on". The proportion a/(a+c) is the accuracy rate among months "on" spell, while d/(b+d) is the accuracy for months that are actually "off" spell.

Figure 4: Relationship between actual and imputed month.

	Actual		
Imputed	On	Off	Total
On Off	a c	b d	a+b c+d
Total	a+c	b+d	h

Table 2 shows the accuracy of imputations for the five imputation techniques, separately for months "on" and "off" AFDC, together with the proportion of months that would have been imputed "on" or "off" spell correctly just by chance alone (under a simple model of independence between imputed and actual values). The estimates are shown only for the total number of months imputed. Months that were missing due to attrition nonresponse starting at wave 2, to single wave interim nonresponse, and other patterns of wave nonresponse were also examined, but the results are not shown here since they are essentially the same.

The probability of obtaining a correct imputation is, by chance alone, high when the month is actually "on" spell. Yet each of the imputation methods exceeds chance, handling correctly in all but one case (the Food Stamps hot-deck) more than one-half of the imputations. On the other hand, the probability of predicting "off" spell is, by chance alone, quite low. All of the imputation procedures predict correctly far in excess of chance. The simple and modified carry-over procedures perform the best, and just about identically well. The hot-deck procedures perform considerably poorer, with, again, the Food Stamps hot-deck having the lowest level of accuracy.

It is clear that the imputation of AFDC recipiency is least accurate when Food Stamps is used as the matching criteria. Further, given that the joint hot-deck performs nearly as well as the AFDC hot-deck, it appears that the critical feature of the match is the last month (which is used in both the AFDC and joint hot-decks) rather than the first month.

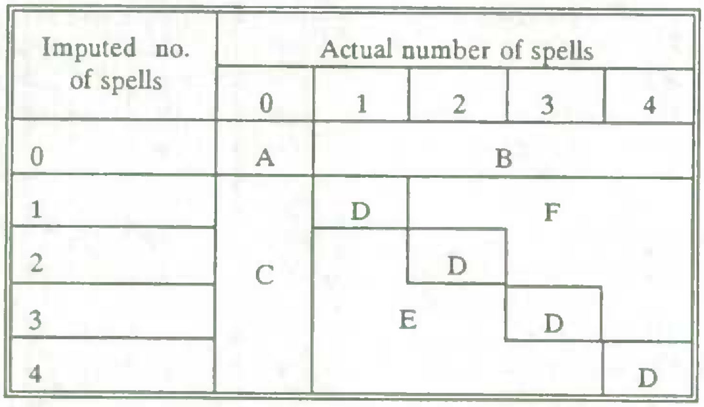
Table 2: Probability of correct imputation by method for AFDC recipiency.

Imputation method	Pr{correct|On}		Pr{correct|Off}	
	a/(a+c)	Deviation from chance	d/(b+d)	Deviation from chance
All months				
Chance/expectation	0.957	-	0.043	-
Simple carry-over	0.987	+0.030	0.868	+0.825
Modified carry-over	0.988	+0.031	0.867	+0.824
AFDC hot-deck	0.988	+0.031	0.701	+0.658
FS hot-deck	0.974	+0.017	0.489	+0.446
Joint hot-deck	0.987	+0.030	0.697	+0.654

5.2 Imputing Spells

The carry-over imputation method does not create spells. Figure 5 presents a summary comparison of the actual and imputed spells for each person. We expect by chance alone that, because most persons will have no AFDC recipiency, most persons will fall in cell A (agreement given no actual spells). The methods probably also do not vary greatly in their ability to impute correctly when there actually is a spell, corresponding to type D cells (agreement given one or more actual spells). Carry-over methods miss spells, imputing no spells when some exist as in type B cells (missed spells given one or more actual spells) or fewer spells than actually occur as in type F cells (missed spells given two or more actual spells). On the other hand, carry-over imputation methods cannot create spells, and can have no persons classified in cells of type C (created spells given no actual spells) or type E (created spells given one or more actual spells) for carry-over imputations. Because the carry-over methods cannot create spells, they are expected to under-estimate the total number of spells.

Figure 5: Actual and imputed number of spells: Classification of cells for AFDC recipiency by person.



Imputed no. of spells	Actual number of spells				
	0	1	2	3	4
0	A	B			
1	C	D	F		
2		E	D		
3				D	
4					D

Table 3 presents the rate of these various types of spell agreement, missed spells, and created spells. While carry-over imputations have higher levels of agreement they are biased by their inability to create new spells. Carry-over imputations completely miss short duration spells (*i.e.* those lasting less than one wave), and they create no other spells to compensate for them. On the other hand, hot-deck imputations miss some short duration spells, capturing others, and they also create other spells that compensate for the loss of short duration spells. The net effect should be a greater accuracy for number of spells for hot-deck methods overall because they have the capacity to create spells while carry-over methods do not.

Table 3: Actual and imputed number of spells for AFDC recipiency by imputation method.

Imputation method	Agreement | 0 spells (A)	Agreement | 1 + spells (D)	Missed spells | 1 + spells (B)	Created spells (C)	Created spells | 1 + spells (E)	Missed spells | 2 + spells (F)
All sample persons						
Simple carry-over	0.967	0.018	0.013	0.000	0.000	0.002
Modified carry-over	0.967	0.018	0.013	0.000	0.000	0.002
AFDC hot-deck	0.956	0.019	0.010	0.013	0.001	0.002
FS hot-deck	0.932	0.019	0.009	0.034	0.014	0.002
Joint hot-deck	0.954	0.020	0.009	0.013	0.002	0.002

Table 4 provides several summary rates of the various types of errors. Carry-over imputation methods cannot by definition have "false positive" imputations or created spells, but they do have false negative and missed imputations. The net effect is that carry-over imputations under-estimate the total number of spells. Hot-deck imputations have approximately the same rate of false positive and false negative imputations, but a slight excess of missed spells. Thus, contrary to expectation, the hot-deck imputation methods have a tendency to under-estimate the number of spells as well. Given the small number of AFDC spells that are involved, the under-estimation for both types of imputation is expected to be small.

Table 4: Error rates for number of imputed spells of AFDC by imputation method: All sample persons.

Imputation Method	False positive C/ (A+C)	False negative B/ (A+B)	Missed spells | 1+ spells (B+F)/ (B+F+D+E)	Created spells (C+E)/ (C+E+D+F)
Simple carry-over	0.000	0.013	0.443	0.000
Modified carry-over	0.000	0.013	0.443	0.000
AFDC hot-deck	0.012	0.010	0.264	0.545
FS hot-deck	0.036	0.010	0.164	0.779
Joint hot-deck	0.013	0.010	0.247	0.562

The under-estimation of number of spells by the carry-over methods is further illustrated in Table 5 in which the distribution of the number of spells is shown for the complete data (*i.e.* the actual data for the full panel respondents) and by imputation method. The carry-over imputation methods show a marked tendency to under-estimate the total number of spells because they have more single spell persons and fewer two spell persons than the complete data. The AFDC and joint hot-decks have a distribution of spells that is closer to that in the complete data.

5.3 Spell Length Estimation

The ability to accurately estimate the total number of spells is related to a second estimation problem, spell length. Table 6 presents spell length estimates for the complete data and for each imputation method. Shown are the Kaplan-Meier estimates of AFDC spell length computed using appropriate weights (see Miller, Lepkowski and Kalton 1992, for a discussion of the estimation procedure). The Kaplan-Meier estimates for the complete data represent the "unbiased values" that are to be estimated using data with imputations made by the various methods. The results are consistent with those presented in the last section. Carry-over imputation methods show a tendency to over-estimate spell length. For example, at six months the true proportion of the sample with AFDC spells which lasted longer are 0.539. Both carry-over procedures estimate higher proportions with longer spells. On the other hand, the hot-deck procedures provide estimates that at six months are slightly lower than the actual. There is, if anything, a slight tendency for the hot-deck procedures to under-estimate spell length. The hot-deck under-estimation does not appear to be quite as severe as the over-estimation of the carry-over procedures.

Table 5: Percent distribution of AFDC spells by method.

Imputation Method	Number of spells			
	1	2	3	4
Complete data	79.1	17.7	2.7	0.5
Simple carry-over	81.2	16.0	2.2	0.6
Modified carry-over	81.2	16.0	2.2	0.6
AFDC hot-deck	78.3	18.3	2.7	0.5
FS hot-deck	76.1	20.6	3.0	0.5
Joint hot-deck	78.7	18.3	2.5	0.5

Table 6: Kaplan-Meier estimates of AFDC spell length by imputation method.

Imputation Method	Percent of AFDC recipiency lasting more than...				
	1 month	3 months	6 months	12 months	24 months
Complete data	85.4	74.0	53.9	35.6	27.2
Simple carry-over	85.6	75.0	55.2	38.2	31.4
Modified carry-over	85.4	74.7	54.8	38.2	31.5
AFDC hot-deck	85.9	75.0	52.7	35.4	26.5
FS hot-deck	87.9	78.1	52.0	35.5	26.6
Joint hot-deck	86.0	76.0	52.9	36.3	27.5

6. CONCLUDING REMARKS

The imputation exercise reported here is part of work in progress on simple longitudinal imputation procedures for compensating for entire missing waves in a panel survey. The work has been limited to imputing for recipiency of income from a small number of programs. Results indicate that carry-over methods are more accurate for imputing the spell status of individuals months, but since carry-over methods cannot create spells, they are biased with respect to estimating number of spells and spell length. The hot-deck procedures perform poorly for imputing monthly spell status, but are somewhat better for imputing number of spells and spell length.

The investigation will shift emphasis from recipiency to amounts received. The plan is to impute amounts using a similar approach to that illustrated in Figure 3, matching on categories of income in the first and last month. Joint imputation matching will be on categories of income, and possibly categories of income for one program and recipiency status for another. Measures such as mean-squared deviations and difference between complete and imputed data for total annual income will be used to assess accuracy. Simple product moment correlations will be used to examine attenuation of associations.

REFERENCES

Coder, J., and Ruggles, P. (1988). *Welfare Recipiency as Observed in the SIPP*. SIPP Working Paper 8818. U.S. Bureau of the Census, Washington, DC.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Joumal of Official Statistics*, 2, 303-314.

Kalton, G., Hill, D., and Miller, M. (1990). *The Seam Effect in Panel Surveys*. SIPP Working Paper No. 9011. U.S. Bureau of the Census, Washington, DC.

Kalton, G., and Miller, M. (1986). Effects of adjustments for wave nonresponse on panel survey estimates. *Proceedings of the Survey Research Methods Sections, American Statistical Association.*

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys*, D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, Eds. New York: J.W. Wiley and Sons, Inc., 348-374.

Miller, D.P., Lepkowski, J.M., and Kalton, G. (1992). Estimating duration of food stamp spells from the SIPP. *Proceedings of the Survey Research Methods Section, American Statistical Association*, (forthcoming).

Murray, T.S., Michaud, S., Egan, M., and Lemaître, G. (1991). Invisible seams? The experience with the Canadian Labour Market Survey. *Bureau of the Census 1991 Annual Research Conference Proceedings.*

Short, K.S. (1985). *Survey of Income and Program Participation: Uses and Applications*. SIPP Working Paper 8501. U.S. Bureau of the Census, Washington, DC.

Singh, R., Weidman, L., and Shapiro, G. (1989). *Quality of the SIPP Estimates.* SIPP Working Paper 8901, U.S. Bureau of the Census, Washington, DC.

SESSION 4

Using Longitudinal Structure in Estimation

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

LONGITUDINAL SMOOTHING OF PRICE INDEX VARIANCES

R. Valliant[1]

ABSTRACT

This paper develops generalized variances for price indexes by applying nonparametric scatterplot smoothers to time series of point variance estimates. The goal here is to formulate smoothed variances which are approximately unbiased, which provide acceptable confidence interval coverage, and which are more stable than the point variance estimates. Smoothing methods are applied to time series of point variance estimates in a simulation study using data from the U.S. Consumer Price Index program.

KEY WORDS: Generalized variance function; Laspeyres price index; Linearization variance estimator; Loess; Super smoother.

1. INTRODUCTION

Index series are characterized by seasonal and irregular fluctuations in addition to underlying trends. The literature is replete with methods for decomposing and smoothing such time series. Point estimators of variance, obtained by linearization, replication, or another method, may be subject to the same types of seasonal and irregular variations as the index series themselves. The variable nature of point variance estimates was illustrated by Leaver (1990) for indexes. This paper explores the possibility of developing generalized variances for price indexes by applying nonparametric scatterplot smoothers to series of point variance estimates. The goal here is to formulate smoothed variances which are approximately unbiased, which provide acceptable confidence interval coverage, and which, most importantly, are more stable than the point variance estimates.

The approach taken here is somewhat different than that which is sometimes used in household surveys for estimating generalized variance functions (*GVF's*). That method is described in Wolter (1985) with some justifying theory given in Valliant (1987). The general idea is to use models to approximate variances. Given a set of survey variables whose variances all follow the same model, parameters of the model are estimated by least squares. The parameter estimates are then provided to users rather than individual variance estimates in order to condense survey publications. Ideally, the models will also lead to more stable estimates of variance. Applications of *GVF*'s in two particular surveys can be found in Hanson (1978) and Johnson and King (1987). In the case of price indexes, finding multiple indexes whose variances follow the same model may be difficult. However, smoothing the variances of a particular index series over time is a practical alternative. For a given index series this is a two-step process consisting of estimating variances at a number of points in time and of smoothing the series of point variance estimates. As will be illustrated, this approach can produce more stable variance estimates that are approximately unbiased and that provide near nominal confidence interval coverage.

Section 2 defines the population Laspeyres price index, a class of index estimators, and a superpopulation model which is used to study the variance of the index estimators. In section 3, an approximation to the variance of a long-term price change estimator is discussed. The fourth section presents the methods that were tested for estimating generalized variances. A simulation study, described in section 5, was conducted using data from the U.S. Consumer Price Index to determine how well the proposed variance estimators would work in practice. Finally, section 6 gives conclusions.

[1] R. Valliant, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Avenue NE, Washington DC 20212, U.S.A.

2. INDEX ESTIMATORS AND A SUPERPOPULATION MODEL

The population is divided into H strata with stratum h containing N_h establishments. Establishment (hi) contains M_{hi} items, and the total number of items in all establishments in stratum h is $M_h = \sum_{i=1}^{N_h} M_{hi}$. At time t the price of item j in establishment (hi) is p_{hij}^{t}, and the price relative between time t and the base period time 0 is $r_{hij}^{t,0} = p_{hij}^{t} / p_{hij}^{0}$. The quantity of item (hij) purchased in the base period is q_{hij}^{0}. The finite population value of the long-term fixed base Laspeyres price index for comparing period t to period 0 is

$$
\begin{aligned}
I^{t,0} &= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} p_{hij}^{t} q_{hij}^{0} \Big/ \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} p_{hij}^{0} q_{hij}^{0} \\
&= \sum_h \sum_i \sum_j W_{hij}^{0} r_{hij}^{t,0} ,
\end{aligned}
\tag{1}
$$

where $W_{hij}^{0} = p_{hij}^{0} q_{hij}^{0} / \sum_{h,i,j} p_{hij}^{0} q_{hij}^{0}$ is the fraction of total base period cost or value accounted for by item (hij). For later reference, it is also convenient to define the stratum index $I_h^{t,0} = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} W_{hij}^{0} r_{hij}^{t,0} / W_h^{0}$ where $W_h^{0} = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} W_{hij}^{0}$. Using long-term indexes, the population short-term index for comparing periods t_2 and t_1 $(t_1 < t_2)$ is defined as $I^{t_2,t_1} = I^{t_2,0} / I^{t_1,0}$. Monthly, quarterly, semiannual, and annual changes are commonly published by index programs.

In order to analyze the properties of index estimators, we will consider the superpopulation model defined below, which was also used in Valliant (1991).

$$
\begin{aligned}
r_{hij}^{t,0} &= \alpha_{th} + \omega_{thi} + \epsilon_{thij} \\
\epsilon_{thij} &= \rho_h \epsilon_{t-1,hij} + \xi_{thij} ,
\end{aligned}
\tag{2}
$$

where $E(\omega_{thi}) = E(\omega_{t_1hi}\omega_{t_2h'i'}) = 0$ for all t, h, i, and $(t_1hi) \neq (t_2h'i')$; $E(\omega_{thi}^2) = \sigma_{\omega h}^2$; $E(\xi_{thij}) = E(\xi_{t_1hij}\xi_{t_2h'i'j'}) = 0$ for all t, h, i, j and $(t_1hij) \neq (t_2h'i'j')$; $E(\xi_{thij}^2) = \sigma_{\xi h}^2$; and $-1 < \rho_h < 1$. By convention, define $\alpha_{0h} \equiv 1$ and $\epsilon_{0hij} \equiv 0$. Considering times only back to the base period and not beyond, (2) implies that $\epsilon_{thij} = \sum_{k=0}^{t-1} \rho_h^k \xi_{t-k,hij}$. Using this expression and the properties of ξ_{thij}, the covariance structure implied by model (2) is

$$
\operatorname{cov}(r_{hij}^{t_2,0}, r_{h'i'j'}^{t_1,0}) =
\begin{cases}
\sigma_{\omega h}^2 + (1 - \rho_h^{2t_1}) \Delta_h^2 & t_2 = t_1, h = h', i = i', j = j' \\
\rho_h^{t_2 - t_1} (1 - \rho_h^{2t_1}) \Delta_h^2 & t_1 < t_2, h = h', i = i', j = j' \\
\sigma_{\omega h}^2 & t_2 = h_1, h = h', i = i', j \neq j' \\
0 & \text{otherwise}
\end{cases}
\tag{3}
$$

where $\Delta_h^2 = \sigma_{\xi h}^2 / (1 - \rho_h^2)$. Expression (3) implies that price relatives for a particular item are correlated over time. At a given time period, items within a particular establishment are also correlated, while other items are not.

The sample design addressed here is a rotating panel survey in which establishments are sampled as the first-stage units. Establishments are retained in the sample for a specified period of time and then rotated out and replaced by new units. At each time $t (t=1, ..., T)$, we have a sample s_{th} of n_h establishments from the N_h establishments in stratum h and a sample s_{thi} of m_{hi} items from the M_{hi} items in sample establishment (thi). A two-stage sampling plan, often approximated in practice, is one in which establishments are selected with probabilities proportional to $W_{hi}^{0} = \sum_{j=1}^{M_{hi}} W_{hij}^{0}$. Items within establishments are then selected with probabilities proportional to W_{hij}^{0}. Surrogate measures of size, felt to be closely related to W_{hi}^{0} and W_{hij}^{0}, such as current sales values or employment are often used in practice. At each time period, the total establishment sample size is assumed to be constant at $n = \sum_h n_h$ with the total number of sample items in stratum h being $m_h = \sum_{i \in s_{th}} m_{hi}$. At each time period a proportion δ_h of the sample establishments is rotated out in stratum h and an equal number rotated in. The size of the overlap, $s_{tuh} = s_{th} \cap s_{uh}$, between samples from time t and $u\,(t \geq u)$ is $\max\left\{0, n_h\left[1-(t-u)\delta_h\right]\right\}$.

The class of estimators considered here was introduced in Valliant and Miller (1989) for one-stage sampling and generalized in Valliant (1991). For the long-term index, define

$$\hat{I}^{t,0} = \sum_h \bar{z}^{t}_{th} \prod_{u=1}^{t-1} \left[\frac{\bar{z}^{u}_{uh}}{\bar{z}^{u}_{u+1,h}}\right]^{\gamma^{tu}_h}, \tag{4}$$

where $\bar{z}^{u}_{kh} = \sum_{i \in s_{kh}} \lambda_{hi} \bar{r}^{u,0}_{khi}$, $\bar{r}^{u,0}_{khi} = \sum_{j \in s_{khi}} r^{u,0}_{hij} / m_{hi}$ for k=u or u+1 (u=1,...,t−1) and γ^{tu}_h is a real number. The term λ_{hi} is a coefficient which does not depend on the model random variables $r^{t,0}_{hij}$. For the two stage probability-proportional-to-size design mentioned above, for example, $\lambda_{hi} = W^0_h / n_h$. We restrict consideration to cases where

$$\sum_{i \in s_{th}} \lambda_{hi} = W^0_h$$

in which case $E(\bar{z}^{u}_{kh} - W^0_h I^{u,0}_h) = 0$. When all within-stratum samples of establishments are large, $\hat{I}^{t,0}$ is approximately model-unbiased under (2). Short-term estimators are defined by taking ratios of long-term estimators. The price change from time t_1 to $t_2 (t_1 < t_2)$ is estimated by $\hat{I}^{t_2,t_1} = \hat{I}^{t_2,0} / \hat{I}^{t_1,0}$.

A number of estimators in class (4) are listed in Valliant (1991). Three are of particular interest. If $\gamma^{tu}_h \equiv 1$, then (4) is the product estimator, which can be written as

$$\hat{I}^{t,0}_1 = \sum_h \prod_{u=1}^{t} \left[\frac{\bar{z}^{u}_{uh}}{\bar{z}^{u-1}_{u,h}}\right]$$

with $\bar{z}^0_{1h} \equiv 1$. If $\gamma^{tu}_h \equiv 0$, (4) reduces to the simple index estimator

$$\hat{I}^{t,0}_2 = \sum_h \bar{z}^{t}_{th}.$$

A third choice of γ^{tu}_h is the one which minimizes the approximate variance of $\hat{I}^{t,0}$ under model (2). The optimum is complicated in general, but in the special case of a constant number of sample items per establishment, $m_{hi} = \bar{m}_h$, and λ_{hi} a constant for all sample establishments in stratum h, then the optimum reduces to

$$\overset{*}{\gamma}{}^{tu}_h = \frac{1}{2} \frac{\alpha_{uh}}{\alpha_{th}} \rho^{t-u}_h \left[1 + \bar{m}_h \frac{g_{uh}}{1 - g_{uh}}\right]^{-1}$$

for $1 \leq u \leq t-1$ and $g_{uh} = \sigma^2_{\omega h} / \left[\sigma^2_{\omega h} + (1 - \rho^{2u}_h)\Delta^2_h\right]$.

3. APPROXIMATE VARIANCES UNDER THE MODEL

When the establishment sample size n_h is large in each stratum, the long-term index estimator can be approximated, as shown in Appendix A of Valliant (1991), by

$$\hat{I}^{t,0} \doteq \sum_h \left\{\bar{z}^{t}_{th} + \sum_{u=1}^{t-1} \gamma^{tu}_h \frac{\alpha_{th}}{\alpha_{uh}} \left(\bar{z}^{u}_{uh} - \bar{z}^{u}_{u+1,h}\right)\right\}. \tag{5}$$

Before presenting the variance of this approximation, it is instructive to compare (5) to a similar one for composite estimators. In repeated surveys a common form of composite estimator (Cantwell 1990) is

$$\hat{x}_c^t = (1-k)\hat{x}_{s_t}^t + k\left(\hat{x}_c^{t-1} + \phi_{s_{t,t-1}}^t\right),$$

where $\hat{x}_c^t$ is the time t composite estimator of a total, k is a weight between 0 and 1, $\hat{x}_{s_t}^t$ is an estimated total based on the sample at time t only, and $\phi_{s_{t,t-1}}^t = \hat{x}_{s_{t,t-1}}^t - \hat{x}_{s_{t,t-1}}^{t-1}$ is an estimate of the change between $t-1$ and t based on the units in the sample at both times t and $t-1$. Repeated substitution gives

$$\hat{x}_c^t = \tilde{\hat{x}}_c^t + \sum_{u=1}^{t-1} k^{t-u}\left(\tilde{\hat{x}}_c^u - \hat{x}_{s_{u+1,u}}^u\right),$$

where $\tilde{\hat{x}}_c^t = (1-k)\hat{x}_{s_t}^t + k\hat{x}_{s_{t,t-1}}^t$. Thus, the composite estimator can be written as a time t estimator $\tilde{\hat{x}}_c^t$ plus a sum of estimators of 0. Judging from (5), the same is approximately true for an estimator in the γ class. As a result, the method of variance smoothing discussed below should also apply to certain types of composite estimators.

Using (5) and results in the appendix of Valliant (1991), we can write the approximate variance of the long-term estimator as

$$\operatorname{var}(\hat{I}^{t,0}) \doteq \sum_h \left\{\sum_{u=1}^{t-1} a_{tuh}(\iota_h^{t,u})^2 + 2\sum_{u=1}^{t-1} b_{tuh}\iota^{t,u} + c_{th}\right\}, \tag{6}$$

where

$$\iota_h^{t,u} = \alpha_{th}/\alpha_{uh} = E(I_h^{t,0})/E(I_h^{u,0}),$$

$$a_{thu} = (\gamma_h^{tu})^2\left[\sum_{i\in C_{uh}} \frac{\lambda_{hi}^2}{m_{hi}} v_{uhi} + \sum_{i\in D_{uh}} \frac{\lambda_{hi}^2}{m_{hi}} v_{uhi}\right],$$

$$b_{tuh} = \gamma_h^{tu}\rho_h^{t-u}(1-\rho_h^{2u})\Delta_h^2\left[\sum_{i\in s_{tuh}} \frac{\lambda_{hi}^2}{m_{hi}} - \sum_{i\in s_{t,u+1,h}} \frac{\lambda_{hi}^2}{m_{hi}}\right],$$

$$c_{th} = \sum_{i\in s_{th}} \frac{\lambda_{hi}^2}{m_{hi}} v_{thi},$$

where $v_{uhi} = v_{uh}\left[1 + (m_{hi}-1)g_{uh}\right]$, $v_{uh} = \sigma_{uh}^2 + (1-\rho_h^{2u})\Delta_h^2$, $C_{uh} = s_{uh} - s_{u+1,h}$, *i.e.* the part of s_{uh} that is not contained in $s_{u+1,h}$, and $D_{uh} = s_{u+1,h} - s_{uh}$.

An expression similar to (6) can also be worked out for the approximate variance of the short-term index estimator $\hat{I}^{t_2,t_1}$.

4. GENERALIZED VARIANCE FUNCTIONS FOR INDEXES

Judging from (6), the approximate variance is a second order polynomial in the stratum superpopulation short-term indexes $\iota_h^{t,u}$. This is analogous to the relationship between an estimator $\hat{T}$ of the population total T in two-stage sampling and its approximate variance derived in Valliant (1987) for a particular class of models in which the variance of a unit was a quadratic function of the unit's mean:

$$\operatorname{var}(\hat{T}) \doteq aE(T)^2 + bE(T). \tag{7}$$

The terms a and b are coefficients that depend on various quantities, such as intracluster correlations, numbers of population and sample units within clusters, and coefficients in the estimator $\hat{T}$. In fitting the *GVF* model defined by (7), the usual procedure is to select a group of variables which all have the same a and b coefficients, calculate point estimators of variance for each of the variables, and then to estimate a and b by some form of least squares. Application of this course to (6) would be fraught with practical difficulties. In (7) there are only two regression coefficients to be estimated - a and b. In (6) there are $2(t-1)+1$. As t increases so does the number of coefficients. The components of the coefficients, a_{tuk}, b_{tuk}, and c_{tk}, are also complex, so that identifying different indexes that all follow model (6) would be a problem.

An alternative approach is to work with a particular index series and attempt to model the behavior of its variance over time. If $\iota_k^{t,u}$ is a smooth function of time, *e.g.* a polynomial in $t-u$, then the variance (6) will also be a smooth function of time, say, $f(t)$. If an unbiased, or approximately unbiased, variance estimator is used for $\hat{I}^{t,0}$, then its expectation can also be described by $f(t)$. As data are accumulated over time, a time series of point variance estimates is developed and the function $f(t)$ can be fitted by a scatterplot smoother without having to know the explicit form of the function. A number of such smoothers are available, and we will consider two that have proved to be useful in other situations.

The two smoothers used here are the super smoother (Friedman 1984) and loess (Cleveland 1979, Cleveland, Cleveland, McRae, and Terpenning 1990). The two algorithms are fairly complex to describe in detail, so that only rough sketches will be given here. Both methods use local linear fits in neighborhoods around each point t. A critical parameter in both algorithms is the span, the size of the neighborhood around t, which is used to estimate $f(t)$. In loess the span is fixed while for the super smoother, spans can be variable. Of the two, loess explicitly incorporates features to reduce the effects of outlying values and tends to produce a smoother looking curve of estimates. The variable span used by super smoother allows it to adapt more readily to changing curvature in $f(t)$. Super smoother also has the advantage of being computationally faster than loess.

5. AN EMPIRICAL STUDY

A simulation study, using a population derived from data collected for the U.S. Consumer Price Index program by the BLS, was undertaken to test the usefulness of the proposed method of calculating *GVF*'s. The population was composed of establishments and items and was described in detail in Valliant (1991). Its main features are briefly recounted here. The six hundred and fifty-nine establishments in the population were divided into the five strata. Each establishment contained an average of just under 10 items with each item having prices for 42 consecutive months.

Two sets of 500 stratified two-stage samples were selected with the number of sample establishments allocated to each stratum being roughly proportional to W_h^0. The total establishment sample sizes in the two sets of samples were $n = 50$ and 100. Samples were selected in such a way that 20% of the sample establishments were rotated in each 12-month period. This was done by first selecting a large systematic, random-start sample of establishments in each stratum with probabilities proportional to W_{hi}^0. For samples of size $n = 50$, the initial, large sample size was 84 and was 168 for the samples of size $n = 100$. These initial samples were large enough to accommodate all 42 months accounting for the amount of establishment rotation. The initial sample from each stratum was then sorted in a random order. For a particular time period t, the stratum establishment sample consisted of establishments $1+(t-1)n_h\delta_h, \ldots, n_h+(t-1)n_h\delta_h$, where δ_h was the proportion of establishments rotated in a month. For both the cases of $n = 50$ and $n = 100$, $\delta_h = 1/60$ which resulted in an annual turnover of $12(n_h/60) = n_h/5$ establishments of 20%. From each sample establishment, $\bar{m}_h = 2$ sample items were selected systematically with probability proportional to W_{hij}^0.

From each sample, the long-term product estimators $\hat{I}_1^{t,0}$ $(t=1,\ldots,42)$, and the short-term estimators of 1-month and 12-month change were computed. The special case of $\lambda_{hi} = W_h^0/n_h$ was used, which produces a design-unbiased estimator under the simulation study sampling plan. More extensive results from this simulation are reported in Valliant (1992).

Point variance estimates were obtained by the linearization method and were described in detail in Valliant (1991). It should be emphasized that the results here do not depend on the use of any particular method of point variance estimation. Estimates obtained by balanced repeated replication, the jackknife, or another approach would work just as well as long as consistent or approximately unbiased variance estimates were used. For each sample the linearization variance estimate was computed for each of the long-term and short-term index estimates and time periods named above. Two *GVF*'s – super smoother and loess – were then computed for each index series. For example, for the product long-term index estimate, a series of 42 point variance estimates was produced for each sample. The super smoother and loess estimates were calculated in each sample by applying those methods to the series of 42 linearization estimates for each index estimate. The simulation calculations were performed in double precision using Borland's *Turbo Pascal*. *GVF*'s were calculated with the software package *S-Plus for DOS* by Statistical Sciences Inc.

Summary statistics were then calculated across all 500 samples. The square roots of the empirical mean squared errors were computes as $\left[\sum (\hat{I}-I)^2/500\right]^{1/2}$ with the summation being over the 500 samples, $\hat{I}$ being one of the long- or short-term estimators, and I being the population index defined in section 2. Square roots of the average of the variance estimates were computed at each time period as $\sqrt{\bar{v}}$ where $\bar{v}=\sum_{s=1}^{500} v_s/500$ and v_s is one of the three types of variance estimate (linearization, super smoother, or loess) at a particular time period from sample s. This was done separately for the product estimates for long-term and 1-month price change.

Summary results across all samples and time periods are listed in Table 1. The ratios (in percent) of the square root of the average variance estimate to the root of the empirical mean squared error (RMSE) are generally somewhat less than 100 in all cases, *i.e.* both the point variance estimate ($\hat{v}$) and the *GVF*'s are underestimates, but the problem is minor. In all cases, the *GVF*'s are more stable than $\hat{v}$. For example, in Table 1, the standard deviation of the super smoother *GVF* is 61% of that of $\hat{v}$ for 1-month change when n=100. For the same case the loess GVF has a standard deviation which is 57% of that of $\hat{v}$. The biggest gains in stability are for 1-month price change while the smallest gains occur for long-term change. The loess estimates are generally more precise than the super smoother estimates with the improvement compared to the linearization estimate being somewhat less for the larger sample size. Table 1 also lists empirical coverage of 95% confidence intervals across the 42 time periods. Normal approximation confidence intervals were computed in the usual way as $\hat{I}\pm 1.96\sqrt{v}$ where $\hat{I}$ is one of the long- or short-term indexes and v is one of the variance estimates. Although all variance estimates provide slightly less than the nominal 95% coverage, the smallest percentage in Table 1 is 92.0%, and the *GVF*'s are quite competitive with $\hat{v}$.

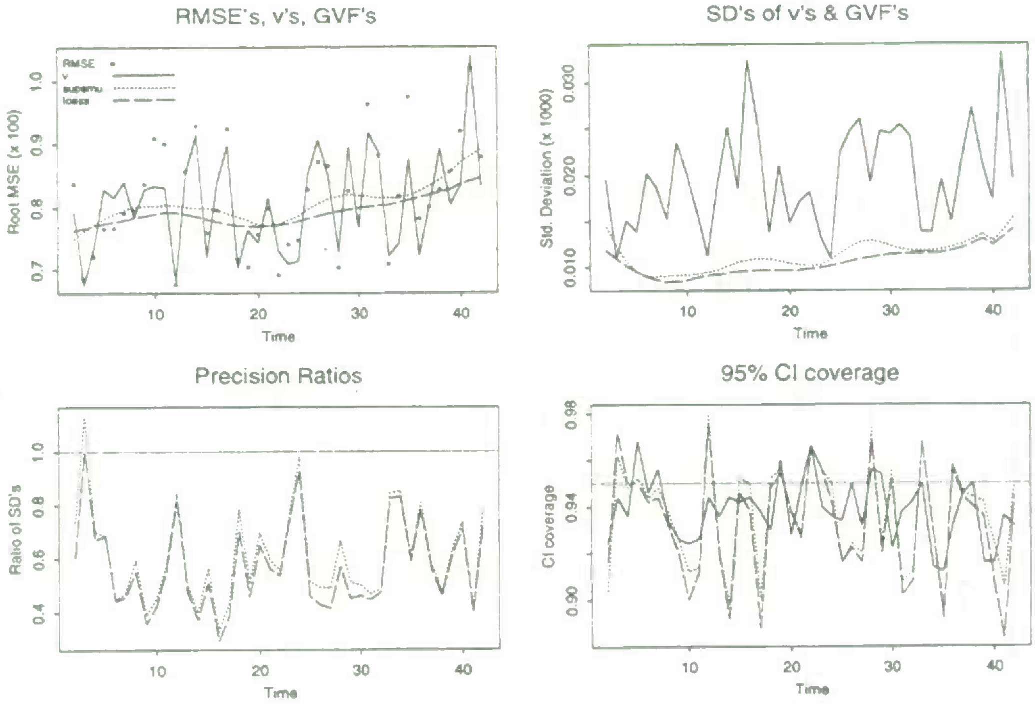
Figure 1 contains plots of summary statistics over the 500 samples by time period for n=100. Plots are given only for the 1-month product estimator. The upper left panel in each figure plots empirical RMSE's and the square root of the average of each *GVF* versus time. The *GVF*'s are much smoother than $\hat{v}$, as might be expected. Although neither smoother is inordinately influenced by outliers among the $\hat{v}$'s, the super smoother does follow the fluctuations of the $\hat{v}$ curves more closely than does loess. This leads to the super smoother's generally having a larger standard deviation than loess, as shown in the upper right-hand panel of each figure. The lower left-hand panel shows the ratio of the *GVF* standard deviation over the 500 samples to the standard deviation of $\hat{v}$. This again simply illustrates that the two *GVF* are more precise than the linearization estimate with gains being especially large for 1-month change. The lower right-hand panel of each figure charts the coverage of 95% confidence intervals over time. The *GVF*'s give reasonably good coverage which is almost equal to that of the point variance estimates. The time periods where the *GVF*'s provide noticeably poorer coverage than $\hat{v}$ are ones where the smoothers do not closely follow upward fluctuations in $\hat{v}$.

The modelling of time series is an extensively studied area, and there are many other choices for time series smoothers that could perform just as well as ones studied here. A recent discussion is found in Kohn, Ansley, and Wong (1992). A further possibility, which we have not pursued, would be to calculate a weighted average of a smoothed variance and the point variance estimate at each time point. This could be advantageous when the point variance estimators are felt to be more nearly unbiased than the smoothed estimates because of failure of the approximate variance (6) to be s smooth function of time.

Table 1: Simulation results for the product estimator from 500 two-stage cluster samples averaged over 42 time periods. All figures are in percent. $\hat{v}$ denotes the linearization variance estimate.

	$\sqrt{\hat{v}}$/RMSE			Std. dev. (*GVF*) /Std. dev. ($\hat{v}$)		95% CI coverage		
	$\hat{v}$	Supsmu	Loess	Supsmu	Loess	$\hat{v}$	Supsmu	Loess
n=50								
LT	99.5	99.0	98.0	90.8	86.8	93.7	93.1	92.8
1-month	99.3	99.4	96.8	56.8	50.8	93.3	93.6	92.9
12-month	95.0	94.8	93.9	74.3	71.0	92.0	92.3	92.1
n=100								
LT	99.8	99.2	98.5	96.1	92.1	94.2	93.4	93.3
1-month	99.3	99.8	98.1	61.0	57.0	93.8	93.6	93.3
12-month	96.1	95.8	96.0	79.1	79.6	93.2	93.4	93.3

Figure 1: Summary plots of simulation results for the 1-month product estimator from 500 samples of size *n*=100 establishments. Legend for each of the panels is in the upper left-hand panel. *v* denotes the linearization estimator; supsmu denotes the super smoother.



6. CONCLUSION

In continuing surveys which produce time series of estimates, the methods studied here for smoothing variance estimates appear to be quite useful. For continuing surveys in which the sample design and sample size are the same for long period of time, users expect variances to be smooth over time, a feature which point variance estimates generally do not have. Such expectations by users may seem, at first, to be statistically unreasonable since actual mean square errors may vary over time. However, for price indexes we have shown, using large-sample theory and simulations, that smoothed, approximately unbiased variance estimates can be obtained which are more stable than point variance estimates for both long-term and short-term price change, and which also provide near nominal confidence interval coverage. Thus, in the situation studied here, smoothed variances have a statistical justification, in addition to having considerable cosmetic appeal to data users. Consequently, smoothed variances are worth considering both for internal analysis and for publication.

ACKNOWLEDGEMENT

Any opinions expressed are those of the author and do not constitute policy of the Bureau of Labor Statistics.

REFERENCES

Cantwell, P. (1990). Variance formulae for composite estimators in rotation designs. *Survey Methodology* 16, 153-163.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., and Terpenning, I. (1990). STL: A seasonal decomposition procedure based on loess. *Journal of Official Statistics*, 6, 3-32.

Friedman, J.H. (1984). A variable span smoother. Technical Report No. 5. Laboratory for Computational Statistics, Stanford University.

Hanson, R.H. (1978). *The current population survey: Design and methodology*. Technical Paper 40, Washington DC, U.S. Bureau of the Census.

Johnson, E.G., and King, B.F. (1987). Generalized variance functions for a complex sample survey. *Journal of Official Statistics*, 3, 235-250.

Kohn, R., Ansley, C., and Wong, C.-M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, 79, 335-346.

Leaver, S.G. (1990). Estimating variances for the U.S. consumer price index for 1978-1986. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 290-295.

Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.

Valliant, R. (1991). Variance estimation for price indexes from a two-stage sample with rotating panels. *Journal of Business and Economic Statistics*, 9, 409-422.

Valliant, R. (1992). Smoothing variance estimates for price indexes over time. *Journal of Official Statistics*, 9, in press.

Valliant, R., and Miller, S.M. (1989). A class of multiplicative models for Laspeyres price indexes. *Journal of Business and Economic Statistics*, 7, 387-394.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

NEW DEVELOPMENTS IN COMPOSITE ESTIMATION FOR THE CURRENT POPULATION SURVEY

P.J. Cantwell and L.R. Ernst[1]

ABSTRACT

In January, 1994, the U.S. Bureau of the Census will incorporate several procedural changes in the Current Population Survey. These changes may affect the measurement of level for labor force characteristics, or the bias patterns associated with the eight rotation groups. We are investigating whether changes to the current composite estimator are desired or necessary. Several key issues which affect the estimator are addressed: (1) Which estimator is "best" for January 1994 and the following months? (2) Should we change the rotation design at a later date? (3) What form might the composite estimator take in the long run? By considering appropriate error measures, we present recommendations for the first two issues. For the third, a method is briefly discussed where separate composite estimators are produced for different characteristics, while retaining consistency of subgroups across totals.

KEY WORDS: Labor force status; Rotation design; AK composite estimator; Month-in-sample bias; Demographic controls.

1. INTRODUCTION

1.1 Changes in CPS and Scope of This Paper

The Current Population Survey (CPS), sponsored by the Bureau of Labor Statistics and conducted by the Census Bureau, measures labor force status in the United States. Presently, about 20% of interviews are completed at one of two centralized computer assisted telephone interviewing (CATI) facilities. This portion is expected to increase in the 1990s. The remaining 80% are conducted with a paper-and-pencil instrument in the field regions.

In January, 1994, the CPS is planning to eliminate all paper-and-pencil questionnaires in favor of computer assisted personal interviewing (CAPI) on laptop computers using a new questionnaire. The revised questionnaire will measure several new characteristics, as well as redefining or expanding established ones. The change in procedures may effect measurements of labor force status in various ways. To help prepare for this event, the Census Bureau began phasing in an experimental set of households, called the CATI/CAPI Overlap (CCO) Panel, in January, 1992. This panel will operate through December, 1993.

Except for necessary modifications, the CCO's instrument is the same as will be used in regular CPS in January, 1994, and subsequently. Further, when the CCO Panel is fully phased in, the 15,000 sample households (per month) will be interviewed in the same rotation pattern as CPS. It should be noted that the CCO Panel, during its duration, is completely separate from regular CPS; none of the data from the CCO are used in estimates or publications, nor will any respondents continue in sample in 1994.

Because these one-time changes will occur in January 1994, corresponding changes to the current composite estimator are being investigated. We provide information on the current CPS rotation design and composite estimation in Sections 1.2 and 1.3, respectively. In Section 2, the best estimator for January, 1994 and subsequent months is sought. We consider variance and month-in-sample bias when estimating monthly level and month-to-

[1] P. Cantwell, Principal Researcher, and L. Ernst, Assistant Division Chief, Statistical Research Division, Bureau of the Census, Washington, DC, U.S.A. 20233.

month change under several options: (i) continuing with the usual "AK composite" estimator, (ii) not compositing in January, 1994, and (iii) compositing with the CCO Panel.

Looking beyond 1994, Section 3 explores alternative rotation plans--those where households are interviewed for six or eight consecutive months. In Section 4, more comprehensive changes to the estimator are considered. We review other forms of composite estimators, and describe a procedure (suggested by Fuller) which develops separate composite estimators for different characteristics, and uses these labor force controls to ensure consistency of totals across subgroups.

1.2 The Current Rotation Design in CPS

In the CPS, sample households are interviewed for four consecutive months, rotated out of sample for the next eight months, and finally returned to the sample for four more months. For more information, on this 4-8-4 design or the CPS in general, see U.S. Bureau of the Census, Technical Paper No. 40 (1978). By comparison, participants in the Labour Force Survey (LFS) conducted by Statistics Canada are contacted for six consecutive months and then retired.

There are several cost advantages of repeated sampling. Some overhead costs are encountered only once per household; telephone interviews--much cheaper than personal visits--might be more appropriate after the first contact. However, our focus is strictly on bias and variance effects. The rotation overlap in the CPS and the LFS--75% and 83% respectively--reduces the variance of change estimates.

1.3 Composite Estimation

To take greater advantage of the household overlap in consecutive months, the CPS uses a composite estimator. For a specified characteristic, let $x_{h,i}$ be the estimate of total for month *h* arising from the rotation group which is interviewed for the *i*th time that month. In each month, there are eight such estimates. A simple ratio estimate takes the form $Y_h = (1/8)\ \Sigma\ x_{h,i}$. (This is a ratio estimator because the weight attached to each respondent is derived after several stages of adjustment.)

About twenty years ago, the CPS started using a "simple composite estimator." Let us define $\Delta = (1/6)\ \{\ \Sigma\ x_{h,i} - \ \Sigma\ x_{h-1,i-1}\ \}$, where the sums are over rotation groups common to months *h* and *h-1*, to estimate the difference between the months. An alternative to the ratio estimate of total for month *h* (called a "change estimate" here) would take the previous month's estimator and add to it the difference estimate Δ. Then the simple composite estimator for month *h* is defined as $Y_h' = (1\text{-K})\ Y_h + \text{K}\ (Y_{h-1}' + \Delta)$, a linear combination of the simple ratio estimate and the change estimate, both for month *h*. CPS formerly used this estimate with K = .5.

Further refining the composite estimate, while adding only one parameter, is the AK composite estimator, developed by Gurney and Daly (1965). Let $\beta = (1/8)\ \{\ \Sigma\ x_{h,i} - (1/3)\ \Sigma\ x_{h,j}\ \}$, where $i = 1{,}5$, and $j = 2{,}3{,}4;\ 6{,}7{,}8$. The AK composite estimator is defined as $Y_h'' = (1\text{-K})\ Y_h + \text{K}\ (Y_{h-1}'' + \Delta) + \text{A}\ \beta$. Introducing the A parameter allows the rotation group coefficients to more closely resemble those of minimum variance linear estimators, and to lessen the effects of month-in-sample bias (described in Section 2.1). CPS converted to this estimator with A = .2 and K = .4 in the mid-1980s.

For additional developments in composite estimation, see Gurney and Daly (1965), Wolter (1979), Kumar and Lee (1983), Breau and Ernst (1983), Adam and Fuller (1992), or Yansaneh and Fuller (1992).

2. WHICH ESTIMATOR IN JANUARY 1994?

2.1 Changes and Other Considerations in 1994

Section 1.1 described several one-time changes in the CPS to be implemented in January, 1994. Long-term modifications to the estimators would take place at a later date, perhaps early 1996. During the switch to the new questionnaire, we hope to disrupt the time series of estimators as little as possible, bearing in mind that there may be a "jump" in January, 1994, due to the new procedures.

Three estimators were investigated in this section. They are derived as follows:

(E1) Composite as usual in 1/94 and in succeeding months, with A = .2 and K = .4.

(E2) Start with a ratio estimate (a simple average of the estimates from the eight rotation groups) in 1/94; composite 2/94 with 1/94 (A = .2, K = .4); composite 3/94 with 2/94 and 1/94; *etc.*

(E3) Composite 1/94 with the 12/93 composite estimator obtained from the CCO Panel, with A = .2 and K = .4; composite 2/94 with 1/94 (A = .2, K = .4); composite 3/94 with 2/94 and 1/94; *etc.* To estimate month-to-month change in 1/94, subtract the 12/93 CCO estimate from the 1/94 estimate.

When evaluating estimators for 1994, several factors may have an effect. First, the old and new questionnaires are not completely compatible. That is, some questions will be introduced in the new instrument, and the choice of responses will be changed in others. This question of compatibility is important if we use a composite estimator which combines data from 1993 with 1994, and when we consider month-to-month change in January 1994.

Second, it is possible that the new procedures will introduce a one-time "jump" in the level of labor force characteristics, especially unemployed (UE), in January, 1994. Any estimator, including the simple ratio estimator, may be biased. However, even if there is no real change in level from December 1993 to January 1994, the application of a new mode may introduce an additional shift in the expected value of the ratio estimator or others.

Third, there is evidence to imply that, with the new procedures, the patterns of month-in-sample bias will change. Bailar (1975) defines and discusses the concept of month-in-sample effects caused by panel conditioning in the CPS. Briefly, for any given month and characteristic to be estimated, the expected values of the seven rotation group estimates are generally not equal, but reflect the number of previous interviews or other influences. Using the notation of the last section, the bias *index* for the *i*th month in sample can be defined as $E(x_{k,i}) / E(\Sigma x_{k,j}/8)$, so that an index greater than 1 implies an overestimate in that month relative to the other seven months. In our study, the simple ratio estimator is assumed to be unbiased, to create a reference point for comparison.

Recent analysis by Adams (1991) on CPS responses from 1980 to 1987 provides estimates of the bias indices for UE and civilian labor force (CLF) under the current procedures. Yet computations from the CATI phase-in study (Shoemaker 1992) yield significantly different indices for UE. (We have not yet obtained CATI indices for CLF.) Because all interviewing will be computer assisted in 1994, the new indices may well be closer to those obtained under CATI.

Further, in the new instrument, certain questions will be asked of discouraged workers in each month. It is thought that these questions tend to increase the level of UE slightly. Before 1970 these questions were asked only in months-in-sample 1 and 5. Since 1970, they have been asked in months 4 and 8. Bailar (1975) provides estimates of the bias indices before and after 1970. This allows us to estimate the effect of the questions on month-in-sample bias, and to project what the indices (labeled "DWQ indices," for discouraged worker questions) might be when the questions are asked in each month.

Table 1 displays for the eight months in sample the current, DWQ, and CATI bias indices for UE, and the current and DWQ indices for CLF.

Table 1: Bias Indices Used in Study of Competing Estimators.

Month in Sample	1	2	3	4	5	6	7	8
Unemployed								
Current Indices	1.070	1.004	.988	1.002	1.004	.956	.965	1.012
DWQ Indices	1.125	1.030	.993	.949	1.037	.987	.954	.925
CATI Indices	.98	1.07	1.08	.98	.94	1.03	.96	.96
Civilian Labor Force								
Current Indices	1.016	1.002	.996	1.002	.998	.992	.993	1.000
DWQ Indices	1.017	1.004	.999	.997	1.002	.994	.994	.993

Note that the introduction of the new procedures will produce a one-time mix of bias effects for January 1994 respondents. For example, one rotation group will be interviewed in January for the fourth time, but for the first time with the new instrument. Indeed, there may be other unanticipated effects on bias, which we cannot hope to measure beforehand.

Fourth, as usual, 75% of sample households from the regular CPS in December 1993 will continue in sample in January 1994. An estimator such as E1 would take advantage of the correlated estimates from common rotation groups. On the other hand, the experimental CATI/CAPI Overlap (CCO) Panel, operating through December 1993, will have no households in sample in 1994.

Finally, there is the potential change in correlation patterns of common rotation groups. If $x_{h,i}$ and $x_{h-r,\ j}$ represent estimates from the same rotation group *r* months apart, the values are correlated as a function of *r*. When measuring UE, previous studies (Breau and Ernst 1983, Adam and Fuller 1992) yield correlations of about .50 when *r* is 1, decreasing to about .20 when *r* is 15. Corresponding values for CLF are considerably larger, dropping from about .80 to .55. In our study, we used correlations slightly smoothed from those obtained in the given references.

Because the new instrument relies more heavily on dependent interviewing, some of the correlations may increase in 1994. Still, we assumed that correlations among months in 1994 would remain the same with the new questionnaire. However, correlations for a common rotation group *between an estimate measured in 1993 and a second estimate measured in 1994* will likely decrease--if only slightly--because of the different modes. In our analysis, a 10% decrease in the usual correlations was assumed in such cases.

2.2 Qualities of the Estimators

To make a decision for January 1994, managers are considering three aspects of the competing estimators: operational circumstances; resulting variances; and biases. For the last item, we actually measured in our work the deviation from the expected value of the steady-state composite estimator (*i.e.*, the state of the usual composite estimator reached after new effects are fully phased in). As we have indicated, there may be a one-time shift in the value of labor force characteristics due to the new procedures. Because the actual level will never be known, our hope is to suffer this jump at once (January), and then proceed rapidly to the "steady-state." That is, we want an estimator to reach the bias level of the long-run composite estimator as quickly as possible.

Before displaying computational results, let us briefly examine the three estimators, defined in Section 2.1. E1 continues to composite as usual, combining data from 1993 with January 1994 and later months. As it takes advantages of the correlations among common rotation groups, we expect it to yield the lowest variances in most cases. Nevertheless, combining data from different questionnaires can be operationally complex. In addition, there would be different month-in-sample bias effects--due to the different procedures--for several months into 1994. This could lead to an estimator of month-to-month change whose expectation is nonzero.

The second estimator, E2, is operationally simple. By starting 1994 with a simple ratio estimator, and subsequently compositing only with earlier months in 1994, there is a clean break with the bias effects from the old system. Still, we lose the potential variance reduction by ignoring rotation groups which were in sample in late 1993. Moreover, for all three sets of bias patterns we consider, the new bias effects will take several months to reach those of the steady-state composite estimator.

The third estimator, E3--by compositing January 1994 with the CCO Panel--uses the same instrument before and after 1/1/94. Thus, there is no problem of compatibility of questionnaires. In addition, the steady-state bias effects should be reached almost immediately. (However, see note in Section 2.2 about the one-time mix of bias effects for January respondents.) Unfortunately, E3 suffers from two serious problems: (i) the CCO Panel has about one-fourth as many households as regular CPS, and (ii) there is no overlap between the CCO in 1993 and CPS in 1994. The result is a serious increase in the relevant variances.

2.3 Results of Computations for January 1994 and Subsequent Months

To simplify the research, we started by comparing the variances of the three competing estimators. The problems with E3 were stated briefly in the last section. Indeed, its variance for January 1994 is about 35% (50%) greater than the steady-state composite when estimating monthly level of UE (CLF), and about 4 (7) times greater when estimating month-to-month change. In light of these results, we continued the investigation comparing only E1 and E2.

To evaluate E1 and E2 in the first six months of 1994, we computed their variances, and their deviations from the expected value of the steady-state composite estimator. Tables 2 and 3 present the variances and the mean squared deviations (MSD, variance plus squared deviation), each divided by the variance of the steady-state composite. Two columns portray monthly level, two more month-to-month change. For ease of comparison, in each cell, the value for E1 ("continue compositing as usual") is placed directly above that for E2 (ratio estimator in 1/94).

For all computations, the current bias indices were used for months before 1994. For months in 1994, we inserted the current (no change), DWQ, and CATI indices to monitor each situation. Only DWQ indices were used in the cases shown in Tables 2 and 3.

A specified increase in the level of the characteristic was incorporated for January 1994, representing the effect of the new CPS procedures. The cases shown assume the expected value of the ratio estimator jumps 10% for UE (Table 2) and 1% for CLF (Table 3). These "jumps"--based on preliminary data--affect biases and variances. No further changes in the levels were assumed for later months in 1994.

From the tables, it is seen that E1 realizes almost no increase in variance of monthly level compared to the steady-state composite, E2 slightly more (6% in Table 2, 20% in Table 3). In the case of CLF (Table 3), both estimators suffer a greater variance increase in month-to-month change, 23% and 30%, respectively. This follows because, where the usual correlations are high--80% for two groups one month apart--they are assumed to dip 10% from December to January. Applying other bias index patterns or shifts in level does not seriously affect the variance comparison between E1 and E2.

Table 2: Variances and Mean Squared Deviations (as Compared to the Steady-State Composite Estimator) for *UE*, using *DWQ* Biases, *10%* Shift in Level; Values for E1 Placed Above Those for E2.

Month	Monthly Level		Month-to-Month Change	
	Variance	Mean Squared Deviation	Variance	Mean Squared Deviation
1/94	1.0154 1.0673	2.0700 2.8724	1.0283 1.0087	22.30 24.99
2/94	1.0044 1.0246	1.1731 1.3134	1.0022 1.0028	1.3097 1.5292
3/94	1.0011 1.0087	1.0281 1.0549	1.0008 1.0014	1.0500 1.0856
4/94	1.0002 1.0025	1.0045 1.0099	1.0003 1.0012	1.0082 1.0147
5/94	1.0000 1.0004	1.0007 1.0016	1.0001 1.0007	1.0013 1.0029
6/94	1.0000 1.0000	1.0001 1.0002	1.0000 1.0001	1.0002 1.0005

Table 3: Variances and Mean Squared Deviations (as Compared to the Steady-State Composite Estimator) for *CLF*, using *DWQ* Biases, *1%* Shift in Level; Values for E1 Placed Above Those for E2.

Month	Monthly Level		Month-to-Month Change	
	Variance	Mean Squared Deviation	Variance	Mean Squared Deviation
1/94	1.0289 1.1966	1.5145 2.4065	1.2259 1.2962	35.06 40.85
2/94	1.0080 1.0722	1.0857 1.2657	1.0073 1.0133	1.2477 1.6122
3/94	1.0020 1.0234	1.0144 1.0543	1.0026 1.0114	1.0411 1.1072
4/94	1.0003 1.0063	1.0023 1.0113	1.0010 1.0062	1.0071 1.0215
5/94	1.0000 1.0010	1.0004 1.0018	1.0002 1.0032	1.0011 1.0057
6/94	1.0000 1.0001	1.0001 1.0002	1.0000 1.0006	1.0002 1.0010

The deviations from the steady-state composite are quite large in January. For monthly level, this reflects the change to DWQ biases in these examples. For month-to-month change, there is the added effect of the jump in level--10% in Table 2, 1% in Table 3. (The steady-state change from month to month is assumed to be 0.) The monthly-level deviations are much smaller if CATI biases are used in 1994 (for UE), and almost 0 if the biases do not change in 1994. The month-to-month change deviations for 1/94 are only slightly smaller under the latter bias patterns, because of the strong influence of the shift in level.

From our computations with other sets of parameters, several trends are worth noting. Although the variances and MSDs for E2 are generally larger than those for E1, the difference is usually not serious. Also, regardless of the size of the variance or MSD in January, the values in February are typically much closer to the steady-state composite. By March or April, the values are generally only 1% or 2% away. If we speculate that there will likely be a more serious increase in other nonsampling errors through this transitional period, the increases due to the estimators may be minor by comparison.

With these observations in mind, the decision between these two estimators will probably be based mostly on operational or processing concerns. If so, the current leaning is more toward E2, starting with a simple ratio estimate in January, and resuming a composite in February. Even here, the A and K coefficients might be manipulated to more quickly bring the variances or deviations in line with the steady state.

3. SHOULD WE CHANGE THE ROTATION DESIGN?

3.1 Interviewing in 6 or 8 Consecutive Months

The current 4-8-4 rotation design, as described in Section 1.2, produces reductions in variance of change from month to month and year to year. This occurs because about 75% of the households in any month are interviewed again the following month, and 50% again one year later. What type of reductions can be obtained by interviewing households in, say, six or eight consecutive months? While speaking at the Census Bureau recently, Wayne Fuller discussed some results he had obtained with Adam and Yansaneh investigating these alternative rotation plans. Although the overlap from month to month would increase to 83% or 87%, respectively, there would be no overlap in months one year apart. Fuller's comparisons assume a minimum variance linear estimator is applied to the data.

We compared the variance of AK composite estimators, varying several parameters in the analysis. Three rotation designs were studied, 4-8-4, the 6-consecutive month plan, and the 8-consecutive month plan. Under each design, we computed the variance of estimators for monthly level, month-to-month change, and annual average. For the characteristics studied--the number of people unemployed (UE) and the number in the civilian labor force (CLF)--we used the same sets of correlations as in Section 2. Finally, although we did not investigate general linear estimators in this analysis, we computed variances for AK composite estimators, allowing A and K to assume all values 0, .1, .2, ..., .9.

3.2 Results for Three Rotation Designs

We computed the change in variance realized for AK composite estimators under a 6- or 8-consecutive month rotation design, *comparing in each case to the current 4-8-4 scheme with A = .2 and K = .4.* The consecutive-month plans have minor effects on the variance of monthly level. For UE, the variance increases between 0% and 10% for most A,K pairs; for CLF, the alternative plans can reduce the variance as much as 11% by using values of A and K around .7 or .8.

More notable are the variance reductions for month-to-month change when the correlations are higher. When estimating UE, the 6- and 8-month designs lower the variance from 3% to 7% for most A,K pairs. Because of the higher rotation groups correlation patterns estimating CLF, this reduction grows to 23% (26%) for the 6- (8-)month design when K = .8 (but A is low, .1 or .2).

We should mention here that comparing optimal results for the 6- and 8-month designs is somewhat unfair to the 4-8-4 rotation design. After all, the currently used values, A = .2 and K = .4, are a bit of a compromise; they work well for UE, not badly for CLF. (See discussion in Section 4.1.) Estimating CLF while retaining the 4-8-4 design, we could reduce the variances of monthly level and month-to-month change by as much as 12% and 17%, respectively, by selecting other A,K pairs.

Unfortunately, interviewing only in consecutive months increases the variances of annual average substantially. Even selecting the optimal A,K pair, the variance for UE (CLF) increases 14% (18%) under the 6-month design and 28% (36%) under the 8-month design.

Although estimates of monthly level and month-to-month change are typically considered more important than those of annual average, the latter are important to the 40 smaller states (including the District of Columbia). Whereas monthly labor force estimates are produced for the 11 largest states, only annual estimates are required for the remaining 40 due to their small monthly sample size. Depending on the correlations, the consecutive-month designs have a smaller effective sample size than 4-8-4 for measuring average over a 12-month period, taking into consideration the greater overlap of households. As computations confirm, this produces higher variances.

Because the smaller states' coefficients of variation for annual average are larger than that for the national estimates on a monthly basis, managers at the Bureaus of Labor Statistics and the Census felt that interviewing in six or eight consecutive months would seriously harm the smaller states' estimates. It was decided that the current 4-8-4 rotation design would be retained for the 1990s.

4. LONG-TERM ISSUES IN COMPOSITE ESTIMATION

4.1 Choice of Composite Coefficients

Before changing the system of composite estimation, data from 1994 will be studied to determine the pattern of biases and other effects resulting from the new procedures. Different A,K pairs or general coefficients are optimal under various sets of rotation group biases and correlations. We hope to implement any changes around 1996.

The coefficients currently used in the CPS estimator, K = .4 and A = .2, represent a bit of a concession. Although they are close to optimal among AK estimates for measuring UE, characteristics with higher correlations, such as CLF and the number employed, would realize greater reductions in variance with values of K closer to .7 or .8, and A selected correspondingly larger. The number of unemployed, however, is often deemed to be the more important characteristic. Other factors can influence the choice of K and A as well. While higher K values generally reduce variance when estimating month-to-month change, they tend to increase the variance when estimating annual average--important to the smaller states (as mentioned in Section 3.2).

Even the AK composite estimator itself is a compromise in terms of data storage. A simple ratio estimate uses information only from the current month. At the other extreme is the minimum variance linear estimator, requiring us to save rotation group values from many previous months. Although the AK estimator uses data from many months in the past, that information is summarized in the previous month's composite estimator. Only rotation group estimates from this month and last, plus last month's AK composite estimator, need to be saved to produce the new estimator.

The approach of using a general linear estimator is attractive in that, subject to the required storage of past data, we can seek a minimum variance estimator. Further, optimal coefficients for estimating the change from last month are obtained by subtracting optimal coefficients from the current and previous months, using all available months for each estimate. However, as others have pointed out, it would be necessary to revise the labor force estimates of the previous month. The Bureau of Labor Statistics has looked unfavorably on this practice, being hard to explain to data users.

Without expanding data storage needs, we can still improve on the variance of the current AK estimator. Breau and Ernst (1983) investigate *generalized composite* coefficients in an estimator of the form $Y_k = \Sigma a_i x_{k,i} - K \Sigma b_i x_{k,i} + K Y_{k-1}$. In the authors' summary, they note that this estimator can effect greater variance reductions in measuring annual average, smaller reductions in measuring monthly level and month-to-month change. Further, it achieves much of the reduction realized by the minimum variance linear estimator.

However, there is a trade-off. As we move from AK composite to generalized composite to minimum variance linear estimators, previous research has shown that reduced variances are often accompanied by greater month-in-sample bias in the estimates. This has created a reluctance to institute these alternatives to the AK composite. After we measure the bias effects of the new procedures in 1994, there may be further incentive to change the form of the estimator. Because of our distinct interest in limiting the variance of annual average, we are continuing to explore the generalized composite and minimum variance linear estimators.

4.2 Different Coefficients for Different Characteristics

Another option to improve labor force estimates, discussed by Wayne Fuller at the Census Bureau, is to apply different coefficients for the several key characteristics. This idea has been rejected in the past because of problems with data consistency. Current policy dictates that we produce estimates of total which are consistent with their components as well as across time. For example, the number of employed and unemployed people must add to the number in the civilian labor force; an estimate of annual average must equal the average of its components.

Fuller suggested a new approach where different A,K pairs or coefficients would produce composite estimate totals for some or all of the chief labor force characteristics. These totals would then be used as another set of controls, similar to age-race-sex categories, in the weighting adjustment. When all adjustments are completed, the person weights will sum to the composite-estimated labor force totals as well as to the proper demographic totals.

The procedure might work as follows. Apply all stages of adjustment (nonresponse, demographic controls, *etc.*) Using the optimal A,K pair for UE, obtain the composite estimator for UE. Using the optimal A,K pair for the number of employed people (EMP), obtain its composite estimator. Add these estimates to obtain the estimate for CLF. Subtract the result from population controls to obtain the number of people aged 16+ not in the labor force. These labor force totals now function as controls, while retaining consistency across the labor force estimates. At this point, we would repeat the weighting adjustment to demographic controls, which may be off due to the compositing.

Several appealing features would result. The weights on the data files would incorporate the effects of composite estimation. This would allow CPS data users to reproduce final estimates before the seasonal adjustment is applied, something not currently done. In addition, many fewer data would have to be stored.

Several issues arise. First, for which two labor force characteristics should we apply composite estimation (UE and EMP?), and which two should be obtained by deduction? Should this issue be decided merely on the importance of the characteristics, or should statistical implications be considered?

Second, at what level should we apply the controls which arise from labor force estimates: national totals, the margins of demographic (age-race-sex) categories, or the individual (cross-classified) demographic cells? As we move toward compositing in smaller cells, we simplify the raking procedure, while increasing the variability of the weights and the risk of obtaining negative cell frequencies.

Finally, the composite coefficients would be selected to improve national estimates of labor force. How will this procedure affect (1) estimates for subgroups *below the level of compositing* (such as the number of employed Blacks, if we do not composite by race), (2) labor force estimates at the state level, and (3) items other than labor force? These and other questions are being investigated before we can recommend significant changes in CPS composite estimation.

ACKNOWLEDGMENTS

We thank Wayne Fuller for motivating parts of this research, and for his helpful discussions.

REFERENCES

Adam, A., and Fuller, W. (1992). Covariances of estimators for the current population survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, in press.

Adams, D. (1991). Memorandum for documentation, CPS Month-In-Sample (MIS) bias index research, 10/21/91. U.S. Bureau of the Census, Washington, DC.

Bailar, B. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Breau, P., and Ernst, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 397-402.

Gurney, M., and Daly, J. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 242-257.

Kumar, S., and Lee, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 403-408.

Shoemaker, H. (1992). Memorandum for documentation. CATI phase-in analysis: A look at month-in-sample bias indexes for unemployed (CC_ALYS-8), 10/30/92, U.S. Bureau of the Census, Washington, DC.

U.S. Bureau of the Census (1978). The current population survey: Design and methodology. Technical paper No. 40, Washington, DC: U.S. Government Printing Office (Department of Commerce).

Wolter, K. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Yansaneh, I., and Fuller, W. (1992). Alternative estimators for the current population survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, in press.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

SESSION 5

Longitudinal Studies in Health Research

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

ANALYSIS OF BINARY LONGITUDINAL DATA

G.A. Darlington[1]

ABSTRACT

When binary responses are observed over time, the dependence between observations for an individual must be considered. Two specific modelling schemes for binary longitudinal data are discussed. Each approach is appropriate when a primary focus is to identify the relationship between the marginal probability of an event and a set of explanatory variables and when there is also specific interest in studying the dependence of the correlational structure on explanatory variables. Examples are presented to illustrate the modelling techniques.

KEY WORDS: Binary longitudinal data; Explanatory variables; HIV infection.

1. INTRODUCTION

Logistic regression is commonly used to model the relationship between independent binary responses and a set of covariates (Cox 1970). When the binary responses correspond to observations on individuals over time, the dependence between observations for an individual must be considered. Thus, the logistic regression approach must be modified to allow for correlated responses.

Important initial work on the analysis of binary longitudinal data was performed by Zeger, Liang and Self (1985) where working likelihoods and robust variance estimates were utilized. In this paper, the approach of Zeger *et al.* (1985) will be used to explore models that are designed for situations where a primary focus is on the marginal probability of success but where there is also specific interest in studying the dependence of the correlational structure on explanatory variables. The incorporation of parameters to account for dependence among observations for an individual will be discussed in Sections 2 and 3. Numerical examples related to the AIDS epidemic are discussed in Section 4.

2. MODELLING AUTOCORRELATION AS A FUNCTION OF COVARIATES

Let Y_{it} represent a binary response observed for individual i at time t, $i = 1, ..., K$, $t = 1, ..., n_i$. Let z_i represent the $s \times 1$ vector of time-independent explanatory variables for individual i and let $\pi_i = Pr(Y_{1t} = 1|z_i)$. Zeger *et al.* (1985) assume, first, that logit $(\pi_i) = \beta' z_i$ where β is an $s \times 1$ vector of parameters. They also assume that corr $(Y_{it}, Y_{it-1}|z_i) = \rho$. That is, they assume that the first lag autocorrelation is constant.

For many longitudinal studies it is expected that ρ will lie in the range of $0 < \rho < 1$. This restriction will allow a simple logistic representation of the first lag autocorrelation with dependence on covariates. Assume that

$$\text{logit}\,(\rho_i) = \tau' z_i ,$$

where $\rho_i = \text{corr}(Y_{it}, Y_{it-1}|z_i)$ and τ is an $s \times 1$ vector of parameters. If Markov dependence is assumed then the likelihood function is given by

[1] G.A. Darlington, Division of Epidemiology and Statistics, Ontario Cancer Treatment and Research Foundation, 620 University Avenue, Toronto, Ontario, Canada, M5G 2L7.

$$L(\beta,\tau) = \prod_{i=1}^{K} \left[\pi_i^{y_{i1}} (1-\pi_i)^{1-y_{i1}} \prod_{t=2}^{n_i} \pi_{it}^{y_{it}} (1-\pi_{it})^{1-y_{it}} \right], \tag{2.1}$$

where

$$\pi_i = Pr(Y_{it} = 1 | z_i) = \frac{e^{\beta' z_i}}{1 + e^{\beta' z_i}},$$

$$\pi_{it} = Pr(Y_{it} = 1 | Y_{it-1}, z_i) = E(Y_{it} | Y_{it-1}, z_i)$$

$$= \pi_i + \rho_i (Y_{it-1} - \pi_i)$$

and

$$\rho_i = \frac{e^{\tau' z_i}}{1 + e^{\tau' z_i}}.$$

Note that the likelihood as specified in (2.1) does not depend on the assumed model. Note also from (2.1) that the unrestricted range for ρ_i is

$$\max \left\{ -\frac{\pi_i}{1-\pi_i}, -\frac{1-\pi_i}{\pi_i} \right\} < \rho_i < 1.$$

Recall that the likelihood (2.1) is based on a Markov assumption. With the possibility of more general dependence, following Zeger *et al.* (1985), this likelihood is referred to as the working likelihood since it results from working with a Markov assumption. The working Markov log likelihood is therefore given by

$$l(\beta,\tau) = \sum_{i=1}^{K} l_i$$

$$= \sum_{i=1}^{K} [y_{i1} \beta' z_i - \log \{1 + \exp(\beta' z_i)\} \tag{2.2}$$

$$+ \sum_{t=2}^{n_i} \{y_{it} \log \pi_{it} + (1 - y_{it}) \log (1 - \pi_{it})\}].$$

Parameter estimates $\hat{\beta}, \hat{\tau}$ can be obtained by maximizing this log likelihood. An iterative procedure is necessary to compute these estimates.

A test for constant first lag autocorrelation corresponds to testing whether all components of τ, except for the constant term, are equal to zero. As in Zeger *et al.* (1985), the estimates $\hat{\beta}, \hat{\tau}$ are, under certain regularity conditions, consistent for β, τ and asymptotically normal under more general conditions, since

$$\sum_{i=1}^{K} \frac{\partial l_i}{\partial \theta} = 0,$$

are unbiased estimating equations, where $\theta = (\beta', \tau')'$.

Thus, under certain regularity conditions (Inagaki 1973), given that Y_{it} is a stationary binary time series such that

$$\text{logit}[Pr(Y_{it} = 1 | z_i)] = \beta' z_i$$

and

$$\text{logit}[\text{corr}(Y_{it},\ Y_{it-1}|z_i)] = \tau' z_i,$$

$i = 1, ..., K$, $t = 1, ..., n_i < \infty$, the results of Inagaki (1973) indicate that if $\theta = (\beta', \tau')'$ and $\hat{\theta} = (\hat{\beta}', \hat{\tau}')'$, then $\sqrt{K}(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and covariance matrix $W^{-1} V W^{-1}$ where

$$K^{-1} \sum_{i=1}^{K} E_T \left[\frac{\partial^2 l_i}{\partial \theta \partial \theta'} | z_i \right] \to W, \tag{2.3}$$

$$K^{-1} \sum_{i=1}^{K} E_T \left[\left[\frac{\partial l_i}{\partial \theta} \right] \left[\frac{\partial l_i}{\partial \theta} \right]' | z_i \right] \to V, \tag{2.4}$$

as $K \to \infty$ and $E_T(\cdot)$ represents expectation with respect to the true underlying distribution.

Robust estimates of the standard errors of the parameter estimates can be obtained from the covariance matrix estimate $K^{-1} \hat{W}^{-1} \hat{V} \hat{W}^{-1}$ where

$$\hat{W}_{uv} = \frac{1}{K} \sum_{i=1}^{K} \left[\frac{\partial^2 l_i}{\partial \theta_u \partial \theta_v} \right] |_{\hat{\theta}} \tag{2.5}$$

and

$$\hat{V}_{uv} = \frac{1}{K} \sum_{i=1}^{K} \left\{ \left[\frac{\partial l_i}{\partial \theta_u} \right] \left[\frac{\partial l_i}{\partial \theta_v} \right] \right\} |_{\hat{\theta}}. \tag{2.6}$$

Note that if the Markov assumption is valid, $W^{-1} V W^{-1}$ becomes $-W^{-1}$, and therefore model based estimates of the standard errors of the parameter estimates can be obtained from $-K^{-1} \hat{W}^{-1}$.

3. MODELLING A TRANSITION PROBABILITY

The model of Zeger *et al.* (1985) defines the marginal distribution of Y_{it} as logit $(\pi_i) = \beta' z_i$. Farewell (1982) also uses this marginal specification but instead of assuming a form for the correlation, it is assumed that

$$\text{logit } \{Pr(Y_{it} = 1 | Y_{it-1} = 1, z_i)\} = \gamma' z_i,$$

where γ is an $s \times 1$ vector of parameters. Note that there is an asymmetry in the dependence structure which may not be appropriate in all applications.

For this model, the working Markov likelihood function is specified by (2.1) where again $\pi_{it} = \pi_i + \rho_i (Y_{it-1} - \pi_i)$ but

$$\rho_i = \text{corr}(Y_{it}, Y_{it-1}|z_i) = \frac{e^{\gamma' z_i} - e^{\beta' z_i}}{1 + e^{\gamma' z_i}}.$$

Because the likelihood must be maximized at values $0 < \hat{\pi}_{it} < 1$ and $0 < \hat{\pi}_i < 1$, ρ_i is restricted to lie in the range

$$\max \left[-\frac{\pi_i}{1 - \pi_i}, -\frac{1 - \pi_i}{\pi_i} \right] < \rho_i < 1.$$

The working Markov log likelihood is given by (2.2). Parameter estimates can be obtained by maximizing this function with respect to β and γ. Under the Markov assumption, standard maximum likelihood estimation results apply. When the Markov assumption is not valid, the estimates $\hat{\beta}, \hat{\gamma}$, obtained by maximizing (2.2), may still be of value since the estimating equations are unbiased. Thus, under certain regularity conditions presented by Inagaki (1973), if $\theta = (\beta', \gamma')'$ and if n_i are bounded, $i = 1, ..., K$, then $\hat{\theta}$ is consistent and $\sqrt{K}\,(\hat{\theta} - \theta)$ is asymptotically normal with mean 0 and covariance matrix $W^{-1} V W^{-1}$ where W and V are given by (2.3) and (2.4) respectively. The robust estimate of the covariance matrix is given by $K^{-1} \hat{W}^{-1} \hat{V} \hat{W}^{-1}$ where $\hat{W}$ and $\hat{V}$ are presented in (2.5) and (2.6) respectively.

4. NUMERICAL EXAMPLES

4.1 New York AIDS Study

Results from a study of risk factors for HIV infection among intravenous drug users in New York City are presented by Des Jarlais *et al.* (1987) and by Marmor *et al.* (1987). A subset of the data used to obtain these results will be considered to investigate whether intravenous drug use behaviour differs between HIV positive individuals versus HIV negative individuals.

Two hundred and twenty-five male intravenous drug users from New York City were interviewed yearly for 4 years. Monthly frequency of intravenous drug use was determined. All individuals in this data subset were either HIV positive or HIV negative throughout the 4 year period.

The methods outlined in Sections 2 and 3 were employed to investigate changes in behaviour as a result of HIV status. Thus the response is defined as $Y_{it} = 1$ if individual i at time t is a frequent intravenous drug user, $i = 1$, ..., 225, $t = 1$, ..., $n_i, n_i \leq 4$, where frequent is defined as greater than 5 injections per month. Otherwise, $Y_{it} = 0$. The time-independent covariate is an indicator of whether individual i is HIV positive $(z_i = 1)$ or HIV negative $(z_i = 0)$.

The results of fitting the model described in Section 2 are presented in Table 1. When the parameter estimates are compared with the robust or model based estimates of the standard error, it is concluded that the first lag autocorrelation is constant and the probability of frequent intravenous drug use is greater for HIV positive individuals.

Since the first lag autocorrelation is constant, the model of Zeger *et al.* (1985) is considered. The results of fitting this model are presented in Table 2 yielding the same overall conclusions as the first analysis.

Table 3 lists the results of fitting the model described in Section 3 to the data. From this table, it is concluded that the probability of frequent drug use, given previous frequent drug use, does not differ for HIV positive and HIV negative individuals. However, it is also concluded that the overall probability of frequent drug use is greater for the HIV positive individuals.

This model was also considered for the response defined as infrequent drug use instead of frequent drug use. The results of fitting this model to the infrequent drug use data are listed in Table 4. The conclusions here are that the conditional probability of infrequent drug use given previous infrequent drug use depends on HIV status where this conditional probability is lower for HIV positive individuals compared to HIV negative individuals. Also, the overall probability of infrequent drug use is lower for HIV positive individuals compared to HIV negative individuals.

Thus, it appears quite feasible to introduce a dependence on covariates into the correlational structure. The asymmetry of the second model is illustrated in the different characterizations of the results evidenced in Tables 3 and 4. Specific models will be sensitive to specific relationships between the covariates and the general correlational structure. For example, only the model in Table 4 provides a straightforward demonstration of covariate dependence although the models in Tables 1, 3 and 4 all provide identical fits. It is therefore important to consider the form of the model adopted and its relationship to the specific questions of interest.

Table 1: Estimation results for model that assumes

$$\text{logit }[Pr(Y_{it}=1\,|\,z_i)] \;=\; \beta_0+\beta_1 z_i \text{ and}$$
$$\text{logit }[\text{corr}(Y_{it},Y_{it-1}\,|\,z_i)] \;=\; \tau_0+\tau_1 z_i .$$

Parameter	Estimate	Estimated S.E.$_{MB}$[1]	Estimated S.E.$_{R}$[2]
β_0	-0.120	0.178	0.238
β_1	0.879	0.279	0.411
τ_0	-0.494	0.487	0.653
τ_1	-0.567	0.818	1.247

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

Table 2: Estimation results for reduced model that assumes

$$\text{logit }[Pr(Y_{it}=1|z_i)] \;=\; \beta_0+\beta_1 z_i \text{ and}$$
$$\text{logit }[\text{corr}(Y_{it},Y_{it-1}|z_i)] \;=\; \tau .$$

Parameter	Estimate	Estimated S.E.$_{MB}$[1]	Estimated S.E.$_{R}$[2]
β_0	-0.151	0.166	0.206
β_1	0.927	0.244	0.257
τ	-0.775	0.409	0.620

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

Table 3: Estimation results for model that assumes

$$\text{logit }[Pr(Y_{it}=1\,|\,z_i)] \;=\; \beta_0+\beta_1 z_i \text{ and}$$
$$\text{logit }[Pr(Y_{it}=1\,|\,Y_{it-1}=1,z_i)] \;=\; \gamma_0+\gamma_1 z_i .$$

Parameter	Estimate	Estimated S.E.$_{MB}$[1]	Estimated S.E.$_{R}$[2]
β_0	-0.120	0.166	0.169
β_1	0.879	0.247	0.255
γ_0	0.712	0.312	0.326
γ_1	0.458	0.414	0.441

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

Table 4: Estimation results for model that assumes

$$\text{logit }[Pr(Y_{it}=0|z_i)] \;=\; \beta_0+\beta_1 z_i \text{ and}$$
$$\text{logit }[Pr(Y_{it}=0|\,Y_{it-1}=0,z_i)] \;=\; \gamma_0+\gamma_1 z_i .$$

Parameter	Estimate	Estimated S.E.$_{MB}$[1]	Estimated S.E.$_{R}$[2]
β_0	0.120	0.166	0.169
β_1	-0.879	0.247	0.255
γ_0	0.887	0.237	0.234
γ_1	-0.910	0.377	0.403

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

4.2 Toronto AIDS Study

Results from a cohort study of HIV infection among male sexual contacts of individuals with AIDS or ARC are presented by Calzavara *et al.* (1991) and by Calzavara *et al.* (1993). A subset of data was obtained from this study to investigate whether frequency of high risk sexual behaviour is associated with the use of recreational drugs. The data subset included information obtained from 176 cohort members. Information regarding high risk sexual behaviour was obtained every three months for up to five years. A score was developed to quantify exposure to high risk sexual behaviour (Calzavara *et al.* 1993) by considering both type and frequency of sexual contact. In the data subset, an individual was classified as high (low) risk with respect to sexual activity if his score value was above (below) the median. Information on recreational drug use at the time of the initial visit also was included.

The methods outlined in Sections 2 and 3 were employed to investigate the relationship between sexual behaviour and use of recreational drugs. Thus the response is defined as $Y_{it} = 1$ if the sexual behaviour risk score for individual i at time t was above the median, $i = 1, ..., 176$, $t = 1, ..., n_i, n_i \leq 20$ (a maximum of 4 observations per year for 5 years). Otherwise, $Y_{it} = 0$. The covariate is an indicator of whether individual i used recreational drugs $(z_i = 1)$ or not $(z_i = 0)$.

The results of fitting the autocorrelation model of Section 2 are presented in Table 5. If the estimate of τ_1 is compared to the robust standard error estimate, it is concluded that the first lag autocorrelation does not depend on recreational drug use. If the model based standard error estimate is used, evidence of such dependence is observed and therefore the potential for erroneous conclusions resulting from the use of the model based standard error estimates is clearly illustrated. It also is concluded that the marginal probability of high risk sexual behaviour depends on recreational drug use in that the probability of high risk sexual behaviour is greater for individuals with a history of recreational drug use compared to individuals with no history of recreational drug use. Since the result of the robust approach does not reveal evidence of non-constant correlation, the model of Zeger *et al.* (1985) can be considered (results not presented).

If the conditional probability approach of Section 3 is considered (Table 6), the following results are obtained. As expected, the conclusions regarding the marginal probability of high risk sexual behaviour are the same as previously outlined for the autocorrelation model. With respect to the conditional probability, it is concluded that the probability of high risk sexual behaviour given previous high risk sexual behaviour is greater for individuals with a history of recreational drug use compared to individuals with no history of recreational drug use.

It should be noted that the data subset also contained an indicator of HIV status for each individual. Initial models included this variable but the results of fitting these models are not included since HIV status did not contribute significantly to any of the models and the inclusion or exclusion of this variable did not affect the estimates of the effect of recreational drug use.

Table 5: Estimation results for model that assumes

logit $[Pr(Y_{it}=1\,|\,z_i)] = \beta_0 + \beta_1 z_i$ and
logit $[\mathrm{corr}(Y_{it}, Y_{it-1}|z_i)] = \tau_0 + \tau_1 z_i$.

Parameter	Estimate	Estimated $S.E._{MB}$[1]	Estimated $S.E._{R}$[2]
β_0	-1.069	0.188	0.211
β_1	0.534	0.195	0.273
τ_0	1.259	0.207	0.278
τ_1	0.240	0.104	0.151

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

Table 6: Estimation results for model that assumes

logit $[Pr(Y_{it}=1\,|\,z_i)] = \beta_0 + \beta_1 z_i$ and
logit $[Pr(Y_{it}=1\,|\,Y_{it-1}=1, z_i)] = \gamma_0 + \gamma_1 z_i$.

Parameter	Estimate	Estimated $S.E._{MB}$[1]	Estimated $S.E._{R}$[2]
β_0	-0.887	0.165	0.233
β_1	1.444	0.198	0.275
γ_0	0.617	0.182	0.260
γ_1	1.117	0.210	0.298

[1] From model based estimate of covariance matrix.

[2] From robust estimate of covariance matrix.

ACKNOWLEDGEMENT

The author is grateful for the assistance of Dr. Vernon T. Farewell and Dr. Janet M. Raboud. This research was supported by the Natural Sciences and Engineering Research Council of Canada and by a grant to the Societal Institute of the Mathematical Sciences (SIMS) from the National Institute on Drug Abuse (NIDA Grant DA-04722). The illustrative data from the New York study were collected by Drs. Don Des Jarlais, Michael Marmor and colleagues under the sponsorship of the National Institute on Drug Abuse (NIDA Grant DA-03574). The illustrative data from the Toronto study were collected by Drs. Randall A. Coates and Stanley E. Read and colleagues under the sponsorship of the Ontario Ministry of Health (Grant No. 01184) and the National Health Research and Development Program, Health and Welfare, Canada (Grant No. 6606-2587-54).

REFERENCES

Calzavara, L.M., Coates, R.A., Johnson, K., Read, S.E., Farewell, V.T., Fanning, M.M., Shepherd, F.A., and MacFadden, D.K. (1991). Sexual behaviour changes in a cohort of male sexual contacts of men with HIV disease: A three-year overview. *Canadian Journal of Public Health*, 82, 150-156.

Calzavara, L.M., Coates, R.A., Raboud, J.M., Farewell, V.T., Read, S.E., Shepherd, F.A., Fanning, M.M., and MacFadden, D. (1993). Ongoing high risk sexual behaviours in relation to recreational drug use in sexual encounters: Analysis of five years of data from the Toronto sexual contact study. *Annals of Epidemiology* (accepted for publication).

Cox, D.R. (1970). *Analysis of Binary Data*. London: Methuen.

Des Jarlais, D.C., Friedman, S.R., Marmor, M., Cohen, H., Mildvan, D., Yancovitz, S., Mathur, U., El-Sadr, W., Spira, T.J., Garber, J., Beatrice, S.T., Abdul-Quader, A.S., and Sotheran, J.L. (1987). Development of AIDS, HIV seroconversion, and potential co-factors for T4 cell loss in a cohort of intravenous drug users. *AIDS*, 1, 105-111.

Farewell, V.T. (1982). Alternatives to the proportional hazards model. In *Environmental Epidemiology: Risk Assessment*, R.L. Prentice and A.J. Whittemore, Eds., 216-229.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Annals. Institute Statistical Mathematics*, 25, 1-26.

Marmor, M., Des Jarlais, D.C., Cohen, H., Friedman, S.R., Beatrice, S.T., Dubin, N., El-Sadr, W., Mildvan, D., Yancovitz, S., Mathur, U., and Holzman, R. (1987). Risk factors for infection with Human Immunodeficiency virus among intravenous drug abusers in New York City. *AIDS*, 1, 39-44.

Zeger, S.L., Liang, K.Y., and Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72, 31-38.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

STATISTICAL ANALYSIS OF PARALLEL TIME SERIES: AIR POLLUTION EFFECTS ON HOSPITAL ADMISSIONS

R. Burnett, S. Bartlett, D. Krewski, G. Roberts and M. Raad-Young[1]

ABSTRACT

The potential adverse health effects of ambient air pollution are examined by relating daily admissions to acute care hospitals in Ontario for respiratory illnesses and daily levels of ozone. The data consist of 197 parallel time series of counts, corresponding to the 197 acute care hospitals used in the analysis. Estimates of ozone levels in the vicinity of each hospital are determined from air pollution monitoring stations maintained by the Ontario Ministry of the Environment. Estimating equation methods are used to examine the effects of ozone on respiratory admissions and the stochastic nature of the responses. The admission data display little evidence of serial correlation. However, admission rates vary considerably among hospitals. This latter source of variation needs to be taken into account in examining the effects of air pollution.

KEY WORDS: Generalized estimating equations; Overdispersion; Air pollution; Respiratory health; Hospital admissions.

1. INTRODUCTION

Epidemiological studies are often used to examine the potential adverse health effects of ambient air pollution (Office of Technology Assessment 1984). Due to the relatively low levels of air pollution in most parts of North America today, any adverse health effects are likely to be subtle, thus requiring large sample sizes to achieve sufficient power for their detection. Many of the protocols that have been used to study the possible adverse health effects of air pollution involve some form of cluster sampling, and are based on longitudinal data on health and air quality.

Administrative health records, such as daily mortality records or hospital records of morbidity have been used as possible indicators of the effects of ambient air pollution on human health. Bates and Sizto (1989) examined daily hospital admissions for respiratory problems in 79 acute care hospitals in southwestern Ontario in relation to daily levels of ambient air pollution. In this investigation, the total number of daily respiratory admissions from 79 hospitals was used as an aggregate response, and related to several air pollutants, including ozone, sulphur dioxide, nitrogen dioxide and sulphates. Because of the longitudinal nature of these data, some form of serial correlation may be present. Due to the varying sizes of the hospitals and the nature of their role in health care delivery, appreciable variation in admission rates among hospitals is also anticipated. This induces a positive correlation between observations within the same hospital. Studies of this type generally involve a large number (several hundred or more) of observations per hospital.

Recent years have witnessed considerable work on the development of statistical methods to analyze such data. Stiratelli *et al*. (1984) used a computationally intensive empirical Bayes approach to analyze longitudinal binary data from a panel study of the effects of particulate air pollution on asthma attack rates. Other methods based on generalized estimating equations are somewhat less burdensome computationally (Liang and Zeger 1986; Prentice and Zhao 1991; and Liang *et al*. 1992). However, these methods of analysis have been implemented primarily with research designs involving only a few observations per cluster. Here, we are interested in clusters

[1] R. Burnett, S. Bartlett and D. Krewski, Environmental Health Directorate, Health & Welfare Canada, Ottawa, Ontario, Canada, K1A 0L2. G. Roberts, Business Survey Methods Division, and M. Raad-Young, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

(hospitals) with hundreds or thousands of observations. Existing methods of statistical inference are not practical for designs with such large cluster sizes.

In this paper, computationally simple adaptations of the estimating equation approach of Liang and Zeger (1986) for regression models with correlated data are explored for research designs involving large cluster sizes. Our methods are developed in such a way that they may be implemented using standard statistical software packages. The proposed methods are illustrated using daily hospital admissions for respiratory illnesses in 197 hospitals in Ontario in relation to daily levels of ambient ozone during the period 1983-1988.

2. REGRESSION MODELS FOR LONGITUDINAL COUNT DATA

Let y_{kt} represent the number of urgent or emergency respiratory admissions on the t^{th} day in the k^{th} hospital $(t = 1,\dots,T; k = 1,\dots,K)$. Consider the model for longitudinal count data

$$E(y_{kt} \mid \epsilon_k, \eta_t) = \lambda_{kt}\, \epsilon_k \eta_t,$$

where λ_{kt} is the unconditional expectation of y_{kt} that is functionally dependent on a $(p \times 1)$ vector of covariates x_{kt}, such as ambient air pollution and climate values, with unknown regression vector β. The ϵ_k are independent scalar random variables with unit expectation and common variance τ, representing the random effect of hospital, and the η_t are random variables with unit expectation and variance-covariance matrix $\phi\Omega$ that induce serial correlation on the responses. Here, $\Omega = ((\omega_{ts}))$ is a $(T \times T)$ correlation matrix whose elements are specified by an $(s \times 1)$ vector ρ of unknown parameters. For example, an AR[1] correlation structure is specified by setting

$$\omega_{t,t-l} = \rho^l,$$

$(l = 0,\dots,T-1; |\rho| < 1)$.

We further assume that the conditional variance is given by

$$\mathrm{Var}(y_{kt} \mid \epsilon_k\, \eta_t) = \theta\, \lambda_{kt}\, \epsilon_k\, \eta_t,$$

$(\theta > 0)$, with $\theta = 1, \theta > 1$, and $\theta < 1$ representing Poisson, extra-Poisson, and intra-Poisson variation respectively. Given ϵ_k and η_t, the conditional covariance between daily admissions within the same hospital is assumed to be zero.

Averaging over the η_t, we have

$$E(y_{kt} \mid \epsilon_k) = \lambda_{kt}\, \epsilon_k,$$

$$\mathrm{Var}(y_{kt} \mid \epsilon_k) = \theta\, \lambda_{kt}\, \epsilon_k + \phi\, (\lambda_{kt}\epsilon_k)^2 \equiv v_{kt},$$

$$\mathrm{Cov}(y_{kt}, y_{ks} \mid \epsilon_k) = \phi\, \omega_{ts}\, \lambda_{kt}\lambda_{ks}\epsilon_k^2$$

and

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \epsilon_k) = \phi\, \omega_{ts}\lambda_{kt}\lambda_{ks}\,\epsilon_k^2\, (v_{kt}\, v_{ks})^{-1/2}.$$

Note that the serial correlation is a function of the length of time between responses and of the conditional expectation $E(y_{ks} \mid \varepsilon_k) = \lambda_{ks}\varepsilon_k$. The dependence on this conditional expectation is minimal if the number of daily admissions is large. In this case, $\theta \lambda_{kt} \varepsilon_k << \phi(\lambda_{kt}\varepsilon_k)^2$, and the serial correlation may be approximated by

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \varepsilon_k) \approx \omega_{ts}.$$

For small $\phi, \theta\lambda_{kt}\varepsilon_k >> \phi(\lambda_{kt}\varepsilon_k)^2$, and the correlation may be approximated by

$$\mathrm{Corr}(y_{kt}, y_{ks} \mid \varepsilon_k) \approx (\phi/\theta)\,\omega_{ts}\,(\lambda_{kt}\lambda_{ks})^{1/2}\,\varepsilon_k.$$

If ϕ is small, then the correlation may also be small even for large values of ω_{ts}.

Averaging over the ε_k we have

$$E(y_{kt}) = \lambda_{kt}.$$

The within hospital covariance matrix $\mathrm{Cov}(Y_k) \equiv V_k$ of $Y_k = (y_{k1}, \ldots, y_{kT})'$ is defined by

$$V_k(\tau, \phi, \rho) = \theta \Lambda_k + \Lambda_k(\tau J_k + \phi(\tau + 1)\Omega)\Lambda_k, \tag{1}$$

where $\Lambda_k = \mathrm{diag}(\lambda_{k1}, \ldots, \lambda_{kT})$ and J_k is a $(T \times T)$ matrix of ones.

Zeger (1988) considered a single time series of count data with covariance given by (1) setting $\tau = 0$. Thall and Vail (1990) also used the covariance function (1) in their analysis of longitudinal count data with no serial correlation (*i.e.*, $\phi = 0$). Burnett *et al.* (1992a) examined error structures similar to those proposed here under the additional assumption that the random effects ε_k and η_t are log-normally distributed.

Two other sources of variation in the admissions series are evident. The admission rates vary by the day of the week with the highest rates on Mondays, followed by a decline in admissions through the week, with the weekend experiencing the lowest rates. Seasonal variation in admissions is also evident with higher rates observed in the winter months, followed by a decline in rates from March until August, and an increase during the fall. The deterministic component λ_{kt} of the model can accommodate these temporal patterns and the influence of air pollution and climate on admission rates with the specification

$$\lambda_{kt} = D_t\, S_t\, f(x_{kt}; \beta).$$

Here $(D_t; t = 1, \ldots, T)$ is a time series with seven unique values representing the ratio of the average number of urgent or emergency admissions on each of the seven days of the week to the average daily admission rate. The factor S_t defined by

$$S_t = \sum_{l=-9}^{9} \Psi_l\, \bar{y}_{t-l'}$$

is a 19 day symmetric linear filter designed to remove temporal trends and slow moving serial correlation. The weights $(\Psi_0, \ldots, \Psi_9)$ are given by (0.087, 0.086, 0.081, 0.073, 0.063, 0.052, 0.040, 0.030, 0.020, 0.012), and $\bar{y}_t$ is the average number of admissions on the t^{th} day over all hospitals. The function f relates the environmental variables x_{kt} to the daily admission series y_{kt} after adjusting for the seasonal and day-of-week temporal trends.

The purpose of the present analysis is to make inferences on the unknown regression parameter vector β, and to obtain estimates of the correlation parameter vector $\alpha = (\tau, \phi, \rho, \theta)'$ that defines the variance-covariance structure of the data.

3. ESTIMATION

Since the distribution of the data has not been specified, likelihood methods for statistical inference cannot be used. Liang and Zeger (1986) suggest the use of estimating equations as a means of making inferences on the regression parameters if the joint distribution of the T observations on the k^{th} cluster is not specified.

Following Liang and Zeger (1986), the regression parameter vector β would be estimated in the following manner. Given a consistent estimate $\hat{\alpha}$ of α, the updated estimate $\hat{\beta}$ of β is given by

$$\hat{\beta} = \beta + \left[\sum_{k=1}^{K} X_k' V_k^{-1} X_k\right]^{-1} \sum_{k=1}^{K} X_k' V_k^{-1} (Y_k - \lambda_k), \tag{2}$$

where X_k is the $(T \times p)$ matrix of derivatives of λ_{kt} with respect to β and $\lambda_k = (\lambda_{kl}, \dots, \lambda_{kT})'$. The right hand side of (2) is evaluated at the current parameter estimates of β and α.

Implementation of this iterative estimation procedure requires specific computer software to be developed; the inversion of V_k, moreover, can pose computational problems for large cluster sizes.

The SAS procedure NLIN (SAS 1988), may be used to obtain parameter estimates for nonlinear regression models by the method of iteratively reweighted least squares. The estimate $\tilde{\beta}$ of β is equivalent to the estimate $\hat{\beta}$ based on (2) if V_k is replaced by the *working* covariance matrix $\tilde{V}_k = \text{diag}(v_{kl}, \dots, v_{kT})$, with the NLIN weight for y_{kt} given by v_{kt}^{-1}. In this paper, we use the weights λ_{kt}^{-1}.

To illustrate our methods, consider an AR[1] error structure for the η_t with scalar parameter ρ. Given the consistent estimate $\tilde{\beta}$ of β, consistent estimators of ϕ, ρ and τ are given by

$$\hat{\rho} = \frac{\hat{C}_3 - \hat{C}_2}{\hat{C}_2 - \hat{C}_1},$$

$$\hat{\tau} = \hat{C}_1 - \frac{\hat{C}_1 - \hat{C}_2}{1 - \hat{\rho}} \tag{3}$$

and

$$\hat{\phi} = \frac{\hat{C}_1 - \hat{C}_2}{\hat{\rho}(1 - \hat{\rho})(\hat{\tau} + 1)}, \tag{4}$$

where

$$\hat{C}_l = \frac{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{r}_{kt} \hat{r}_{k,t-1}}{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{\lambda}_{kt} \hat{\lambda}_{k,t-1}},$$

$(l = 1,2,3)$, with $\hat{r}_{kt} = (y_{kt} - \hat{\lambda}_{kt})$ and $\hat{\lambda}_{kt} = \exp(x_{kt}'\tilde{\beta})$. A consistent estimator $\hat{\theta}$ of θ is given by

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{r}_{kt}^2 - \hat{\lambda}_{kt}^2(\hat{\tau} + \hat{\phi}\,(\hat{\tau}+1))}{\sum_{k=1}^{K} \sum_{t=l+1}^{T} \hat{\lambda}_{kt}}.$$

If serial correlation is not present ($\rho = 0$), τ may be estimated by

$$\hat{\tau} = \frac{\sum_{k=1}^{K} \left(\hat{S}_k^2 - \sum_{t=1}^{T} \hat{r}_{kt}^2\right)}{\sum_{k=1}^{K} \left(\hat{G}_k^2 - \sum_{t=1}^{T} \hat{\lambda}_{kt}^2\right)}, \tag{5}$$

where $\hat{S}_k = \sum_{t=1}^{T} \hat{r}_{kt}$ and $\hat{G}_k = \sum_{t=1}^{T} \hat{\lambda}_{kt}$. (Note that the estimator of $\hat{\tau}$ in (5) is based on all possible cross products of residuals within each cluster, whereas the estimator in (3) is based on only the first and second off-diagonals of the matrix of cross products of residuals).

If $\rho = 0$, an estimate of ϕ may not be obtained using (4). In this case, estimates of both ϕ and θ may be obtained by noting that given $\tau = \hat{\tau}$ and $\lambda_{kt} = \hat{\lambda}_{kt}$,

$$\mathrm{Var}(y_{kt}) = \theta\hat{\lambda}_{kt} + (\hat{\tau} + \phi(\hat{\tau}+1))\hat{\lambda}_{kt}^2. \tag{6}$$

The parameters θ and ϕ are estimated by simple linear regression using the right hand side of (6) to predict the observations $(y_{kt} - \hat{\lambda}_{kt})^2$.

The covariance matrix of $\tilde{\beta}$ is given by

$$\mathrm{Cov}(\tilde{\beta}) = \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} X_k\right]^{-1} \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} V_k \tilde{V}_k^{-1} X_k\right] \left[\sum_{k=1}^{K} X_k' \tilde{V}_k^{-1} X_k\right]^{-1} \tag{7}$$

(Liang and Zeger 1986). An estimate of the covariance of $\tilde{\beta}$ is obtained by evaluating $\mathrm{Cov}(\tilde{\beta})$ at $(\hat{\beta}, \hat{\alpha})$.

The multiplication of very large matrices such as $X_k' \tilde{V}_k^{-1} V_k \tilde{V}_k^{-1} X_k$ can pose difficulties for the SAS matrix language PROC IML, due to RAM memory limitations on microcomputers. However, the matrix multiplications in (7) may be written as sums-of-squares. Using this formulation, the required matrix multiplication may be performed by the SAS procedure PROC CORR, which calculates the sums-of-squares matrix for several variates.

4. AIR POLLUTION EFFECTS ON HOSPITAL ADMISSIONS

Bates and Sizto (1989) examined the relationship between daily hospital admissions for respiratory problems to 79 acute care hospitals in southwestern Ontario and ambient air pollution in the months of January, February, July and August, 1976 to 1983. A subsequent study (Burnett *et al.* 1992b) is based on the number of daily urgent and emergency respiratory admissions to acute care hospitals in Ontario during the period January 1, 1983 to December 31, 1988, totalling over 400,000 hospital days. We were able to obtain reasonable estimates of ozone exposure in the vicinity of 197 hospitals. Because the observations within hospitals are ordered in time, the possibility of serial correlation exists.

For illustrative purposes, the relationship between urgent or emergency respiratory hospital admission rates (*International Classification of Disease Codes* 466, 480-486, 490-496, 786) in the months of May to August and the daily maximum one hour average ozone level x_{kt} is examined, using the model

$$f(x_{kt};\beta) = \exp(\beta_0 + \beta_1 x_{kt}). \tag{8}$$

Since $\hat{\phi} \approx 0$, there is little evidence of serial correlation. Estimates of θ, τ, β_1 and the standard error of β_1 for *population average* and *hospital specific* models are given in Table 1 using data either for Ontario as a whole or for Toronto alone. (*Population average* models describe changes in admission rates for all of Ontario associated with changes in levels of ozone while *hospital specific* models characterize these changes in each hospital.) Since $\hat{\theta} \approx 1$, there is little evidence of over or under-dispersion relative to Poisson variation. This is because hospital admissions for respiratory illnesses are rare events, with only 154 daily admissions for Ontario summers with approximately nine million people potentially at risk of being admitted to hospital each day.

Table 1: Parameter Estimates and Standard Errors for Population Average and Hospital Specific Regression Models.

Parameter (units)	Ontario (197 hospitals)		Toronto (24 hospitals)	
	Population Average	Hospital Specific	Population Average	Hospital Specific
θ (adms^2)	1.04	1.04	1.08	1.08
τ (adms^2)	0.75	≈ 0	0.31	≈ 0
$\beta_1 \times 10^{-4}$ (adms/ppb)	24.4	1.78	1.75	1.73
$s.e.(\hat{\beta}_1)_{Ind} \times 10^{-4}$ (adms/ppb)	1.74	1.42	3.52	2.97
$s.e.(\hat{\beta}_1)_{Dep} \times 10^{-4}$ (adms/ppb)	10.4	1.42	2.97	2.97

Under the *population average* model for Ontario, the between hospital variance is $\hat{\tau} = 0.75$, and the intra-hospital correlation is $\hat{\tau}\hat{\lambda}(1 + \hat{\tau}\hat{\lambda})^{-1} = 0.36$, evaluated at the average rate of $\hat{\lambda} = 0.78$ admissions per day. The estimate $\hat{\beta} = 24.4 \times 10^{-4}$, of the regression coefficient for ozone corresponds to 24.8 fewer daily admissions in all 197 hospitals for a decrease in ozone from its mean value of 51.6 ppb to 0 ppb. This result implies that ozone is associated with 24.8/154=16% of all urgent or emergency respiratory admissions in Ontario summers.

If the observations are assumed to be independent ($\tau = 0$) then the standard error of $\hat{\beta}_1$ is 1.74×10^{-4} (see Table 1, $s.e.(\hat{\beta}_1)_{Ind}$), suggesting strong evidence for a positive ozone effect. However, if τ is assumed to be positive, the standard error increases to 10.4×10^{-4} (Table 1, $s.e.(\hat{\beta}_1)_{Dep}$). This 6-fold increase in error is due to an ozone signal arising from differences in admission rates between hospitals. Higher ozone levels are observed in the south-western portion of the province, which is more heavily populated than the northern and eastern regions. These differences in population density induce differences in admission rates which are also positively correlated with ozone levels.

The purpose of the present analysis is to examine the effects of short term episodes of air pollution on admission rates. To properly examine daily fluctuations in admissions in relation to air pollution within each hospital, the differences in rates among hospitals should be removed. This is accomplished by considering the conditional expectation

$$E(y_{kt} \mid \varepsilon_k) = D_t S_t \exp(\beta_0 + \beta_1 x_{kt})\ \varepsilon_k .$$

Under this model, a common effect of air pollution, β_1, is regressed on the ratio of the daily admissions for a given hospital, y_{kt}, to the corresponding hospital effect, ε_k. An estimate of ε_k is given by the ratio of the average number of daily admissions for the k^{th} hospital to the average daily admission rate for the entire dataset.

Although the estimate of θ under this *hospital specific* model is the same as under the *population average* model, $\hat{\tau} \approx 0$, indicating that all variation in rates among hospitals has been removed. This adjustment removes any cross-sectional signal of air pollution on admissions, leaving only the longitudinal air pollution signal. The estimate $\hat{\beta}_1 = 1.78 \times 10^{-4}$ under the *hospital specific* model is much less than that obtained under the *population average* specification, indicating that much of the relationship between air pollution and respiratory admissions originates from variation in rates among hospitals.

Under the *hospital specific* model $\hat{\tau} \approx 0$ and thus τ is set to zero. Under this error structure, the two specifications of standard error (*Ind* and *Dep*) are equivalent since the data are independent. An advantage of this approach over the *population average* model is that standard computer software may be used for both estimation and inference. The estimates of the standard errors obtained from PROC NLIN, for example, are consistent under the *hospital specific* model specification.

It is well known that an analysis which includes blocking will yield more power to detect effects that vary only within a block and are balanced between blocks, than an analysis assuming the data are independent. To illustrate this point, consider data from the 24 acute care hospitals in Toronto. Here, patients admitted to all Toronto hospitals are assumed to have been subjected to the same level of exposure to tropospheric ozone. Thus, this covariate is common among hospitals and varies over time within each hospital under the *population average* and *hospital specific* models for Toronto. The estimates of β_1 are similar and, interestingly, comparable to the results for the entire province based on the *hospital specific* model. The standard error of $\hat{\beta}_1$ assuming independent responses is 3.52×10^{-4}. A 15% reduction in the error is obtained $(s.e.(\hat{\beta}_1) = 2.97 \times 10^{-4})$ by recognizing the dependence of observations within a hospital. This represents only a modest gain in power since the variation in rates among Toronto hospitals, $(\hat{\tau} = 0.31)$, is less than that observed for all Ontario hospitals $(\hat{\tau} = 0.75)$. This is due to the limited range in the size of Toronto hospitals. The standard error of $\hat{\beta}_1$ based on the *hospital specific* model is similar to that obtained for the *population average* model.

5. DISCUSSION

Computationally simple methods for estimation and inference are presented for longitudinal count data with large cluster sizes. The form of the covariance structure accounts for variation in daily admissions between hospitals and autocorrelation in the counts over time. The bias in the estimate of the effects of air pollution on respiratory hospital admissions due to the positive association between population density and air pollution levels may be removed by considering *hospital specific* regression models. This specification permits the use of standard statistical software to be used for both estimation and inference on regression parameters.

The estimation procedure for the overdispersion parameters presented in section 3 is motivated by computational considerations. Although estimating β and α with separate equations yields highly efficient estimates of β a considerable loss in efficiency may be anticipated for α. Liang *et al.* (1992) show that for correlated binary data, estimating β and α by separate equations can yield as much as a 50% loss in efficiency for the estimates of α, compared to joint estimation of α and β. Joint estimating equations could be used to avoid this loss in efficiency, although their implementation is not practical for large cluster sizes.

ACKNOWLEDGEMENTS

The authors are grateful to the Ontario Ministry of Health for providing the hospital admissions data and to the Ontario Ministry of the Environment for providing the air pollution data.

REFERENCES

Bates, D., and Sizto, R. (1989). The Ontario air pollution study: identification of the causative agent. *Environmental Health Perspectives*, 79, 69-72.

Burnett, R.T., Shedden, J., and Krewski, D (1992a). Nonlinear regression models for correlated count data. *Environmetrics*, 3, 211-222.

Burnett, R.T., Dales, R.E., Raizenne, M.E., Krewski, D., Summers, P.W., Roberts, G.R., and Dann, T.F. (1992b). The relationship between hospital admissions and ambient air pollution in Ontario, Canada: a preliminary report. In Proceedings of the Air and Waste Management Association 85th Annual Meeting & Exhibition, Kansas City, Missouri, June 21-26, 1992. Reprint 92-146.05.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society Series* A, 54, 3-40.

Office of Technology Assessment (1984). Acid rain and transported air pollutiants - implications of public policy. U.S. Government Printing Office, Washington, DC.

Prentice, R.L., and Zhao, L.P. (1991). Estimating equations for parameters in means and covariates of multivariate discrete and continuous responses. *Biometrics*, 47, 825-839.

SAS Institute Inc. (1988). *SAS/STAT User's Guide, Release 6.03 Edition*. Cary, NC: SAS Institute Inc., 1028.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40, 961-971.

Thall, P.F., and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657-671.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika*, 75, 621-629.

SESSION 6

General Applications I

Proceedings of Statistics Canada Symposium 92
Designing and Analysis of Longitudinal Surveys
November 1992

THE "ÉCHANTILLONDÉMOGRAPHIQUE PERMANENT" PANEL: A FRENCH EXPERIENCE IN LONGITUDINAL FOLLOW-UP ON PERSONS

M. Isnard[1]

ABSTRACT

Since 1968, INSEE has been doing longitudinal follow-up on persons. This follow-up involves matching, at the individual person level, data derived from successive censuses and statistical civil status reports. The file, which covers more than 700,000 individuals, living or deceased, can be used for differential studies (mobility, mortality) and methodological studies, notably with respect to certain census variables. In 1995, a new system of computerized management will replace the current manual system of management, which is cumbersome and results in errors.

KEY WORDS: Census; Civil status report.

1. INTRODUCTION

Most demographic statistics use results from censuses and civil status reports. A census has the advantage of providing, for the population as a whole, a certain amount of information on some characteristic (for example, place of residence or social category) for a given date. It could be compared to a photograph or an inventory. A civil status report looks more at flows: it makes possible better knowledge and a better understanding of the various demographic events experienced within a population.

These two complementary sources are used by the demographer with reference to the phenomenon studied. However, these data are insufficient. For example, to measure fertility by social milieu, the traditional method, which uses civil status and census reports, is inadequate. Social category is a difficult characteristic to measure and, depending on the conditions under which the information is collected, a given individual can be placed in different categories.

Moreover, civil status statistics and censuses do not provide detailed information on an individual's history, even on the demographic level. For example, a study of geographical or occupational mobility over twenty or so years is impossible using censuses. To go further, one must use more suitable tools, such as retrospective surveys, where individuals are questioned about their past, or panels - that is, information systems that accumulate information through follow-up on individuals.

In retrospective surveys, people may forget and respond inaccurately. Some events, such as short interruptions of activity or even the birth of children who died at a very young age, may simply be forgotten or not mentioned.

For all these reasons, and to better determine the quality of the censuses, INSEE established a longitudinal file called the "Échantillon Démographique Permanent" (EDP, permanent demographic panel). On the level of individuals, reports from the various censuses since 1968 and the civil status reports for the same person (see Appendix 1: Content of EDP) are matched. To facilitate identification of individuals, the "Répertoire National des Personnes Physiques" (RNIPP, national inventory of natural persons) was used: a registration number in the inventory makes it possible to follow each individual more easily.

[1] M. Isnard, INSEE, 18 Bd Adolphe Pinard, 75675 Paris Cedex 14.

2. INSEE'S DEMOGRAPHIC PANEL

A sample ...

It was difficult to design matching of the civil status data and the census data for the entire population of France, so the decision was made to look at only a subset of the population. To obtain a sampling rate of approximately 1%, 4 dates of birth were chosen. Any person born on one of those four days is included in the EDP as soon as one of the reports described in the appendix and concerning that person has been collected.

... Demographic ...

The purpose of this sample is to make the conducting of demographic studies possible. The content of the census and civil status reports differs from that in other countries: for example, the census makes no reference to income and it is against the law for INSEE to conduct studies taking medical data (causes of death, for instance) into account. Rather, the purpose of this file is to make it possible to conduct demographic or sociological studies that require follow-up on individuals.

... Permanent

The EDP was begun in 1968, and is still being built on. At present, the data from the 1990 census of the population and the civil status reports for the years 1982 to 1989 are being added. A project for continuous enrichment of the EDP is being developed and should be completed during the first quarter of 1994.

3. MANAGEMENT OF EDP

The EDP, which was created in 1968, is characterized by manual management that is cumbersome and delicate to administer. For each individual, there is a paper file containing the various reports on that individual. This file is kept at the regional office of INSEE that corresponds to the individual's place of birth. At present, there are 711,038 files containing some 3 million reports of all kinds. Insertion of reports in files is done entirely by hand and the files are organized in an *ad hoc* manner. The RNIPP is used only when there are identification problems. The 1990 census and 1982-1989 civil status reports are now being inserted. This work will be over at the end of 1992 and will be followed by a coding phase and then a data-capture phase. The complete file containing the reports for the years from 1968 to 1990 will be available toward the end of 1994.

However, the information contained in the individual census reports is insufficient. It was therefore decided that the magnetic EDP file would be partly enriched with the information on housing and families for individuals from the 1975 and 1982 censuses. Similarly, information on the death of individuals was drawn from the RNIPP.

Of course, the information is complete only if the individual was in metropolitan France for the various censuses, completed a census report there and specified his or her date of birth.

4. STUDIES CONDUCTED USING EDP

Despite its limitations, the sample offers many possibilities, especially when one considers the information provided by the civil status reports. For example, the sampling rate of 1% allows for studying residential mobility over four censuses (1962, 1968, 1975 and 1982) with some 315,000 individuals. Thanks to the marriage reports, one can see how a change in marital status influences residential mobility.

In addition, the method used to construct the EDP prevents any distortions of the "memory effect" type and even allows for measuring that effect in some cases. The range of studies that can be done is vast and the description below is by no means exhaustive. A list of the studies published using the file is given in Appendix 2.

Differential demographic studies

The purpose of these studies is, through matching of information from the civil status and census reports, to calculate the demographic rates with respect to fertility, nuptiality and mortality. In particular, the EDP allows for better identification of the links between occupational mobility (change in occupation) and mortality. The traditional mortality samples do not allow for calculating this type of indicator.

Life cycle, mobility, migrations

The EDP can be used to reconstruct an image of the family and occupational life cycle of individuals, particularly with respect to their spatial mobility. Analysis of the changes in individuals' situations between two census reports can provide a measurement of occupational or geographical mobility. Use of civil status reports allows one to better study the link between the two forms of mobility and nuptiality or fertility.

Methodological studies

The EDP makes possible methodological studies related to the quality of censuses. Although this field remains wide open, we can cite studies on omission from the census for persons who died shortly before or shortly after the census, and studies on double counts - that is, people who were included in the census several times in error (and for whom there are therefore several census reports in the file). It has thus been possible to study the quality of the responses to some questions (diploma, nationality) for two successive censuses (see Appendix 3: Methodological Study on Level of Education).

Moreover, matching of information from the census and civil status reports allows for studying consistency between these two sources for "sensitive" variables such as socio-occupational category or nationality.

EDP as sampling frame

The EDP has also been used as a sampling frame for a number of surveys organized by INSEE:

- Lorraine sociodemographic survey. This is a regional operation set up by the *direction régionale de Lorraine* in co-operation with the University of Nancy. It is aimed at providing a social and economic description of households in Lorraine, as well as at analysing the impact of social policies on individuals. The sample was made up of individuals drawn at random from the EDP. The advantage of using this sampling frame is that information regarding the 1968-1982 period does not have to be collected.

- studies on inclusion on electoral lists and participation in elections. These studies were conducted on subsamples from the EDP for which characteristics from the 1975 census were retained.

- survey on estates of decedents. The aim here was to better understand the behaviour of individuals with respect to the transferring of estates upon their death. A sample of individuals who died in 1988-1989 was drawn from the EDP and a survey was conducted in the *trésoreries générales* (General Accounts Dept.). The advantage of the EDP for this survey was that it provided information regarding the individual's situation in 1968 - that is, in a period of activity - without the need for additional questioning.

- survey on social and geographical mobility. The purpose of this survey, which is now being done in the field, is to measure integration of immigrants into French society. To measure the "second generation" effect, two subsamples were drawn. The first is made up of individuals drawn from the 1990 census who were born abroad in or around the 1960s and who immigrated to France between 1982 and 1990. The second is made up of individuals who belong to the same generations, but were born in metropolitan France of a naturalized father or father of foreign nationality. Only a source such as the EDP allows for identifying such individuals.

5. NEW MANAGEMENT OF EDP

Manual management of the 711,038 files was very cumbersome and generated errors. For example, the encoding of the reports for the 1968-1982 period took 700,000 hours, or approximately 460 person-years. The encoding of the reports that is now being done should take some 100,000 hours. In addition, it was difficult to reconcile continual enrichment with the handling of paper files. Finally, more and more civil status reports were coming to INSEE in magnetic form and the storage of paper records did not seem to be a long-term solution.

Consequently, INSEE decided to design a new form of management for the EDP, one that would be much more automatic. This new management approach, which will be implemented at the end of the first quarter of 1994, should allow for continual enrichment of the EDP using the civil status reports and for automatic inclusion in the case of computerized reports. The principle is as follows: the RNIPP will be used to identify an individual by means of the civil status number on the report. Once the identification has been made, the information will be recovered directly from the civil status statistical files. Only in cases where there is a legal dispute respecting identification will the information not be inserted immediately. In such cases, it will be necessary to investigate.

It has been estimated that, under the current system, the time required for complete processing (including the processing of cases involving disputes) for one year of civil status reports would be some 15,000 hours, including encoding, data capture and processing of cases involving disputes. The reports for one year would not be integrated before the end of year n + 2.

APPENDIX 1

CONTENT OF EDP

Reports collected

The individual reports from the 1968, 1975 and 1982 censuses were collected and inserted into the files. Both ordinary reports and reports for persons living in institutional households were included.

The following civil status reports were inserted:

- supplementary information (marriage, death and so on) report;
- acknowledgement report for EDP child or EDP parent;
- stillbirth report for EDP parent;
- death report for EDP person (before 1974);
- birth report if child, father or mother is in EDP;
- marriage report, where spouse or legitimatized child included in EDP.

These same reports are being inserted for the 1982-1990 period, as are the individual reports from the 1990 census.

Computerized file

The computerized file contains, in addition to the above-mentioned reports, information regarding the housing of individuals and their families from the 1975 and 1982 censuses. Matching with the RNIPP revealed the dates of death for persons born in Metropolitan France who died in 1990 or before.

It is stored in the form of an SAS base containing 1,295 variables.

APPENDIX 2

LIST OF MAIN PUBLICATIONS CONCERNING EDP

January 1987: L'échantillon démographique permanent de l'INSEE. Courrier des Statistiques. (O. SAUTORY, INSEE).

1988: Participation électorale. Économie et statistique Nos 152, 165 and 178. (J. MORIN, INSEE).

February 1988: internal report. Étude sur les déclarations des diplômés aux recensements de la population de 1975 et 1982. (O. SAUTORY).

April 1988: Economie et Statistique No 209. Plus de la moitié de la population a changé au moins une fois de commune entre 1962 et 1982 (O. SAUTORY).

October 1989: Fifth meeting of CICRED network (Paris): Mortalité différentielle (M. ISNARD).

September 1990: ISI conference (Cairo): La mobilité en France entre 1962 et 1982 (M. ISNARD).

October 1991: European Conference on Demography: Mobilité géographique d'après l'EDP (G. DESPLANQUES and M. ISNARD).

Being published: Mobilité Sociale (A. CHENU - CNRS/INSEE).

Being published: Les migrations des personnes en Ile de France. (F. CRIBIER and A. KYCH CNRS).

APPENDIX 3

METHODOLOGICAL STUDY ON LEVEL OF EDUCATION

When one is conducting a differential study, it is important that the criteria defining the various subpopulations be as reliable as possible (and, of course, that they relate to the phenomena studied).

One traditional criterion for explaining differences, in demography and sociology, is level of education. This is practically a fixed variable for persons 25 years of age or older. Using the EDP, the responses concerning level of education in the 1975 and 1982 censuses were compared for 56,000 individuals.

The table below summarizes the information regarding the level of education reported in the two censuses for individuals who responded twice. Only 85% of the reports were consistent.

Population category	75 level = 82 level	75 level > 82 level	75 level < 82 level
Overall	84.8%	7.5%	7.7%
Men	82.7	8.5	8.7
Women	86.7	6.6	6.7
Age in 1975			
25-34 years	77.9	11.7	10.4
35-44 years	83.0	8.8	8.2
45-54 years	86.0	6.8	7.2
55-64 years	87.9	5.0	7.0
65 years and older	93.0	2.7	4.3
Nationality in 1982			
French	84.6	7.6	7.8
Foreign	92.2	3.9	3.9
Socio-occupational category in 1982			
Farmer	90.8	5.8	3.4
Skilled tradespeople, merchants	78.4	10.2	11.4
Senior managers	70.8	13.2	16.0
Intermediate professionals	69.0	17.4	13.7
Employees	79.5	10.5	9.9
Labourers	88.4	5.8	5.8
Retires	90.2	4.3	5.5
Others not in labour force	88.2	5.4	6.4

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

METHODOLOGICAL EXPERIMENTS IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

R.P. Singh[1]

ABSTRACT

The Survey of Income and Program Participation (SIPP) a longitudinal survey of households conducted by the U.S. Census Bureau. It provides cross-sectional and longitudinal data on labor force, income, government programs and other person and household characteristics that may influence an individual's economic well being. The U.S. Census Bureau instituted an extensive research and evaluation program for the SIPP. As part of the program we conducted about ten survey methodological experiments including interview mode, collection of missing wave data, employer provided benefits, and gift to respondents. This paper briefly summarizes the SIPP methodological experiments and their results.

KEY WORDS: Response rate; Data quality; Seam problem.

1. INTRODUCTION

The Bureau of the Census has been conducting interviews for the Survey of Income and Program Participation (SIPP) since October 1983. The SIPP is a national survey and is designed to provide improved information on income and participation in government programs for the noninstitutionalized United States population. Person and household characteristics that may influence income and program participation are also available from the SIPP. This information is vital for improving the capability of federal agencies to formulate and evaluate their policies and programs in the areas of income and social welfare.

The SIPP is the Bureau's first large scale longitudinal survey of households and is very complex. The Census Bureau had no experience with a survey of this type and a number of questions and issues were unresolved. Therefore, in 1984, the Bureau started an extensive research and evaluation program to learn strengths and weaknesses of the SIPP. Research also included searching for ways to reduce or eliminate weakness of the data and make the SIPP more efficient. As part of this research, we conducted methodological experiments on the SIPP. This paper presents a brief summary of these experiments and their results.

Section 2 of this paper presents the SIPP sample design. Section 3 briefly discusses the SIPP methodological experiments. Section 4 presents a summary and conclusions.

2. SIPP SAMPLE DESIGN

The SIPP is a multistage stratified systematic sample of the noninstitutionalized resident population of the United States. Only persons who are at least 15 years of age are eligible for interview, although limited data on children is also collected by proxy interviews.

Initially, a sample of housing units in selected Primary Sampling Units (PSUs) is taken. The SIPP sample is divided into four groups of equal size called rotation groups. One rotation group is interviewed each month.

[1] R. Singh, Demographic Statistical Methods Division, U. S. Bureau of the Census, Washington, D.C 20233 U.S.A. This paper reports the general results of research undertaken by the U.S. Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

In general, one cycle of four rotation groups is called a wave. Persons in the sample are interviewed once every four months for 32 months. The reference period for the interview is the four months preceding the interview month. Persons 15 years old and over present as household members at the time of first interview are part of the survey for the entire 32-month period. With certain restrictions, we follow these sample persons if they move to a new address. "New" persons living with sample persons are also part of the sample only while residing with these sample persons. For more details on the SIPP design, see Nelson, McMillen, and Kasprzyk (1985).

The SIPP questionnaire is long and complex. Questions are asked about assets and labor force status, and on specific type of cash and non-cash income by months received and amounts per month. For many types about assets and labor force status, and of income, additional questions are asked of recipients. In most interviews, questionnaires also include additional questions (topical modules) on various subjects.

3. METHODOLOGICAL EXPERIMENTS

Due to the SIPP's longitudinal nature, it's procedures are very different from other surveys of the Census Bureau. Therefore, the results of experiments conducted on other surveys could not be applied directly to the SIPP. Hence, we designed and implemented the following SIPP methodological experiments.

1. Asset and Liability Feedback
2. Employer-Provided Benefits
3. Gift to Respondent
4. Maximum Telephone Interviewing
5. Missing Wave Data
6. Eighth Interview
7. Time-line Calendar Aid to Respondents
8. Cognitive Research
9. Debriefing of Respondents
10. Income Source (ISS) Record Checks

Below, I briefly discuss these experiments along with their results. Some of these results have been presented before in various statistical meetings.

1. Asset and Liability Feedback Experiment

The SIPP collects data on assets and liabilities because of their importance in determining programs eligibility and assessing the economic situation of families. The data on assets and liability were collected in the topical modules of Wave 4 and Wave 7 of the 1984 panel. Therefore, this experiment was designed to determine if a respondent when provided previous year data on assets and liabilities will provide better quality data. The rationale for the experiment was that the respondents would provide more reliable estimates at Wave 7 if they were first reminded of the amount they reported in Wave 4 for the previous year. If a respondent knew the amount change in asset value, then Wave 7 response would be consistent with the amount of change over the year.

For the experiment, we split the sample in two halves. One half used regular interview procedures. For the other half, the field representative (FR) gave each respondent a computer printed asset-feedback form at the beginning of the asset and the liability portion of the interview. The FR also read an introductory statement explaining that the form contained information collected for the respondent one year ago and should be used by the respondent as a reference when answering similar items during the interview. The FR assisted the respondent, if necessary, by referring to a line item of the feedback form showing same from Wave 4. Because of confidentiality concerns, feedback forms were used only when the same person responded in both waves.

Evaluation of the experiment tested if annual changes for mean and median net worth and correlation between two years were different for the two groups. The analyses included population subgroups. Estimates of changes for these estimates gave no statistical evidence of consistent differences between the two groups. For example, the estimated decline in median net worth for the feedback and non-feedback groups, were $590 and $860 respectively. These changes were not statistically significant. Also analysis did not show strong statistical increase in correlations for population groups studied (Lamas and McNeil 1987; Weidman *et al.* 1988).

Even though the feedback procedure appeared to reduce estimates of changes and increases the correlations between the estimates of two years, the results have insignificant analytical implications. This may be due to very small changes occurring in assets and liabilities over a short-term period. Because of additional respondent burden and SIPP's inability to detect short-term changes in assets and liabilities, the collection of such data twice during a panel was dropped.

2. Employer-Provided Benefits Experiments

While the SIPP collects a large amounts of data on labor force, source of income and amount by income source, it does not collect data on employers' contributions to insurance, retirement, *etc.* In 1986, the Bureau considered collecting data on employers' contributions for life, and health insurance plans, and pension plans from the respondent's employer. High quality data from the employer's records would supplement the SIPP data and be extremely useful for researchers and policy makers.

Since the SIPP is very complex and collects very sensitive data, adding another potentially sensitive data request placed respondents and FRs in a more difficult situation with possibly more FR turnover. FRs expect negative reactions from respondents to sensitive and longer questionnaires. The negative reaction may not be limited to only additional questions but can affect the overall FR performance. Due to these concerns, the experiment was designed to estimate response rate for getting employer-provided benefit data from respondents' employer rather than overall response rate.

The experiment was conducted in August of 1987 on one half of the sample from Rotation 4 of Wave 8 of the 1985 Panel. The experiment included only those employed persons who were 18 years of age or older, and completed a Wave 8 interview questionnaire. A total of 1352 persons were eligible for the test at Wave 8.

After completing the Wave 8 interview, the FR determined if the respondent was eligible for the experiment. If a respondent was eligible, the FR asked the respondent to sign an "Employer Questionnaire and Authorization Form" to authorize the Bureau to request the employer to provide information regarding insurance and retirement contributions. The form was left with a proxy respondent and mailed to a telephone interview respondent. We did not follow-up those who did not return a form. The authorized questionnaires were mailed out to employers. We followed up the employers who did not return all their questionnaires.

Only 40% of the eligible SIPP respondents signed the authorization forms. Follow-up of the respondents could have increased this rates. Employers returned 96% of the questionnaires. Item nonresponse rate, in general, was low. The reaction from the FRs was also very positive. For details, see Adams (1988).

3. Gift Experiment

SIPP, being a longitudinal survey, suffers higher non-response rates than Bureau's other demographic surveys'. In an attempt to reduce the nonresponse rate, we presented a token gift (a small solar-powered calculator) immediately after the FR introduced the survey to wave 1 households of the 4th rotation in the 1987 panel. The remaining three rotations did not receive any gift.

Nonresponse rates of the gift receiving rotation were compared with the other three rotations and with the earlier panels -- 1984, 1985 and 1986 panels. Results of the analyses suggested that the gift may help in reducing the nonresponse rate at the national level. However, the wave-to-wave rates of increase during the life of the panel for the recipient and non-recipient rotations did not differ significantly [Butler 1991].

Results from the 1988 and 1989 panels suggested that intervention by Field Division reduced nonresponse to a similar degree as giving a gift. For example, for the 1988 panel, the average nonresponse rate for the first three rotations was 7.71%. Before the interview for the fourth rotation occurred, Field Division, concerned over high nonresponse rates, asked the Regional Offices for reasons for the increase in nonresponse rates over earlier panels. As a result of asking, the noninterview rate for the fourth rotation dropped to 6.73%. A similar situation occurred for 1989 panel.

The gift appeared to have positive effects on the response rate in the beginning of the panel. However, field intervention seemed to be just as effective in lowering nonresponse rates as giving a gift. It was not possible to determine whether both used together would reduce nonresponse rates even more. There was no evidence of the gift having any effect after the first interview. It was possible that giving a gift more than once (preferably in the middle of the panel) could reduce the nonresponse rate.

4. Maximum Telephone Interviewing

Due to the complexity and sensitivity of the SIPP data and the length of interview, personal visit interviewing was generally believed to be the only effective data-collection mode. But, due to various SIPP budget cuts and the Bureau's continued efforts to make the SIPP more efficient, we explored the use of telephone interviewing in the SIPP. The objective of the telephone interviewing experiment was to determine whether the SIPP households could be interviewed by telephone with no or negligible loss in data quality.

To reach our goal, we tested the telephone interviewing mode in two steps: a feasibility test, and a national test. Any case assigned to be telephone interviewed was considered to be in the telephone interviewing mode even if a personal visit had to be conducted to complete an interview. We called it maximum telephone interviewing mode (MTI).

The Bureau conducted the feasibility test in June 1985 to see if the SIPP respondents could be interviewed on the phone without increasing the nonresponse rate. The response rate for the MTI mode was as high as for maximum personal visit (MPV) mode (Durant and Gbur 1988). Overall, the test was very encouraging.

We conducted the national test during the 1986 panel. We designated approximately half of the SIPP 1986 panel sample household to each of the MTI and MPV modes across the nation. A dear friend letter informed the respondents that their next interview would be completed by telephone. The FRs conducted telephone interviews from their homes.

FRs did not receive special classroom training for the MTI mode but completed a self-study prior to beginning an assignment. FRs conducted initial MTI August through November, 1986 (defined as phase I) and second MTI February through April 1987 (defined as Phase II). These interviews were completed during Wave 2 through Wave 4. This design allowed evaluation of wave estimates and transition estimates between two consecutive waves.

Overall, Gbur *et al.* (1989, 1990) found minimal effect on cross-sectional estimates and biases in some longitudinal estimates. The following summarizes results of their analyses.

- Household nonresponse rates at the national level were the same for two interview modes.
- Median family income for 1986 and 1987 by demographic and geographic groups were the same by mode.
- Mean income-to-poverty threshold ratio was different for Hispanics and not for other groups.
- Average household size for the telephone designated sample (2.8 persons) is lower than personal visit designated sample (2.9 persons).
- Percent of self-response for the MTI (62.2) is lower than the 64.7% for MPV cases.
- MPV designated cases reported higher but not statistically different recipiency rates and amounts in government programs.
- Number of recipiency transitions (receiving to not receiving) were different for only 3 income sources out of 26 sources analyzed.
- The percent of persons going into poverty for MTI cases is generally lower than the MPV.
- Higher percentage of total persons and of selected subgroups (race and sex) completed their unemployment spell for MTI than MPV.
- Hourly wage rates for total and Blacks in MTI were higher than MPV.
- MTI Respondents often did not use flashcards provided as interview aids.
- Distribution of children ever born to a woman differed by mode.

- Significantly higher percent of MPV respondents reported second and third youngest child in daycare.

The last two results are from the topical module data. Most of the other estimates examined by mode were not statistically different.

McNeil (1989) compared quarterly income, recipiency, and labor force estimates from the entire 1986 panel with similar estimates from the 1985 panel. He did not find any difference that is unique to the time period covered by the telephone experiment.

The 1986 and 1987 calendar year telephone and personal visit based estimates were also compared to CPS. Relationships between the estimates of change from 1986 to 1987 were also compared. This comparison did not provide any clear cut indication that the telephone based estimates were better or worse.

Comparison of the two modes suggested minimal effects on estimates for total population. Most of the significant differences were concentrated in low income groups. However, these differences were small and should not have significant effect on accomplishing SIPP goals. Therefore, the Bureau decided to use maximum telephone interviewing mode for all interviews except for waves 1, 2, and 6. The personal visit interview in Waves 1, 2 and 6 will help FRs build and maintain rapport with respondents and, hence, provide high response rates.

5. Missing Wave Data Experiment

A household responds to a wave if at least one member of the household responds. We can classify households into three categories with respect to their response behavior:

a. households that respond to zero waves

b. households that respond to some but not all waves; and

c. household that respond to all waves.

We use weighting to adjust for the second category in cross-sectional estimation. However, persons in this category can be handled by imputation or as noninterviews in longitudinal weighting (Singh *et al.* 1990). However, given the analytical importance of transitions and spell lengths, we are uncertain whether missing items should be imputed for whole households.

The SIPP provides the opportunity to obtain retrospective data for the missing wave if a subsequent wave interview is completed. Therefore, the missing-wave-data experiment was conducted to see if retrospective data on transitions would greatly enhance imputation of income and the SIPP longitudinal estimation procedures.

The Bureau collected missing-wave data at the end of the regular interview using an abbreviated questionnaire entitled "Missing Wave Section". This section contained a skeleton set of the SIPP core questions dealing with labor force status, receipt of income, assets, and program participation. Due to concerns about the potential for adverse data quality effect of the longer recall, the missing wave questionnaire was used for persons whose response pattern over a set of three consecutive interviews was response-nonresponse-response. This limited the recall length to a maximum of 8-month. We introduced the missing-wave section in wave 4 of the 1984 panel. The data from the last interview of each rotation in the 1984 panel was analyzed to judge the utility of the missing wave data. Analyses focused on transitions in receiving income, assets, and government assistance. The following is a brief summary. For details, see Huggins (1987).

The missing wave section questions detected only a small number of changes in receipt of income and assets. 512 persons were eligible to respond to the missing wave. 38 of these reported a transition in receipt of one income type and one person reported a change in receipt of two income types. Only two and four changes in receipt of AFDC and social security, respectively, were reported. One person reported a change in alimony payment, 68 persons reported a transition in one asset and only one person reported transitions in two assets. These transitions picked up from the missing wave form data compared to estimated benchmarks were substantially lower.

Since it detected a proportionately small number of transitions in receipt of income and assets, the missing-wave form will not enhance our imputation on longitudinal weighting significantly. Furthermore, the effect of the number of transitions lost due to not collecting missing wave data for imputation and weighting should be negligible. Hence, the Bureau discontinued collection of the missing wave data in 1988.

6. Eighth Interview

We conducted the Eighth Interview (also called Wave-8) Study in August 1988 to determine if we could reduce the seam problem in SIPP by using alternative interviewing procedures. The seam problem is an over-reporting of changes in income sources and amounts between consecutive months covered by two consecutive interviews and under-reporting of changes between other consecutive months.

The Eighth Interview Study (EIS) used a subsample of rotation 4, 1986 panel households interviewed in Wave 7. Data for the study were collected when these households would have been interviewed for a Wave 8 interview using the core section of the Wave 7 questionnaire. Households in the EIS were interviewed by one of three interviewing procedures: R, B, or W. Households assigned to procedure R were interviewed using standard interviewing procedures and were treated as a control group. With the B procedure, the respondent received a feedback form containing his/her monthly income amount responses from the Wave 7 interview. For procedure W, the FR strongly encouraged the respondent to use records throughout the interview. We believed that procedures B and W could improve the accuracy of reporting the timing of transitions.

The analysis of the EIS data did not suggest that procedures B or W reduce the transition problem or that they should be implemented. We expected that procedure B would decrease between wave change, but it resulted in a higher observed rate than the regular interview procedure. Procedure W was expected to increase within wave change. However, the percent of within wave change with procedure W was not significantly different from the corresponding percents with procedures R and B.

Procedures B and W resulted in lower interview rates than procedure R (Gbur 1990). However, interview rates could be different if the field representatives and respondents had not already been exposed to the regular procedures. We held debriefing sessions with the FRs to discuss the Wave 8 experiment. According to FRs, procedure B (the feedback procedure) seemed to be improving the quality of the data (Singh 1988). In some cases, it encouraged the respondents to check records and helped them to provide better responses.

Reaction to procedure W was somewhat disappointing because the FRs felt that it was similar to what they did on regular SIPP. It appeared that the experiment was not carried out the way we intended. Perhaps classroom training instead of a self-study should have been given to emphasize more the use of records by the respondents and emphasize differences in procedures.

7. Time-line Calendar Aid to Respondents

The experiment is an extension of procedure B of the Wave-8 experiment. We used the time-line (event) calendar to display the responses given during the previous interviews for some of the government-program events in the SIPP. Displaying will help respondents to remember events and place them accurately. This in turn will reduce the seam problem.

We conducted the experiment in the Chicago Regional Office (RO) during 1989 panel interviews. For the experiment, the FR handed the respondent the calendar of the person he/she is responding for before beginning the interview. The FR explained to the respondent the purpose of the calendar. After each interview, the FR updates the calendar to reflect the current interview recipiency status and income amount by source.

The preliminary analysis (Kominski 1990) suggested that the time-line calendar has the potential to reduce the seam problem. Kominski also points out that the calendar facilitated longitudinal editing and correction of data either in the current or an earlier wave. In early 1993, Kominski's research will provide more results. These results will help us determine whether to use the approach.

8. Cognitive Research

The seam problem affects most of the SIPP transition and spell estimates. The Bureau initiated cognitive research to reduce the seam effect. As a part of this initiative, Cantor *et al.* (1991, 1992) used the "think aloud" approach to learn about reporting of recipiency and amounts, asked respondents to paraphrase questions and debriefed them to assess respondent's comprehension and understanding of technical terms and acronyms (such as AFDC, SSI) used in the SIPP and to elicit more detailed recall information.

Cantor's research provided a foundation for the Bureau's cognitive research. The Bureau planned its research in three phases. Moore *et al.* (1992) presented the Bureau's research and I will not discuss it in this paper. The planned research at the Bureau will continue for the next several years.

9. Debriefing of Respondents

We believe increased respondent cooperation and use of records during interview will give accurate data and, hence, will increase data quality for micro-level estimates. Therefore, in 1986, the Bureau conducted an experiment to learn from respondents: 1) Why they do not use records during SIPP interviews, 2) reasons for their cooperation and participation in the SIPP, 3) how knowledge gained in earlier interviews by participating in SIPP effects respondents' behavior for later interviews.

As part of the SIPP quality control program, we reinterview some of the respondents after each month's interview. For this experiment, however, we used SIPP reinterview sample assigned for the last 3 months of the 1985 panel to save cost and get results in a shorter period of time. A sample of size 516 households was eligible for the debriefing questionnaire. Only one person per sample household was eligible for reinterview. The person responding to reinterview may have differed from the person who responded in regular interview.

The response rate in general was high. Overall 89.5% persons responded to the debriefing questionnaire. The respondents seemed to like the debriefing. About 60% of the respondents reported using bank records and pay stubs during regular interviews. About the same percent of respondents reported using W-2 and 1986 tax forms during their regular interview. These record use rates from the debriefing were much higher than those reported in the regular interview (about 30%) and raised questions about accuracy of responses to debriefing questions. This large difference between debriefing and regular interview record use rates could be due to the different procedures used or perhaps a different interpretation of questions than intended.

About 2.2 percent of the respondents (12% of the persons on government programs) claimed that they started participating in the government programs after learning of a program through the SIPP. This learning effect caused a change in respondents' behavior and positive bias in our estimates.

The most frequent reasons respondents reported for participating in the survey were "like interviewer(s)" and "patriotic duty".

About 80% of the respondents who did not use records reported that one of the main reasons they did not use records was either "too much effort," "records not available" or "knew information without referring to records." Eighty percent of those who did not use records claimed that nothing could be done to encourage them to use records. For more details, see Petroni *et al.* (1989).

The findings of the experiment led to incorporating a statement into the "Dear Friend" letter that is sent out to the respondent prior to each interview asking them to refer to records. It also led FRs to call respondents before the interview to gather their records and to the initiation of cognitive research.

10. Income Source Summary (ISS) Record Checks

The field staff of the Census Bureau believes that training a respondent to use records during the first interview will increase their record use in SIPP. Therefore, we experimented training the respondent for greater use of records during the SIPP interview.

We emphasized the importance of record use to FRs during the refresher training for wave 1 of the 1990 panel and asked them to urge greater use of records by respondents. During wave 1, we also asked FR observers (usually a senior field representative) to note the use of records by respondents for certain questions on government programs and record their observations on the "SIPP Record Use Study Form". In later interviews, FRs recorded record use by income source on the ISS page of the questionnaire. We continued to emphasize the importance of record use to FRs and we believed recording such information on the questionnaire would not only provide a better estimate of record use but would also send a stronger message to FRs that record use is an important aspect of the SIPP. This would also encourage FRs to make an extra effort to train respondents during the wave 1 interview.

Evaluation of the data indicated that record use during wave 1 of the 1990 panel was low (Kominski 1991). Data from later waves showed record use rate at about the same level. More recent evaluation (Lessard 1992), based on wave 5 of the 1990 panel, showed that the average national rate of record use per income source was 21.44% and the national rate of persons using at least one record was 22.53 percent. The rates for other waves and panels were similar.

In cognitive research, about 70 percent of the respondents used records by income source. This is much higher than in regular SIPP interview rate. Since it is essential to increase record use in SIPP, we are continuously reminding interviewers in training sessions and memoranda to encourage their respondents to use records. In the future, we plan to analyze the rates at the FR level to provide them feedback on record use rate of their respondents. Feedback at the individual level may increase record use.

4. SUMMARY AND CONCLUSIONS

The SIPP experiments contributed significantly to the investigation of potential problems, providing data about the specific SIPP related issues, and helping us focus our research. We continue to analyze some of the SIPP research experiment data. The research experiments conducted at the Bureau contributed significantly to survey methodology, especially, for longitudinal surveys. It provided significant insight about various longitudinal survey related issues and assisted us in improving SIPP procedures and questionnaires. Some of these improvements and changes were discussed above.

The asset feedback experiment has shown that feeding back information on asset and liability data collected a year ago did not provide statistically different results from the data collected without feedback. Also, it suggested the changes in assets and liabilities are small during a short-term period. Research should be conducted to determine if 1) it is worth collecting assets and liability data more than once for a longer panel, 2) feeding back information will improve data quality of estimates of change over a longer period.

Reducing the seam effect is very important for the SIPP. The time-line calendar (Kominski 1990) and cognitive research, Moore *et al.* (1992) suggested that the seam problem can be reduced. Cognitive research identified the concepts and questions that are difficult for the respondents to comprehend. Also, it showed that record use by respondents can be increased drastically. However, current response rates for cognitive research are 80% or less for wave 1 which will decrease as the panel ages. This compares to a current wave 1 rate of 93%. Cognitive research results could have significant positive impact in improving data quality if the response rate is much higher than the current rate. Therefore, it is important to conduct research to identify specific elements (such as: increase the response rate, group interviews and increase record use by respondents) of the cognitive research that improve data quality. Research should also be conducted to use MTI with cognitive approach.

Ninety-six percent of the employers returned the "Employer Questionnaire and Authorization Form", while only 40 percent of the respondents signed this form. This experiment provided the very valuable information that surveyors can obtain supplemental data from employers successfully. However, extra effort is needed to get respondents to sign the form. In addition, research should be conducted to supplement survey data with other administrative record data. Supplementing the survey data will greatly enhance data base with good quality data without increasing response burden significantly.

Giving a token gift has a positive effect on the SIPP response rate. However, this effect is small and can be achieved by some administrative actions. Research to evaluate the effect of a higher-value gift or giving gifts twice or more should be conducted.

MTI research showed that a survey which is complex and sensitive in nature can be successfully conducted by maximizing telephone interviewing. MTI had minimal effect on cross-sectional and longitudinal estimates. The SIPP started using maximum telephone interviewing in February 1992. We are monitoring data closely. We have not observed any significant problem yet except that the respondent use of flashcards has decreased from 100 percent to 20-30 percent. Research to increase use of flashcards and records use in MTI should be conducted.

Furthermore, research into centralized computer assisted telephone interviewing should be conducted to make SIPP operations more flexible without the loss of data quality. Flexibility will help manage fluctuating work load in the field.

Collecting data for the missing wave in a subsequent interview was considered a good alternative but did not provide good quality data for the collected items, probably due to a longer recall (5 months to 8 months) period. We should conduct research using alternative approaches with more feedback or probing. One approach worth researching is to collect data up to the interview time and use it for probing at the next interview. Another is to provide more feedback from the previous interview. Computer Assisted Personal Interview (CAPI) will be a real asset in researching these approaches. The research in these areas will also have the potential to reduce the seam effect.

ISS showed much lower record use in SIPP as compared to cognitive research. Research should be conducted to see if increased record use improved the data quality. If so, alternatives such as a special FR training, modifying FR performance standard to include record use, accepting lower response rate for higher record use, *etc.* to increase record use should be researched.

ACKNOWLEDGEMENTS

I gratefully acknowledge Sandy Carnegie and Kathy Kreilick for typing support and Rita Petroni, Andrea Meier, Daniel Kasprzyk and Nanak Chand for their helpful comments to improve the paper.

REFERENCES

Adams, D. (1988). SIPP 85: Evaluation of the employer-provided benefits study. Internal memorandum to Shapiro from Singh, dated October 28, 1988.

Butler, D. (1991). SIPP 87: Gift experiment results. Internal Census Bureau memorandum to Singh dated April 2, 1991.

Cantor, D., Brandt, S., and Green, J. (1991). Results of first wave SIPP interviews. Memorandum to Bowie, dated February 2, 1991.

Cantor, D., Green, J., Moesinger, K., Brandt, S., and Rose, P. (1992). Revised draft results of second wave of SIPP interviews. Memorandum to Lampe dated September 9, 1992.

Durant, S., and Gbur, P. (1988). Testing telephone interviewing in the survey of income and program participation and some early results. SIPP Working Paper Series No. 8824, U.S. Bureau of the Census.

Gbur, P., (1990). SIPP 86: Eighth interview study data analysis. Internal Census Bureau memorandum for Shapiro from Singh dated May 9, 1990.

Gbur, P., and Petroni, R. (1989). Preliminary evaluation of maximum telephone interviewing on the SIPP. *Proceedings of the Survey Research Section of the American Statistical Association*.

Gbur, P., Cantwell, P.J., and Petroni, R.J. (1990). Effect of maximum telephone interviewing on SIPP topical module and longitudinal estimates. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Huggins, V. (1987). Evaluation of missing wave data from the survey of income and program participation. *Proceedings of the Section on Survey Research Methods Section of the American Statistical Association.*

Kominski, R. (1990). The SIPP event history calendar: Aiding respondents in the dating of the longitudinal events. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Kominski, R. (1991). Record use by respondents. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Lamas, E.J., and McNeil, J.M. (1987). An analysis of the SIPP asset and liability feedback experiment. SIPP Working Paper Series No. 8725.

Lessard, J., (1992). SIPP 90: Wave 5 results of the record check study. Census Bureau's Internal memorandum for SIPP Research and Evaluation Steering Committee from Singh, dated June 15, 1992.

McNeil, J., (1989). Quarterly estimates of core characteristics: 1984, 1985, and 1986 panels. Internal Census Bureau memorandum for The Record, dated July 21, 1989.

Moore, J., Bogan, K., and Marquis, K. (1992). A "cognitive" interviewing approach for the survey of income and program participation: Development of procedures and initial test results. Presented at the Symposium 92: Design and Analysis of Longitudinal Surveys held at Statistics Canada, Ottawa, November 2-4, 1992.

Nelson, D., McMillan, D., and Kasprzyk, D. (1985). An overview of the survey of income and program participation. SIPP Working Paper Series No. 8401, Update No. 1.

Petroni, R.J., Huggins, V.J., and Carmody, T.J. (1989). Research and evaluation conducted on SIPP. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census.*

Singh, R. (1988). General comments based on wave 8 experiment debriefing. Internal Census Bureau Memorandum for The Record, dated September 23, 1988.

Singh, R., Huggins, V., and Kasprzyk, D. (1990). Handling single wave nonresponse in panel surveys, SIPP Working Paper Series No. 9009, U.S. Bureau of the Census.

Weidman, L., King, K., and Williams, T. (1988). Further evaluation of the SIPP asset feedback experiment. *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

METHODOLOGICAL ISSUES IN THE DESIGN OF THE BRITISH HOUSEHOLD PANEL STUDY

P.C. Campanelli and L. Corti[1]

ABSTRACT

There can be no doubt that household panel studies offer unique opportunities for a range of important and innovative methodological research projects. At the same time, panel studies offer difficult design and quality challenges and the complex nature of their data creates new puzzles and problems for both substantive researchers and methodologists. This paper describes how the British Household Panel Study (BHPS) has come to terms with various design issues. The BHPS is a newly formed multi-purpose, multi-wave panel survey of 5,000 households in Britain. Discussed are decisions made by the BHPS with respect to its initial design work, current quality control issues and programme of methodological research.

KEY WORDS: Panel studies; Surveys; Methodology.

1. INTRODUCTION

One type of longitudinal design which has become increasingly popular is the general-purpose, national household panel study. An exemplar for this type of study is the Panel Study of Income Dynamics which started in 1968 at the University of Michigan (see Morgan and Duncan 1986). National panel studies now exist or are being developed in most European countries (including Belgium, France, Germany, Greece, Hungary, Ireland, Luxembourg, the Netherlands, Spain and Sweden). All of these studies have broadly similar aims without being identical. The British Household Panel Study (BHPS) is one of the more recent additions to this line of European national studies.

1.1 The British Household Panel Study

The BHPS is the largest single project ever funded by the UK Economic and Social Research Council and was designed to provide unique analysis capabilities to users in both academic and government settings. It is based at the ESRC Research Centre on Micro-Social Change on the campus of the University of Essex with an interdisciplinary staff of 41 people. Interviewing on the survey began in 1991 and is scheduled, pending review, to continue in annual waves until at least 1998. The achieved probability sample of 5,538 households and 10,303 individuals was drawn from across Great Britain. The survey instrument, which is administered to all adult members of the household, comprises a face-to-face 45 minute interview with every individual, a short self-completion schedule, as well as a short household level questionnaire. These cover various substantive areas; household organization, income and wealth, labour market experience, housing costs and conditions, health issues, consumption behaviour, education and training, and socio-economic values. The data are designed to allow researchers to describe and analyze how individuals, families and households experience changes in their socio-economic environment and how they act in relation to these changes. More information on the objectives of the BHPS can be found in Rose *et al.* (1991).

[1] P.C. Campanelli and L. Corti, The British Household Panel Study, ESRC Research Centre on Micro-Social Change, University of Essex, Colchester, UK CO4 3SQ.

The main purpose of this paper is to give an overview of some of the key methodological issues arising in the design and operation of household panels and the way these are being tackled on the BHPS[2]. It will cover methodological and design issues in a very broad, multidisciplinary sense with examples from all aspects of the survey process from initial design through to and including dissemination. The paper is divided into three sections: Initial, current, and future issues.

2. INITIAL METHODOLOGICAL ISSUES

Starting a new panel study raises several fundamental issues which stem from its theoretical purpose in connection with cost considerations. These include what type of panel design should be adopted; what will be the interval between waves, mode of data collection, respondent rules, sample size, and type of sample design; what topics will the questionnaire cover and how will this be done; how will issues of nonresponse be countered, *etc.* The broad objective of the BHPS was to further our understanding of social and economic change at both the individual and household levels in Britain through the 1990s. For the BHPS, design decisions, research goals, and cost considerations interacted in the following ways:

The Type of Panel Design

Several alternative designs are available for a repeated survey of individuals (see Duncan and Kalton 1987; Kalton 1992). These include the cohort, rotating panel, split panel and repeated cross-section designs, as well as the classic panel design consisting of a representative sample of individuals or units who are followed over a series of waves. The PSID and similar European national studies tend to be of this latter type. The BHPS also adopted this design, which brought us in line with these other household panel studies, and allowed us to meet our objectives to measure various components of individual change and identify transitory and persistent phenomena.

Interval Between Waves

There are many persuasive reasons for spacing interviews 12 months apart (see Rose, Buck and Corti 1991; van de Pol 1988). Shorter intervals than 12 months mean more frequent waves resulting in higher field costs and can lead to a compensating sample size reduction, as well as greater organizational and respondent burden. Similarly, shorter intervals may not allow adequate time for certain changes to occur. On the other hand, concerns about recall error argue against longer intervals (*e.g.* every 2 years). The 12 month compromise position has the added advantage of being a meaningful bounding concept for the respondent. Therefore it is not surprising that the BHPS, the PSID and several European studies chose a 12 month interval.

Mode of Data Collection

Face-to-face interviews were conducted for the first wave of the survey. Several factors were implicit in the decision not to switch to telephone interviewing for later waves despite possible savings. A switch to the telephone would have meant procedural and questionnaire changes and the possibility for mode effects between waves. Similarly, the telephone is not a well accepted mode of interview for surveys in Britain.

Respondent Rules

Research objectives to explore intra-household structures and processes, explore values and attitudes in relation to behaviour, and the complex nature of some of the questions, influenced the decision to conduct face-to-face interviews with every adult member of the household, rather than only a household head as is done in the PSID, and accept proxy reports only as a last resort. This decision was considered essential, despite the major cost implications.

[2] Fuller descriptions of BHPS issues and work can be found in Rose, Buck and Corti (1991) and Rose, Campanelli, Corti and Taylor (1992).

Sample Size

A sample size of 5,000 households and 10,000 individuals was considered a minimum in order to provide the desired precision for estimates for both the population as a whole and major subgroups of interest such as, one parent families and the elderly. This comparatively large sample (by academic survey standards) was also needed to generate sufficient events over time for the analysis of transitions and to allow for attrition.

Type of Sample

Our early work was dominated by the trade-off between sample design needs and the requirements of certain longitudinal models. An intense debate arose among our panel design advisory group over the appropriateness of using a multi-stage clustered sample design, with calls by some econometricians and statisticians for a simple random sample (SRS) design as required by the statistical models they wished to use (see Coxon 1992). Of particular concern was the fact that the dispersion of clusters over time would pose immense difficulties for modelling. Those familiar with field costs of personal interview surveys will not be surprised that implementing an SRS design, even with the easy availability of the Postcode Address File (PAF) in Britain to use as a frame would increase fieldwork costs by 30-40% due to interviewer travel alone. On the other hand, the large economic gains from a clustered sample are offset by only a small loss in standard error efficiency. Furthermore, with a fixed budget for the BHPS, an SRS scheme would have reduced the sample to a size for which many types of statistical analysis would have been critically affected. Ironically, it would have been impossible to implement a truly 'cluster-free' design as clustering due to interviewers and individuals within households would remain.

What Topics to Cover and How to Cover Them

We were committed to certain core substantive areas by our funding agency contract. However, this still left room for debate on the nature and number of questions to include under each topic. Four additional issues were considered important:

1) The Need to Complement Other Major Datasets

Replicating questions can allow comparison with other data sets for cross-validation purposes, as well as comparative analysis, both with cross-sectional and longitudinal data from other countries. The BHPS has, therefore, been designed to maximize the possibility of data linkage, both with other British academic and government surveys and also with other household panel studies in Europe and the US[3].

2) The Need for Continuous Measurement

One of the values of a panel study is its ability to collect measures of change more reliably than through retrospective histories. Many questions, however, have to be concerned with events between interviews and are thus retrospective in nature. Collecting such continuous measures is of key importance (Duncan 1992). Our aim was to produce questions that would enable us to construct continuous measures of income, employment histories and labour market participation, household structure and residential mobility over the life-cycle.

3) The Need for a Variable Component

Not all of the desirable questions can be asked at one time. Dividing the questionnaire into a "core" and "variable" component allows for the addition of more questions and more importantly allows for flexibility in including questions in the future that may be relevant to changing needs as seen at that time. It does, however, raise the issue of determining the proper frequency for intermittent questions.

[3] Question origin, among other question characteristics, is fully documented.

4) **The Need for Special Design Features**

An example of a special design issue is how to establish the relationship of everyone in the household to everyone else in a fully reliable way without reverting to a full matrix design (see Brynin 1992).

Nonresponse

Unit and item nonresponse bias are key concerns for any survey and take on special meaning within a panel study. In order to ensure continuing representativeness, the BHPS made great effort to achieve a high first wave response rate. We were fortunate enough to obtain interviews (either directly or via proxy) with ALL members of a household in 69 percent of all eligible households. This is remarkably successful for a British non-government survey, where a good response rate for obtaining AT LEAST ONE individual in a household is in the region of 65-70 percent.

Our focus was to use well-researched fieldwork practices to enhance response rates and motivate interviewers as well as to consider questionnaire design and respondent tracing strategies to help to minimize both item and unit nonresponse. One example is an early incentives split-ballot experiment designed to investigate the effects of individual incentive payments (a gift voucher redeemable at a national chain store) on initial response. Because of interviewer variation, results have to be interpreted with some caution, but the incentive appeared to have a positive effect on response. This result, plus evidence and advice from other studies (*e.g.* Department for Statistics of Income and Consumption 1984; Jean and McArthur 1987) suggested the adoption of respondent incentives for all fieldwork. A further example, is the design controversy which surrounded the question of how to implement a detailed set of personal income questions with out incurring item or unit nonresponse. After consulting with a panel of experts, the BHPS took an intermediate option which sought a trade off between chronological depth and detail of information on the most recent period. For example, questions on wealth and assets, considered to be the most sensitive, would be asked at a future wave when panel members would be sufficiently committed to the survey. Similarly we integrated questions on earnings with other measures of job characteristics to make them less obtrusive. Although a split-ballot experiment is needed to evaluate these practices, we achieved income nonresponse rates in Wave 1 similar to government surveys of around 10 percent. This figure is encouraging for an academic survey.

Implementation of Data Collection

There were also issues surrounding how our data collection would be accomplished. Could we afford to set up and maintain our own interviewing staff? If we subcontracted the work, how could we ensure quality? The quality issue became critical as our chosen field agency was more familiar with market than academic research. Therefore a series of survey specific procedures were built into the contract to assure quality. Several of these were standard survey practice, such as thorough vetting of interviewers, targeting perceived areas of interviewer weakness for specific training, use of scripted dummy interviews during briefings, early accompaniments of interviewers, bi-weekly chases of interviewer progress, refusal conversion procedures, postal and telephone reinterviews on a sample of respondents, *etc.* Others were more innovative such as the use of special training videos (see Smith 1992) and training all interviewers to tailor their doorstep introductions to meet the special circumstances of the respondent (see Groves and Cialdini 1991).

3. CURRENT METHODOLOGICAL ISSUES

The second wave of fieldwork brought new implementational and quality control issues. In addition we wanted to establish methodological research in its own rig'

3.1 Implementation Issues

Following Rules

The classic panel study design requires good following rules to indicate which units should be retained at each wave, what new units should be added and which units should be dropped (Duncan 1992). In simple terms, we wanted to follow original sample members over time even if they split-off to from new households and we wanted to add new sample members from birth and marriage, so that the sample would maintain some representativeness. However, the issue quickly becomes complex. One example is when 'sample' and 'nonsample' member status is cross-classified with whether the person has been a respondent or a nonrespondent. For example, does one go back to a household where nonsample members cooperated, but original sample members did not? There is also a practical limitation on which original sample members can be followed. Our Wave 2 rules only excluded original sample members who moved out of the country (much to the disappointment of our eager interviewers) or who moved into prison or mental hospitals (other institutions were included).

There is also the "household versus family" issue. Cost savings can be gained by restricting interviews to a 'family' unit from Wave 2 onwards as the strictly economic roles, such as those of boarder and lodger are of little research interest. This would require the establishment of a 'family' definition which is coherent and can be used in the field. However, such a definition seems inherently problematic, as the criteria for inclusion in a family would have to be the nature of the current relationships and an assessment of their likely permanence. For Wave 2 we employed the standard household definition to determine who should be interviewed at an address containing original sample members. The 'family' concept could, however, come into play in the following rules starting with Wave 3 as certain categories of 'nonsample-nonfamily' members who split to form separate households may not be followed.

Coversheet Design

There were several unexpected issues encountered in the design of a coversheet for Wave 2. The first issue originated because the survey was carried out via personal interview: essentially, how does one physically supply the information that this year's interviewer needs, such as addresses and basic demographics? Should it be put on a separate document, printed onto labels to fit onto the coversheet, or printed directly onto the coversheet itself? A second issue was whether individual or household level coversheets were required. However, the concept of a longitudinal household is futile, as from Wave 2 onwards the sample is in essence a sample of individuals. On the other hand interviewers still deal with households. Individual level coversheets would have produced a proliferation of documents for interviewers and procedures to handle new entrants would have become complicated. Eventually, the household level coversheet was adopted. A third issue was how to identify the 'expected' members of each household. This was not always as straightforward as it would seem. For example, if based on a confirmation of address card we found that a marital split had occurred, we assumed that the children went with their mother and listed them as expected members in their mother's household. If, however, the interviewer found them at their father's household, they would appear to that interviewer as "new entrants" because they would not have been pre-listed. Without a special mechanism to handle such occurrences, major data linkage errors could be created.

Feeding Forward of Information (Dependent Interviewing)

Simply asking questions afresh at each occasion can result in an over-estimate of true change. Take for example, the well known spurious change noted in CPS gross-flow data on occupation and industry (Collins 1975). There are also the perplexing seam problems encountered by the SIPP and PSID panels (Moore *et al.* 1992; Duncan and Mathiowetz 1985). This is a tendency for respondents to *over-report* changes in status and in amounts received between adjacent calendar months included in the reference periods for different interviews, and to *under-report* changes between months covered by the reference period for a single interview, *i.e.* a tendency for changes to cluster at the seams between waves. A procedure to minimize these difficulties is to feed forward a respondent's answer from the previous wave and then ask the respondent if there have been any changes.

Feeding forward information has several advantages including reducing measurement error and spurious change, reducing respondent and interviewer burden, cutting costs through reduced interview time and coding, providing a bounding period for recall of recent events, assistance in stimulating recall, and helping respondents to be more thoughtful about their answers. On the other hand, there is the danger of reducing estimates of true change (it is easy to say that nothing has changed); of giving the illusion of broken confidentiality; negative cooperation; and there are also cost and complexity issues. A full discussion of feeding forward issues and practices is given in Corti and Campanelli (1992).

For Wave 2 of the BHPS, we decided on a limited amount of feed forward information. This included basic recontact information such as respondent address; call records for multiple callback households; and enumeration rosters with demographic details. We also included a special check question to indicate whether the previous year's information was correct or incorrect in order to help explain any cross-year inconsistencies. The cost and the complexity of designing a paper and pencil questionnaire which would allow for feed forward data for original sample members, as well as accommodating new entrants, dissuaded us from feeding forward substantive information until we move to CAPI.

The BHPS does, however, include a built in overlap in reference periods. For example, the fieldwork period for both the first and second wave is essentially 4 months long (September through December) and in each case respondents are asked about changes since September of the previous year. Thus for a respondent interviewed in October both years, September information is available with 13 month recall and duplicated with one month recall. This can be used to disentangle duplicate reports. Similarly, the BHPS has a built in check on occupation. Although respondents are asked afresh about their occupation at each wave, they are also asked when they started work in their present position. Thus we can compare their conceptions of whether or not their job has changed with the differences in their occupational descriptions, holding constant variation due to coding interpretations.

Occupational Coding (CASOC Test)

CASOC or Computer-Assisted Standard Occupational Coding was recently developed in the UK between the universities of Warwick and Cambridge. The CASOC software can be used in both computer-assisted and computer-automated coding and has several associated utility files which automatically recode the detailed output to a wide variety of occupational and social classifications. We conducted our own CASOC experiment to compare computer-assisted to manual coding of occupations at the 3-digit level in terms of both reliability and validity. A random sample of 325 descriptions of the respondents' own occupation was independently recoded 8 times: 4 manually by coders with different levels of experience, 3 by the same coders using CASOC in computer assisted mode, and 1 by a coder using the software in fully automated model. The final addition is an 'expert coder' column which will be used as a model of 'truth'. We are still in the process of analyzing the data, but some of the early results suggest fairly high reliability with average agreement rates among the manual coders of .79 compared to .82 for the computer-assisted mode.

Database Design, Documentation and Dissemination Issues

Other important issues surround the best type of database design for longitudinal data which will satisfy a list of conflicting needs such as ease of access for processing, ease of data manipulation for particular kinds of analyses, efficiency of storage, *etc.* As with several other household panel studies, we chose the Scientific Information Retrieval (SIR) relational database software.

Documentation and dissemination practices can also bring up methodological issues. One must remember the need to develop ways of assisting others in the use and exploitation of panel data. This implies a requirement for carefully researched documentation and dissemination practices and good training policies, especially for young researchers, as large and complex datasets often deter users because of both substantive complexity and technical intricacy. As part of this process the Centre has developed a question history bank, computer interfaces for users, established a series of training seminars (in conjunction with two other universities), and established its own *Research Resources Unit* which plays a central role in the management of information within the Centre and provides a variety of services to its staff and to the wider research community.

3.2 Quality Control and Sources of Error

The prime concern for any survey must be the quality of the data at all stages of the survey process. A more general concern is to investigate the extent of various kinds of sampling and nonsampling error, where possible, so that this is documented for future users.

In addition to monitoring our contract agency, the Research Centre was involved in pre- and post survey projects. For example, in addition to standard pretesting techniques, some of the pretests made use of special pretest methods such as the coding of interviewer and respondent interactions (Cannell *et al.* 1989), interviewer debriefing and field-based respondent debriefing studies (Campanelli, Martin, and Rothgeb 1991); and split ballot studies.

Considerable effort was also put into ensuring that the collection of income data was as accurate as possible. Respondents were asked to consult documents where possible. Tax codes were collected where pay slips were available to check the accuracy of respondent reporting. Similarly we had special concern over the implementation and evaluation of panel maintenance and tracking and tracing procedures and a special computer programme was written to assist in this task.

After Wave 1 data collection was complete we were also involved in several standard, although not always straightforward steps, including quality checks on the editing and coding process, the cleaning of the data, the development of derived variables, weighting and imputation issues, and the analysis of nonresponse. As part of Wave 2 data collection, we are attempting to convert within household refusals and to find and interview Wave 1 "noncontacts." As part of Wave 2 analysis, we will investigate panel attrition issues.

A major project to document sources of error was our plan to measure **interviewer variance**. Such work requires the randomization of the allocation of interviewees to interviewers. This method was pioneered by Mahalonobis (1946) under the name of 'interpenetrating samples.' Due to field requirements and travel costs, we adopted a constrained form of randomization in which addresses were allocated to interviewers at random within geographic "pools". For example, all pairs of PSU's within ten kilometres of each other were uniquely clustered. Within a given cluster, addresses were randomly assigned to interviewers. Plans are to document the extent of variability between interviewers and relate it to the characteristics of the questions asked, the characteristics of the interviewers themselves and most importantly to some measurement error indicators that will be available on the data such as recall discrepancies. As this has been implemented as part of Wave 2, an implicit component of the test will be analyzing the effect of sending the same or a different interviewer to a respondent in the second year of a panel study. We look forward to analyzing these data next spring through variance components models using new hierarchical software available in the UK called ML3 (Goldstein 1991).

3.3 Methodological Research

Projects discussed in this section represent those which will hopefully serve to further our knowledge of the advantages and limitations of panel data and stem mainly from the interests of research staff members.

Response Contamination

Are survey interview responses affected by the presence of third parties? For example, would a wife be loath to talk about her previous boyfriends and provide accurate data if her husband is present in the room? This type of response contamination is of major concern to all surveys, but it is especially relevant to household studies, where other members are often present when respondents are interviewed. Similarly, one could also imagine the situation of differential response error over time as the third parties who are present can change from wave to wave. The BHPS questionnaire allows for this kind of research by including an interviewer check question at the end of each section of the questionnaire to indicate who was present. In analysis of these data from the first wave, Corti and Clissold (1992) found some evidence to suggest that the presence of other persons at an interview may indeed effect the answers of the respondent, particulary to sensitive questions.

Calendars and Recall

As mentioned earlier, the plan to collect continuous measures of change requires that many of the survey measures are concerned with retrospective accounts of change during the past year, rather than with the current state at the time of interview. The issue of designing a questionnaire to capture this depth of data to a reasonable degree of quality was one of concern and debate. During the two years of intensive planning for the survey, a number of pre-tests and pilots were carried out to test various ways of collecting this type of recall data including structuring the direction of recall in different ways, that is from past to present or present to past, making use of visual recall aids, integrating the data collection with other life events, and capturing details of change using free but chronological recall or using a month by month, or week by week, accounting calendar, such as those used by other European panels (see Corti 1992).

Analytic Strategies to Account for the Complexity of the Survey Design

Data obtained from complex sample survey designs have traditionally been analyzed using a variety of methods (such as Taylor Series Expansion, balanced repeated replications, jackknife repeated replications) to adjust for the effects of the design on standard errors (see, for example, Kish and Frankel 1974). Recently, multilevel models have been suggested as a possible alternative framework for the analysis of these data (Goldstein 1991). Multilevel models have the advantage of allowing for the explicit substantive investigation of the hierarchical structure of the sample data and may provide more efficient estimates than traditional approaches. Taylor and Campanelli (1992) have begun an empirical comparison of these two approaches for normal and logistic regression models. The aim is to assess the possible advantages and disadvantages of using multilevel models rather than traditional variance estimation techniques. A number of models have been considered where there is a range of design effects for the dependent variable and a range of design effects for the independent variables. Early results suggest that traditional and multilevel approaches yield similar parameter estimates and model conclusions.

Qualitative Interviews

The interview itself is also an area which can be studied in its own right. One of the primary concerns of the Centre is to link the survey to other methods of data collection, such as more in-depth qualitative interviews. One of our Research Associates is currently undertaking a qualitative study of retirement using members of our pilot panel. This study will, therefore, be most useful as a test of the validity of our survey measures against those deriving from a life-history approach. In a similar vein, the household allocative systems project (see, for example, Laurie 1992; Laurie and Sullivan 1991) combines a qualitative approach with the more limited but more representative possibilities of the panel study in this area.

3.4 Substantive Methodological Research

Another strategy with respect to BHPS work has been to explore methodological problems by doing substantive research. Two examples of this type of work come from our staff's participation in the recent International Conference on Social Science Methodology in Trento, Italy. Buck and Scott (1992), for example, discovered some important methodological issues when using event history analyses. Their goal was to model young people leaving home. They discovered this can be an ambiguous concept. They found that slight changes in the definition of their dependent variable led to non-trivial differences in their substantive results. Another example comes from the work of Dex and Laurie (1992) who explored various methodological issues encountered in doing their cross-national comparative analysis of women's labour market behaviour.

4. FUTURE METHODOLOGICAL ISSUES

Several new projects are planned and are currently under review:

CAPI Feasibility Test

In the future we will be looking into the use of Computer Assisted Personal Interviewing (CAPI). CAPI is now in use or being tested on many studies throughout the world, including some highly complex studies (see Costigan and Thompson 1992). Not only will we have to look into implementation issues and complexities, there are data comparability issues. To what extent are computer-assisted data comparable to data collected with more traditional methods? A change to CAPI in the middle of a panel survey can confound true change with a possible mode effect (Olsen 1992). For example, respondents could be more candid under one mode (showing a difference in means over time) or reporting may be more accurate (possibly showing a reduction in variance).

The Reliability of Retrospective Data

Research documenting the quality and reliability of recall data is available (see review by Dex 1992), but for European panel studies it is extremely limited, in part because many of the national household panel studies have only been established since the early- to mid-eighties. The BHPS hopes to improve on this situation. The data offer two different types of opportunity to estimate the magnitude of memory error: 1) events from the distant past are asked at two different survey years and 2) as described above, there is an overlap between waves of the survey due to the length of the field period. This leads to some interesting opportunities for comparison. We can see whether the reporting discrepancies vary by type of question, topic of the question, difficulty of the reporting task, likely salience of the events in question, and various measures of respondent characteristics. We also hope to ascertain the effect of such errors on actual models and to tie this work in with a similar project proposed by other British researchers using 10 year recall data.

Analysis of Large and Complex Datasets Initiative

The UK Economic and Social Research Council has recently had a call for proposals for a research programme on the Analysis of Large and Complex Datasets. The Centre hopes to be involved in collaborative projects within this programme to develop practical guidelines for social scientists wishing to analyze complex survey data, ways to incorporate imputation 'flags' into data analysis, and expanding the capability of discrete-time survival models to cover multi-level random effects. Other projects ideas include the use of new technology in the coding of open-ended questions other than occupation and in a radical reconceptualisation of data dissemination.

5. CONCLUSIONS

Several important types of projects have not been mentioned this far. Of priority for the BHPS will be first a validation study to examine the accuracy of respondent self-reports and second more work focused on nonresponse. In addition it may also be of value to add a fresh cross-section to examine the effects of panel conditioning in the BHPS context. There are other items which apply to more than just the BHPS. As methodologists we would like to see work directed at questionnaire design issues in a longitudinal context. This would include determining what the value of a 'longitudinal question' really is and developing guidelines for designers and users for what to do upon discovering that a Wave 1 question is not working or has changed its meaning over time. Similarly, there is a need for database design work for particular types of analyses (*e.g.* event history analyses). We would also like to see some of the newer tools available to survey research such as cognitive interviewing techniques (Tanur and Fienberg 1992) and detailed coding of interviewer/respondent interactions (Cannell *et al.* 1989) applied to some of the specific problems of collecting longitudinal data. A precedent for this is Moore and colleagues' use of cognitive interviews to improve SIPP procedures (1992). Another example would be to investigate what cognitive laboratory work and behaviour coding could tell us about how, when and if respondents are "conditioned."

We hope this paper has given you a flavour of some of the work which is currently being carried out or is planned as part of the BHPS. As a relatively young panel, we could not have proceeded as far as we have without the unique knowledge acquired from other studies. We are very grateful to all here who have helped or advised us and look forward to your comments and suggestions.

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 programme of the ESRC Research Centre on Micro-Social Change in Britain.

REFERENCES

Brynin, M. (1992). Relating people by computer. In *Survey and Statistical Computing*, (Eds. A. Westlake, R. Banks, C. Payne and T. Orchard), Amsterdam: North-Holland.

Buck, N., and Scott, J. (1992). *Modelling household dissolution: An event history analysis of young people leaving home.* Paper presented at the International Conference on Social Science Methodology, Trento, Italy, June. Colchester: ESRC Research Centre on Micro-Social Change.

Campanelli, P., Martin, E., and Rothgeb, J. (1991). The use of respondent and interviewer debriefing studies as a way to study response error in survey data. *The Statistician*, 40, 253-264.

Cannell, C.F., Oksenberg, L., Kalton, G., Bischoping, K., and Fowler, F.J. (1989). *New Techniques for Pre-testing Survey Questions*, Research Report, Ann Arbor MI: Survey Research Centre, Institute for Social Research.

Collins, C. (1975). Comparison of month-to-month changes in industry and occupation codes with respondent's report of change: CPS job mobility study. *Response Research Staff Report No. 75-5*, Washington, DC: U.S. Bureau of the Census.

Corti, L. (1992). Calendar and life history recall aids. *Discussion Paper 2*, Colchester: ESRC Research Centre on Micro-Social Change.

Corti, L., and Campanelli, P. (1992). The utility of feeding forward earlier wave data for panel studies. In *Survey and Statistical Computing*, (Eds. A. Westlake, R. Banks, C. Payne and T. Orchard), Amsterdam: North-Holland.

Corti, L., and Clissold, K. (1992). Response contamination by third parties in a household interview survey. *Working Paper 13*, Colchester: ESRC Research Centre on Micro-Social Change.

Costigan, P., and Thomson, K. (1992). Issues in the design of CAPI questionnaires for complex surveys. In *Survey and Statistical Computing*, (Eds. A. Westlake, R. Banks, C. Payne and T. Orchard), Amsterdam: North-Holland.

Coxon, A.P.M. (Ed.), (1992). Sample design issues in a panel survey: The case of the British household panel study. *Working Paper 3*, Colchester: ESRC Research Centre on Micro-Social Change.

Department for Statistics of Income and Consumption, (1984). *Notes on the Planning and the Scope of the Socio-Economic Panel Survey*, Netherlands: Central Bureau of Statistics.

Dex, S. (1992). The reliability of recall data: A literature review. *Working Paper 11*, Colchester: ESRC Research Centre on Micro-Social Change.

Dex, S., and Laurie, H. (1992). Comparative analyses using large scale national data sources of women's employment. *ESF Working Paper 37*, Colchester: ESRC Research Centre on Micro-Social Change.

Duncan, G. (1992). Household panel studies: Prospects and problems. Paper presented at the International Conference on Social Science Methodology, Trento, Italy, June.

Duncan, G., and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 1, 97-117.

Duncan, G., and Mathiowetz, N.A. (1985). A validation study of economic survey data. *Mimeo*, Ann Arbor: Survey Research Centre, Institute for Social Research.

Goldstein, H. (1991). Multilevel modelling of survey data. *The Statistician*, 40, 235-244.

Groves, R., and Cialdini, R. (1991). Toward a useful theory of survey participation. In *Proceedings of the Section of Survey Research Methods*, Washington, DC: American Statistical Association.

Jean, A.C., and McArthur, E.K. (1987). Tracking persons over time. *SIPP Working Paper Series No. 8701*, Washington, DC: U.S. Bureau of the Census.

Kalton, G. (1992). Panel surveys: Adding the fourth dimension. In *Proceedings of Symposium 92: Design and Analysis of Longitudinal Surveys*, Ottawa, Statistics Canada, 99-109.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, 1, 1-37.

Laurie, H. (1992). Multiple methods in the study of household resource allocation. In *Mixing Methods: Qualitative and Quantitative Research*, (Ed. J. Brannen), Aldershot: Avebury.

Laurie, H., and Sullivan, O. (1991). Qualitative versus quantitative? The use of qualitative information in the British household panel study. *Sociological Review*, 39, 1, 113-130.

Mahalonobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

Moore, J., Bogen, K., and Marquis, K. (1992). A cognitive interviewing approach for the survey of income and program participation: development of procedures and initial test results. In *Proceedings of Symposioum 92: Design and Analysis of Longitudinal Surveys,* Ottawa, Statistics Canada, 31-40.

Morgan J.N., and Duncan, G.J. (1986). Experience with the panel study of income dynamics. In *Microanalytic Simulation Models to Support Social and Financial Policy*, (Eds. G.H. Orcutt, J. Merz and H. Quinke), Holland: Elsevier Science Publishers.

Olsen, R.J. (1992). The effects of computer assisted interviewing on data quality. *Working Paper 36*, Colchester: ESRC Research Centre on Micro-Social Change.

Rose, D. *et al.* (1991). Micro-social change in Britain: An outline of the role and objectives of the British household panel study. *Working Paper 1*, Colchester: ESRC Research Centre on Micro-Social Change.

Rose, D., Buck, N., and Corti, L. (1991). Design issues in the British household panel study. *Bulletin de Methodologie Sociologique*, 32, 14-43.

Rose, D., Campanelli, P., Corti, L., and Taylor, M. (1992). Methodology for household panels and longitudinal data analysis: Where are we and where do we go from here? *Discussion Paper 1*, Colchester: ESRC Research Centre on Micro-Social Change.

Smith, R. (1992). *Audio-visual aids in interviewer training.* Paper presented to the 47th Annual Conference of the American Association for Public Opinion Research, St. Petersburg, Florida, May. Colchester: ESRC Research Centre on Micro-social Change.

Tanur, J., and Fienberg, S. (1992). Cognitive aspects of surveys: Yesterday, today, and tomorrow. *Journal of Official Statistics*, 8, 1, 5-17.

Taylor, A., and Campanelli, P. (1992). Accounting for the complexity of the survey Design: A comparison of traditional design-based procedures to multi-level modelling techniques. Paper presented at the annual meeting of the Royal Statistical Society, Sheffield, England, September.

Van de Pol, F. (1988). *Design issues in panel studies.* Amsterdam: Sociometric Research Foundation.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

A LONGITUDINAL STUDY BASED ON A CONTINUING NATIONAL SURVEY: THE LONGITUDINAL STUDY OF AGING

M.G. Kovar[1]

ABSTRACT

A supplement to the National Health Interview Survey, a continuing survey of the civilian noninstitutionalized population of the United States, was used to obtain the baseline data for a longitudinal study. That study, the Longitudinal Study of Aging, also took advantage of matches to administrative records to ascertain the fact and cause of death and the use of inpatient medical care. This paper describes the study and discusses some of the advantages and disadvantages of using a study designed for other purpose as the base for a longitudinal study.

KEY WORDS: Aging; Longitudinal; NHIS; LSOA.

1. INTRODUCTION

Many countries, including Canada, conduct national cross-sectional population-based sample surveys. Many countries also conduct longitudinal studies based on a sample of the population. In addition, all countries have administrative records with data that, at least theoretically, can be linked with these surveys and studies. However, the cross-sectional and longitudinal studies are seldom well coordinated, and administrative records are not used as frequently as they could be.

This paper describes using a national cross-sectional survey as the basis for a longitudinal study in which data from administrative records were linked with the study data.

The study is the Longitudinal Study of Aging, which was based on the Supplement on Aging to the 1984 National Health Interview Survey. Records of the study participants were linked with a record file of death certificates and with a file of medical care claims records. As far as I know, it was the first longitudinal study that was designed to be based on an ongoing national cross-sectional survey and designed to take advantage of matches to administrative records to add information to the survey responses.

The paper first describes the study then discusses some of the advantages and disadvantages of this approach.

2. THE STUDY

2.1 The Baseline Survey

The Longitudinal Study of Aging (LSOA) was based on the National Health Interview Survey (NHIS), an ongoing cross-sectional survey of the civilian noninstitutionalized population of the United States. The NHIS consists of a core, or basic, questionnaire that is revised about once a decade and supplements on special topics that change from year to year. In 1984 there were two supplements to the NHIS, a supplement on health insurance and the Supplement on Aging (SOA). The core NHIS and the two supplements provided the baseline

[1] M.G. Kovar, National Center for Health Statistics, Hyattsville, Maryland, U.S.A. 20782.

data for the LSOA. The SOA also provided the sampling frame and included the questions that were asked on the subsequent LSOA interviews.

2.1.1 The basic NHIS questionnaire

The NHIS is the large continuing survey conducted by the National Center for Health Statistics that is designed to obtain information on chronic and acute illnesses, injuries and the use of health services for the civilian noninstitutionalized population of all ages living in households in the United States.

Details about the survey, including the sample design and procedures used in 1984, have been published (Kovar and Poe 1985). There are, however, some characteristics of the NHIS that were not what we might have chosen if we had had control over the entire process.

The NHIS is a national survey, based on a complex sample design, with interviewing in households every week (the weekly samples are independent national samples).

The survey relies on family respondents. Although all adults who are at home are invited to participate in the interview, any adult can respond for all other family members.

The survey is designed to obtain prevalence estimates, not to study relationships. As a result, subsamples are used when a fraction of the sample is sufficient to estimate the prevalence. Subsamples are used, for example, to provide estimates of the prevalence of chronic conditions. There are six lists of conditions each of which is used in one-sixth of the households.

2.1.2 The Health Insurance Supplement

The health insurance supplement contained questions about the public and private health insurance coverage of all members of the family living in each household. It had been used before on the NHIS to monitor changes in health insurance coverage. The family respondent was also the respondent for this supplement.

2.1.3 The Supplement on Aging

The Supplement on Aging (SOA) was new. It was the first national survey in the United States designed specifically to obtain information from a representative sample of middle-aged and older people living in the community.

The SOA can be viewed as an independent national survey of older Americans and the data can be analyzed as if they are from an independent survey. However, the data were collected as a supplement to an ongoing survey and the questionnaire and procedures for the supplement had to be integrated with those for the core NHIS. The design and procedures for the SOA have also been published (Fitti and Kovar 1987).

The response rate for the SOA was also high; 96.7 percent of the people age 55 and older in the NHIS households participated in the SOA either in person or by proxy. Therefore, the effective response rate for the SOA was 93.2 percent.

2.2 The Longitudinal Interviews

The longitudinal interviews included only the SOA participants age 70 and older. They were designed to measure change in functional status and living arrangements, including moving into (and out of) nursing homes. They were also designed to be evenly spaced at two-year intervals. Both parts of the design were held to with remarkable consistency, but both parts also yielded to external pressures. The details, including all of the questionnaires, have been published (Kovar, Chyba and Fitti 1992), but I want to review some of them here.

There were two major changes between the baseline and the subsequent interviews. One was the change in the size of the sample and the other was the change in the interview mode.

One major change was the change in the size of the sample.

Of the 7,527 SOA participants age 70 or older, only 5,151 were in the sample scheduled for interview in 1986. There was little money available for such a new concept as a longitudinal study of older people by telephone at that time. Fortunately, the National Institute on Aging was willing to take a chance, but the only way to do the study was to subsample. We designed the subsample to maximize the analytic potential by including every participant in the SOA who was 80 or older, every participant who was black and all their family members age 70 or older, and half of the rest of the SOA participants age 70 or older.

When the 1988 interview was planned, there was more interest in longitudinal studies of older people and more money. Therefore, it was possible to include all of the 7,527 age 70 and older from the SOA, which added 2,276 people to the 1988 and 1990 samples. Because of the design of the 1986 sample, the additional people were all people age 70-79 who were not living with someone age 80 or older or a black person age 70 or older.

The change in the sample size created analytic problems. Also, response rates for the people added in 1988 were lower than for those who were in the 1986 sample.

The second major change was in the interviewing mode.

The baseline interviews were in the household. The follow-up interviews were telephone interviews with mail questionnaires for those who did not have, or who were not located by, telephone.

The telephone interviews were conducted using computer-assisted telephone interviewing (CATI) at a U.S. Bureau of the Census telephone facility. The decision to use CATI instead of household interviews was based on costs, but it was made only after a feasibility study had shown that telephone interviewing was practical for older people (Fitti and Kovar 1985). The telephone interviewers were trained, and they had a day of training on this survey, plus mock and practice interviews, before they began interviewing the LSOA participants. They did not, however, have the years of training and experience that many of the interviewers on the baseline study had. However, in contrast with household interviewers, they had constant supervision because the supervisors could monitor any interview at any time and correct errors immediately.

There were three noticeable consequences of using the telephone: the interviews had to be shorter than when interviewers were in homes; response rates were lower than in household interviews; and there was less control over who answered the questions.

The content of the interviews was restricted to the two measures of primary interest (changes in functional status and changes in living arrangements) to keep the interviews as short as possible. The interviews focused on the primary measures needed for the purposes of the survey and obtained little information on covariates except for changes in critical covariates.

Because the major emphasis was on change in functional status, the ADL and IADL batteries were asked at each interview without modification. There was also an additional question at the end of the questions for each ADL and IADL: "Was this a change from the last time we talked with you?" The Nagi questions were also asked exactly as they had been at baseline. In addition, there were questions about reasons for change for the two most prevalent problems - difficulty walking one-fourth mile and difficulty climbing ten steps without resting.

The questions on living arrangements and marital status did not need to be repeated each time to detect change because such changes are highly salient, abrupt changes, unlike the subtle changes in functional ability. Instead there were simple questions on change and dates of change.

Thus, questions on the two critical interests of the longitudinal study were treated differently. There was no attempt to date change in functional status because such changes are often gradual, but there was a great deal of effort devoted to detect the fact of change. In contrast, a great deal of effort was devoted to detect the date of change in marital status or living arrangements because such changes are discontinuities, but little effort was needed to detect the fact of change.

The response rates for the telephone plus mail survey in 1986 remained over 90 percent when those known to be deceased were included. However, the response rates were lower in 1988 and 1990, in part because of the 4-year interval for the people added in 1988. Self-response rates were lower for all three follow-up interviews because women were more likely to answer the telephone and because people were less likely to be able to answer for themselves as they grew older.

The design called for interviews every two years. Two years after the midpoint of the baseline study was July 1 1986. However, some older people move with the seasons so interviewing was scheduled for August through October to increase the chances of reaching people. All mail questionnaires, those to people who did not provide a telephone number and those who could not be reached by telephone, were sent after the telephone phase had been completed. Therefore, responses to the 1986 interview were concentrated in August-October but continued through the end of 1986.

The 1988 schedule was the same as the 1986. In 1990, however, the Census facility was needed for Census work in September and October. Interviewing for the LSOA had to end in early September, so interviewing began in July. Therefore, responses to the 1990 interview were concentrated in July-August, with responses to mail questionnaires through October. In consequence, interviews were rarely precisely two years apart.

2.3 Matched Records

2.3.1 The National Death Index

The National Death Index (NDI) is a computerized file derived from death certificates that is maintained by the National Center for Health Statistics (NCHS). Anyone wanting to use it for research (it cannot be used for regulatory purposes) must submit an application to the Division of Vital Statistics, NCHS. After the application has been reviewed and approved, successful applicants submit a file of records to be matched. Matching is most successful for files with all or most of the 10 items suggested in the NDI Users Manual (NCHS 1981).

Because the NDI match was anticipated when the baseline survey was planned, participants were asked for permission to match and for information on all of the 10 items. Almost all of the participants (15,938 out of 16,148 in the SOA and 7,426 out of 7,527 in the LSOA) provided the information.

The NDI produces a list of all-possible-matches in descending order of probability. The study director is responsible for deciding whether the match is correct. The algorithm used in the LSOA to make this decision is described in Kovar, Chyba and Fitti (1992). Four categories denoting the certainty of the match for deaths from 1984-89 are on the public-use tape.

2.3.2 Cause of Death

NCHS also maintains a computerized file of records with multiple cause-of-death data that includes the death certificate number. NCHS project officers wanting to use it must submit an application to the Division of Vital Statistics, NCHS. After the application has been reviewed and approved, successful applicants submit a file of records to be matched[2].

Because the SOA records had been linked with the NDI and only those records where death was certain or nearly certain would be submitted, permission for this linkage was granted.

2.3.3 Medicare Records

The Health Care Financing Administration (HCFA) maintains a file of all bills for services covered by Medicare. LSOA records are linked with those Medicare records only if the participant gave permission and provided a

[2] Access to this file is restricted to surveys conducted by NCHS, which has this privilege because it has a contract with each registration area. These contracts purchase the data needed to publish the national vital statistics and to maintain the NDI.

number that could be used for linking - a Social Security, Railroad Retirement, or Health Insurance Claims number. Only the number is sent to the Health Care Financing Administration to maintain confidentiality.

The Medicare match was more complicated than the NDI match because it was necessary first to match against the enrollment file to make certain that the person was enrolled, then check the enrollment record against the LSOA record to make certain that the numerical match was indeed for the sample person. Only after that was the file of LSOA records submitted to HCFA to obtain the records of service use.

Information, (including the date, diagnostic codes, procedure codes, and the charge) is abstracted from the records of inpatient care (Medicare Part A). There is one record for each hospitalization on this file, which is on the LSOA public-use tape. The main person file has an indicator that there is no, or at least one, record for that person in this file.

No detailed information is abstracted from the records for other services covered by Medicare. There is, however, a file on the LSOA public-use tape with an indicator of whether there was a bill for the specific service in each calendar year.

There were 11,497 persons in the SOA sample who were age 65 or older; records of 10,442 of them including 6,920 in the LSOA sample of persons age 70 and older have been matched.

3. ADVANTAGES, PROBLEMS, AND SOLUTIONS

3.1 Advantages

Using an on-going national survey to obtain baseline data for a longitudinal survey has the obvious advantage of reducing the cost. Since the NHIS would have been conducted regardless, there was no cost of a screening survey to locate people age 70 or older. In addition, the basic NHIS provided demographic, social data, and health data that are needed for any such study.

The high response rates attained by the NHIS provided another advantage. Because it is a continuing survey, interviewers for the NHIS seldom work on any other survey, they are frequently retrained, and they know the basic questionnaire very well. One result is high response rates; in 1984 information was obtained from 96.4 percent of the households selected for the sample. That was a real advantage of using it. We could not have hired and trained interviewers for a one-time survey and attained such high response rates.

The linkage with the NDI files had two advantages, especially for an older population where death rates are high. It reduced the proportion of people who would have been considered as "lost to follow-up" if we had not had that linkage. It also provided a precise date of death, which was useful for survival studies.

The link to the Medicare files provided information that is difficult for household informants to report correctly. It is hard for elderly respondents to date events accurately and few people of any age can report diagnoses or procedures accurately.

3.2 Compromises

Adding the SOA to the NHIS created several procedural problems. We wanted a self-respondent, we wanted conditions for everyone, we wanted more detail on some things that were on the baseline survey, and we wanted to ask questions that had been used on other surveys.

The solution to integration was to ask the SOA questions after the basic NHIS and health insurance supplement had been completed and to use the same recall periods as the NHIS, but to change the respondent rules and add a special list of chronic conditions.

Asking the SOA questions after many other questions about health may have been an advantage if it improved recall as respondents had the chance to think about their health. Using the recall periods of the NHIS (two weeks, one year) meant that the recall period for some questions adapted from other studies had to be changed (Fitti and Kovar 1987).

Changing the NHIS rule that any adult could respond for all other family members to a self-response rule succeeded so well that 90 percent of the participants in the SOA who were age 70 and older (in contrast with 80 percent on the NHIS) responded for themselves. Those who did not were usually very old or incapacitated. Adding a special list of chronic conditions developed from conditions shown to have a high prevalence among older people on earlier NHIS surveys meant that every respondent was required to respond yes or no to each condition on the list.

3.3 Disadvantages

Because we wanted more detail on some topics that were on the basic NHIS, we had to repeat some questions in order to have the proper context for the other questions on that topic. For example, we wanted information on the relationship of other people in the family to the older person, which meant returning to an earlier topic. Another example is that we wanted a complete sample for a special list of chronic conditions known to have high prevalence among older people. Questions on about one-sixth of these conditions had already been asked in each household.

We were not able to use "batteries" of questions or the recall periods that had been used on the original questionnaire for the questions that were adapted from other studies. Therefore, some reviewers have said that the questions were not calibrated.

The baseline survey was conducted by personal interviews in the home. The follow-up interviews were by telephone. The change in interviewing mode may have influenced the estimates of change in functional status between the baseline survey and the first follow-up. There is little research on mode effects in interviewing older people. There is, however, research that shows that proxy- and self-respondents do not give the same answers, and there was a higher proportion of proxy-respondents to the telephone interviews.

The fact that the NHIS is conducted throughout the year meant that the first longitudinal interview could not be two years after the initial interview. We compromised, but it made survival analysis more difficult.

4. SUMMARY

The LSOA has been a very successful study.

It was possible to conduct it at a relatively low cost because the initial household interview avoided expensive screening for the relatively-rare population of people age 70 and over and because that interview obtained a great deal of information that was needed for the longitudinal study. Using the telephone for the follow-up study, which would have been more difficult had we not had that initial personal contact, also kept costs down.

The baseline survey benefitted by having a field staff of U.S. Bureau of the Census interviewers who were extremely well-trained on how to obtain high response rates and the concepts of the survey.

However, the main reason we deem it successful is that it was designed for the research community and that community has used it. It has been widely used for papers in peer-reviewed journal, numerous theses and dissertations, working papers, and has provided national data for health policy.

ACKNOWLEDGEMENTS

The Longitudinal Study of Aging could not have been conducted without my colleagues at the National Center for Health Statistics, Joseph Fitti and Michele Chyba, or without our project officer at the National Institute on Aging, Richard Suzman. The former took responsibility for major aspects of the study, such as interviewing and record matching. The latter provided support and guidance over the entire period of the study.

REFERENCES

Fitti, J.E., and Kovar, M.G. (1987). The supplement on aging to the 1984 national health interview survey. *Vital and Health Statistics*, 1, 21.

Fitti, J.E., and Kovar, M.G. (1987). A multi-mode longitudinal study of aging. *Proceedings of the American Statistical Association Section on Survey Research Methods*.

Kovar, M.G., and Poe, G. (1985). The national health interview survey design 1973-84, and procedures 1975-83. *Vital and Health Statistics*, 1, 18.

Kovar, M.G., Chyba, M.M., and Fitti, J.E. (1992). The longitudinal study of aging: 1984-1990. *Vital and Health Statistics*, 1, 28.

National Center for Health Statistics (1981). *Users Manual, The National Death Index*, Hyattsville, Maryland: Public Health Service.

SESSION 7

General Applications II

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

A LONGITUDINAL SURVEY AND REALITY CHECK FOR THE VALUE OF FINANCIAL ASSETS

C.D. Cowan[1]

ABSTRACT

The Estimated Cash Recovery (ECR) Survey has all the ingredients of a classic longitudinal multipurpose survey. It measures the expected return from the sale of failed financial institutions that have been placed into receivership by the U.S. government.

The survey measures recovery values for 17 types of assets, with strata defined by region and the size of institution. The population of assets under study changes rapidly in successive quarters, with new institutions being placed into receivership over time, and financial assets being sold and removed from the population. The value of the assets in sample change over time also, as economic conditions vary. All of these factors: the multiple uses of the survey, changing economic conditions, and changes in the structure of the population being studied lead to a complicated and interesting sample design and analysis.

KEY WORDS: Asset liquidation and recovery; Receivership.

1. INTRODUCTION

In June of 1991, the Resolution Trust Corporation (RTC) initiated a survey of the institutions it has taken into receivership since the RTC was founded in 1989. The purpose of the survey, called the Estimated Cash Recovery (ECR) Survey, was to provide quarterly estimates to Congress, the Administration, and the public in general of the amount of recovery expected from the sale of assets from failed Savings and Loan Institutions (S&Ls). An asset for an S&L is a loan made by the S&L for one of a variety of purposes (like a commercial loan or a construction loan), or a type of property that the S&L held or had received as collateral for a loan. There were 19 types of assets covered by the survey; the list of assets covered is found in Table 1.

The amount of recovery expected is the total recovery in dollars, summed over the 19 asset categories. Of almost equal interest is the recovery rate, defined as the total amount expected to be recovered divided by the current book value of the assets. The current book value is defined to be the original amount of loan, minus payment received on the loan. On performing loans, loans which are still being paid off by the person or business who took out the loan, the book value continually declines as the principal is paid off.

There are several complicating factors that made this an interesting and complex sample design. First of all, since there is some interest in the recovery rates for each of the asset categories, one has to trade off the design for an estimate of the total with the design to estimate each asset category total separately. Optimizing the sample to answer each of these goals would result in two very different sample designs.

Second, the population for this survey is changing very rapidly. Each quarters a number of institutions are resolved by the Office of Thrift Supervision and are placed into RTC receivership. At time of resolution, the depositors at each institution are paid off and the assets of the institution are taken over by the RTC. So there is a healthy inflow of institutions and assets into the population. At the same time, assets from institutions previously taken over by the RTC are being sold. Assets in different categories are sold at very different rates;

[1] C.D. Cowan, Senior Statistician, Resolution Trust Corporation, Washington, DC, USA 20434-0001.

e.g., loans on single family dwelling units are sold much more rapidly than construction loans. So there is also a healthy outflow of assets, and the flow out is at a very different rate than the inflow of assets.

Table 1: Asset Categories Used in Estimation of the Expected Cash Recovery.

Assets Covered by the Survey

1) 1-4 Family Mortgages - Performing
2) 1-4 Family Mortgages - Nonperforming
3) Multifamily (5+) Mortgages - Performing
4) Multifamily (5+) Mortgages - Nonperforming
5) Raw Land - Performing
6) Raw Land - Nonperforming
7) Construction Loan - Performing
8) Construction Loan - Nonperforming
9) Commercial Mortgage - Performing
10) Commercial Mortgage - Nonperforming
11) Commercial Loan - Performing
12) Commercial Loan - Nonperforming
13) Consumer Loan - Performing
14) Consumer Loan - Nonperforming
15) Real Estate Owned
16) Furniture, Fixtures and Equipment
17) Subsidiary Equity
18) Subsidiary Loans
19) Other Assets

Assets Not Covered by the Survey

20) Junk Bonds
21) Mortgage Backed Securities
22) Other Backed Securities
23) Judgements
24) Charge-Offs

Finally, there is some information available from the accounting ledgers of the RTC that can be used in estimation to help reduce the variance of the estimates, using either a ratio or regression estimator. The recovery rate mentioned earlier is a good example of the type of ratio estimator that would be of interest. However, because of the nature of the processes that relate the recovery to the original book value of the asset, the recovery rate should be treated as bounded below by zero and above by unity. We can use the relationship between expected recovery value and book value to reduce the variance of the estimates (since book value is known for all population members), and achieve further reductions by bounding the estimates so that the range of these estimates is between zero and one.

This paper presents methods used to design and implement the ECR with some discussion of the problems encountered with the conduct of the survey. The paper concludes with some preliminary results from the survey collected over the first four quarters.

2. METHODS

Before describing the design and analysis of the survey, it would be useful to indicate how the data is to be collected. How the data is to be collected has an impact on the design of the study because of cost considerations.

A sample of institutions and assets within institutions is selected, using the sampling methodology described below. Assets within institutions are valued when first sampled, and then rotate in and out of sample according to a prespecified schedule (also described in the next section). Chart 1 describes this process.

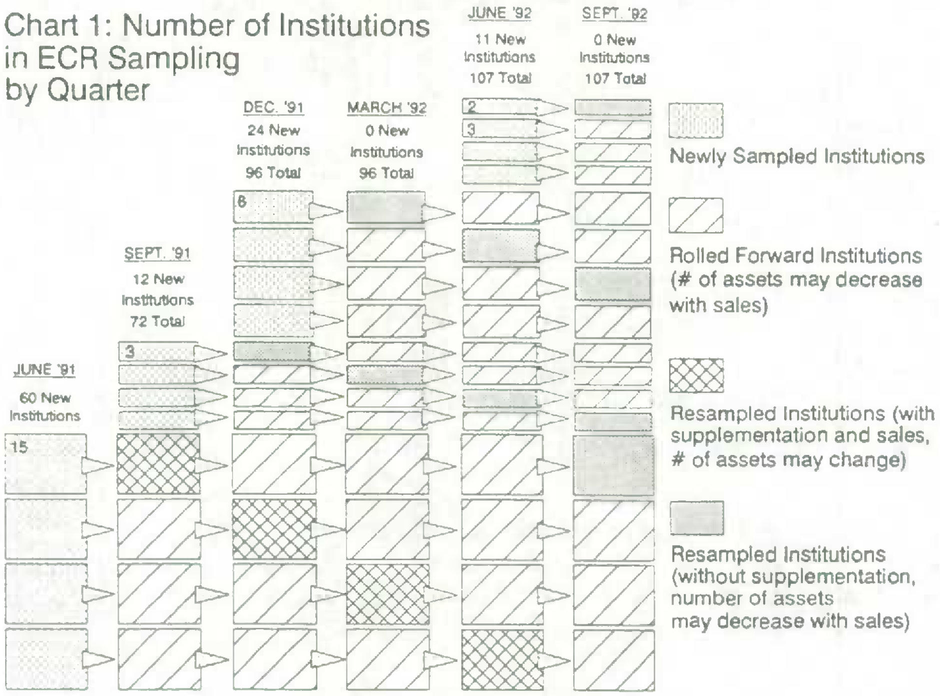


Chart 1: Number of Institutions in ECR Sampling by Quarter

JUNE '91
60 New Institutions
15
SEPT. '91
12 New Institutions
72 Total
3
DEC. '91
24 New Institutions
96 Total
6
MARCH '92
0 New Institutions
96 Total
JUNE '92
11 New Institutions
107 Total
2
3
SEPT. '92
0 New Institutions
107 Total
Newly Sampled Institutions
Rolled Forward Institutions (# of assets may decrease with sales)
Resampled Institutions (with supplementation and sales, # of assets may change)
Resampled Institutions (without supplementation, number of assets may decrease with sales)

Once the sample selected, the list of assets to be valued is given to a contracting firm of accountants who determine when the asset will be sold and the amount the RTC can expect to receive. The accountants also determine the flow of operating income for the asset or property, and the direct expenses incurred in the management of the asset.

The information on the expected flow of operating income, payments on the loan, and direct expenses are recorded on a quarterly basis for the two years following the date of data collection, and on an annual basis for years three, four and five after the date of collection. The quarterly information is used to permit "rolling over" the estimates in subsequent quarters when the sampled institutions have rotated out of sample.

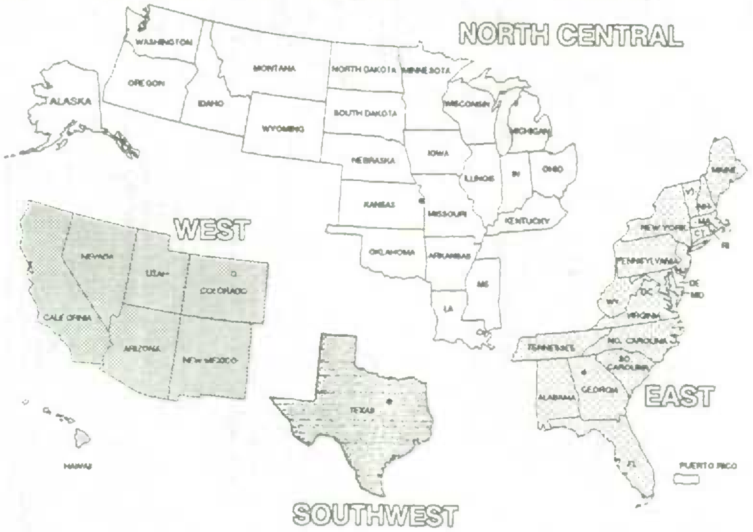
The other major part of the data collection is determining if a sampled asset is sold in subsequent quarters. If an asset is sold, it is no longer part of the survey process because the RTC has already realized the recovery. The survey's purpose is to estimate the future recovery to be realized - for sold assets we know exactly what has already been recovered so it is not necessary to sample these assets. The specialists who value assets in the field are also responsible for reporting the sale value of sold assets.

3. SAMPLE DESIGN

The plan was to develop a sample of approximately 5,000 assets nationally to be valued by the accountants. To make this as efficient as possible, both from a variance standpoint and a cost standpoint, the sample was designed to be a stratified multistage cluster sample. In addition, each asset category was considered to be a separate stratum in the second stage of selection.

Stratification was conducted at the first stage by constructing a two way table of institutions, a table that would change at each quarter as the number of resolved institutions grew. The stratification variables were region, with categories East, North Central, Southwest and West - defined as in Chart 2 - and initial size of institution. Categories for size of institution were "Less than $100 million", "between $100 and $500 million", and "Greater than $500 million".

Chart 2: Regions for the Resolution Trust Corporation



Strata were not balanced in terms of population size of institutions because there are many more small (in terms of initial size) institutions than large. The strata are also not balanced in terms of totals size (defined by book value) and number of assets, but the imbalance goes in the other direction, with the four strata with initial size greater than $500 million holding up more than half the total assets.

There were several conflicting factors to be accounted for in the stratification and estimation:

1) the distribution of the assets was heavily skewed to the larger institutions;

2) the number of institutions was heavily skewed to the smaller institutions;

3) the primary goal of the survey was to produce a single national estimate of recovery;

4) not all institutions have all types of assets, so a sample of institutions that was sparse in some strata may completely miss some asset types;

5) recovery rates by region, size of institution, and asset category were of equal importance and a close second to the national estimate in terms of how the data from the survey were going to be used.

Ordinarily, the most efficient procedure would be to set up an objective function (like a variance function) to minimize subject to a fixed cost. However, because so little was known about the relative costs of valuing the assets and the variation in the estimates of recovery by stratum, the most efficient procedure devolved to selecting an equal number of institutions per stratum, and an approximately equal number of assets within each sampled institution. In hindsight, because of problems in getting complete lists of assets from each sampled institution, this procedure was the correct one to choose. Any other procedure that would have involved more complicated sampling procedures at the asset sampling stage would have greatly slowed down the survey process.

For June of 1991, 60 institutions were sampled, about five per stratum (because of the initial distribution of institutions, one stratum had only four institutions sampled, and another had six sampled to compensate). These 60 were then assigned to four rotation groups, each group of size 15, to be recontacted in subsequent quarters. The 60 sample institutions were assigned so that all 12 strata were represented at least once in each quarter, and no stratum was represented more than two times. Within each institution, a list of all assets in the 19 asset categories was obtained for sampling at the second stage, stratified by asset category. A sample of a minimum of five assets ultimately was selected in each asset category if there were five assets to be selected. If fewer than five assets were available in a category, all the assets in that category were taken into sample. Assets were ordered by book value (size) within a category and systematically sampled.

In September of 1991, the design became more complicated. There were now three sources or lists of institutions that were available. The first was the set of institutions that were originally contacted in June of 1991 that would now be recontacted in September. The same assets valued three months previously would now be revalued (because the status of the loan may have changed, the economic conditions affecting the sale price may have changed, or other factors may have had an impact in valuing the loan).

The second source was the set of institutions that were originally contacted in June of 1991 that would not be recontacted in September. Assets in this group of institutions would be "rolled forward", as describe earlier. Institutions in sources one and two represented all institutions resolved by June 1991.

For December 1991, we faced very much the same situation, except the first source of data was now the 15 institutions in the second rotation group from June, 1991 plus the 3 institutions assigned to rotation group 2 from September, 1991. From the second source of data, we roll forward the estimates from the remaining three rotation groups from June and September. The third source was again new sample from resolutions that occurred between August 1991 and November 1991, with assets sampled in the same way as in June of 1991.

March and June of 1992 proceeded in exactly the same fashion, but there were no new institutions added in March of 1992 because there were too few new resolutions. In June of 1992 we added 11 new institutions, again one per stratum. There were no institutions sampled in stratum four because there were no resolutions of institutions in that stratum.

Finally, for each administration of the survey, each asset was checked to determine whether it had been sold. This was done for rotation groups both in and out of sample so that information on sold assets could be obtained more rapidly and also to counter any biases that might occur for rotation groups if asset groups sold at differential rates.

4. ESTIMATION

The final piece of the project is developing an estimation scheme. Standard estimation methods were used for most of the survey estimates, but some adaptations were made for determining confidence intervals.

For most of the recovery rate estimates, we have quite a bit of information we can use to form ratio or regression estimators. Specifically, we know the book value of every asset, and for the longitudinal portions of the survey we know change in book value. At a minimum, we can use book value for all sampled and population assets, and the projected recovery for all sampled assets, and form the classical stratified ratio estimator for a two-stage clustered sample (Cochran 1963). The ratio estimator and variance estimator for the ratio are both well-defined and have been known for a long time.

Because some of the recovery rates estimated are unusually low or high (close to zero or close to one respectively), it turns out that the tails of the confidence intervals produced as part of the survey estimation procedure go below zero (implying a negative recovery) or above unity (implying a recovery greater than the original value of the asset). While this may happen for an individual asset under very unusual circumstances, it cannot be true for the population value of the recovery rate for procedural reasons related to the methods used for selling the assets. This limitation means that the normal approximation so commonly used for construction of confidence intervals is appropriate for this survey.

As an alternative, we tried a Bayesian approach (Box and Tiao 1973). We assumed that the recovery rate was a parameter drawn from a prior distribution, the Beta distribution. We used method of moments to estimate the parameters of the Beta distribution, using the mean and variance of the sample estimates of the recovery rates. Finally, from the Beta distribution directly we determined the lower and upper bounds of the confidence interval so that we had the tightest confidence bound possible, with the restriction that the lower and upper bounds were in the interval on zero to one. This procedure in all cases reduced the confidence intervals over what would have been calculated using the normal distribution, but kept the mean and variance of the estimate the same.

5. FURTHER RESEARCH

Like any good selfperpetuating statistical program, more research is needed, specifically in the area of producing composite estimates and making better use of the longitudinal nature of the data. Furthermore, some research is needed on whether the Bayesian approach using the Beta as a prior can be improved by looking at the declining trend associated with most of the recovery rates (as a function of the overall declining economy).

REFERENCES

Cochran, W.G. (1963). *Sampling Techniques, Second Edition*, John Wiley and Sons, Inc. New York, NY.

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company, Reading, MA.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

PANELS OF ENTERPRISES AND CONFIDENTIALITY: THE SMALL AGGREGATESMETHOD

D. Defays and P. Nanopoulos [1]

ABSTRACT

The need for networks for the rapid collection of data that can easily be adapted to new requests for information is increasingly being felt in Europe.

Since collectors of information (usually national institutes of statistics) are not always able to communicate individual data, it is important to propose methods of pooling this information and of defining virtual units that preserve the confidentiality of data while maximizing the micro-economic information provided.

Various techniques are examined and the results obtained are compared.

KEY WORDS: Confidentiality; Micro-aggregation; Principal components; Fuzzy clustering.

1. INTRODUCTION

Basing economic analyses on macro-economic data alone is problematic to say the least. Trends can be difficult to interpret: variation at the micro-economic level (which can frequently be a source of valuable information) are concealed by the aggregation of data and the methods of aggregation applied to the data can determine the nature of the analyses that can be carried out and the nature and number of the models that can be tested.

It is generally impossible, once a set of tables has been definitively fixed, to respond to questions requiring aggregation of the data on different lines from those laid down; when data on enterprises have been aggregated on the basis of size-of-enterprise classes, for example, with size defined in terms of numbers of employees, it is impossible to focus on a sub-population of units defined in terms of turnover. Exploratory analysis is almost impossible. How, then, can we take a statistical interest in small innovating enterprises or units which sub-contract if these characteristics were not included when the statistical tables were first defined?

Eurostat, the statistical office of the European Community, is currently lacking individual data in the business sector. In order to address that problem, a project which aims at the creation of a network of co-ordinated national panels of enterprises has been launched.

2. AN EUROPEAN PANEL OF ENTERPRISES

Specifications for a common approach to business panels in Europe are currently being prepared. The objective is to help the existing national panels in Europe to converge towards common standards. But before starting any measure of implementation, different problems have still to be solved. One of them will be to study how existing panels could be tuned in order to meet Community requirements. Indeed, some countries already have panels of enterprises; the problem will be to see how compatible they are with what is recommended and to examine how they could improve their comparability with other national panels. Another problem is to explore under which conditions some type of microdata could be transmitted to Eurostat. The national statistical institutes (NSI) of the Member States of the European Community are not always in a position to supply

[1] D. Defays and P. Nanopoulos, Eurostat.

individual data because they have to protect the confidentiality of the data transmitted to them by the persons and legal entities that supply them with information. NSIs only transmit tables of aggregated data and the analyses carried out at a European level can generally be based only on this source of data.

The aim of our article is to focus on the question of how we can maximize the quantity of data transmitted to Eurostat by the NSIs while at the same time preserving confidentiality as much as possible.

3. THE CONFIDENTIALITY ISSUE

This is not a new problem, of course. But there are certain aspects of the way which the question is posed for Eurostat which merit attention and possibly justify the adoption of a new methodological approach. Firstly, the range of uses of the data transmitted to us is particularly extensive and certainly wider than that of the traditional type of research institute which is generally interested in a specific type of analysis. Secondly, statistics on enterprises must not only meet the needs of economic analysis but also help the Commission to manage and monitor certain Community projects and evaluate the impact of its programmes. This obviously entails the usage of information in many different ways.

Moreover, the preservation of confidentiality necessitates the pre-processing of large volumes of information in 12 Member States. These operations must be as cost-effective as possible. Finally, because the preservation of confidentiality is a particularly sensitive issue and because the transmission of data to Eurostat which is an internal organ of the Commission of the European Communities is a controversial matter, the data protection procedure must be simple, easy to understand and air-tight. It must be explainable to non-specialists.

This article presents a few theoretical results relating to aggregations of units in classes with fixed sizes and describes a simple method of minimal aggregation of individual data which does not hamper the analyses to be carried out, which maximizes the information transmitted and which ensures the preservation of statistical confidentiality. Its impact on the analysis of longitudinal data is examined. Several simulations with data relating to enterprises have been carried out and analysed.

4. STATE OF THE ART

As already mentioned, the protection of individual data is obviously not a new problem. Various methods have been examined and used. A comparative analysis can be found, for example, in G. Paass (1988). These existing methods consist in perturbing the data by adding "noise", constructing units by switching blocks of information (*i.e.* values for subsets of variables) between the original units, or, constructing micro-aggregates on lines that differ, however, from those proposed in the present article.

The traditional methods of protection of individual data seem to us to be inappropriate in our situation for various reasons. The methods of perturbation do not provide an adequate guarantee of confidentiality, particularly with distributions as asymmetrical as those manipulated in the field of business statistics. They also carry the risk of bias in the estimation of a certain number of parameters (Adam & Wortmann 1989). Their application to the analysis of longitudinal data could also cause problems.

Data swapping enables the preservation of certain characteristics of the original multivariate distribution but the application of this method can be far from simple. It appears to be an efficient means of preserving the confidentiality of statistical data; but its use for exploratory analyses involving investigation of the multidimensional structure of the data may give rise to problems, because this structure has been broken up, by definition, by the construction of synthetic units. There is also a risk that the application of this type of method to longitudinal data will introduce a considerable measure of bias.

Generally speaking, the rules permit the transmission of aggregated data when the number of individuals aggregated exceeds a threshold value k (normally k = 2) and when none of the individuals represents the quasi-totality of the aggregate. A strict application of this rule enables us to obtain, in place of individual data, small aggregates or the averages of these micro-aggregates. They can play the role of fictitious individuals which we propose to call "pivots".

For highly homogeneous populations of units, the advantages of this formula are obvious. On the other hand, micro-aggregation would appear to have a significant effect, in certain cases, on the statistical properties of the data. It also inevitably involves the use of methods of classification that can prove expensive when applied to large volumes of data.

Eurostat has also studied a technique for the definition of prototypes (Bragard *et al.* 1988). The method consists, essentially, in selecting, within a population of units, a certain number of virtual and real units considered to be "representative" of the population. The techniques used in these studies is inspired by a method of fuzzy clustering proposed by M. Roubens which consists, essentially in reducing the dimensions of the data by principal component analysis followed by selection of representative units based, inter-alia, on fuzzy clustering. In each class, the element with the maximum "belonging" function is selected for use as a prototype. The degrees of "belonging" are then used to estimate the population parameters. A simulation has already produced promising (but so far unconfirmed) results, but the method is also rather complicated and some of the criteria for the choice of prototypes are still quite arbitrary.

5. EXAMPLES OF NEEDS

A question one is often asked is to define the size of a firm. From the administrative point of view the notion of "size" of a firm relates to the concept of the unit (legal unit, enterprise, local unit, ...) and the measure of several relevant variables like "Number of Employees", "Turnover", or "Fixed Assets" or even other variables related to the input or output of the unit.

There are several ways one can try to answer to this question. We give here 3 illustrative ones.

(a) Defining classes of the relevant variables

The problem with this simple approach is that use of different variables gives different results on the classification of units. The usual utilization in the EC legislation is to define a threshold for the variable X = "Employment " or the variable Y = "Turnover". The problem we face then is that of linking the X threshold with the Y threshold on a reasonable way.

The analysis of this problem needs a rather big sample of units in order to compare methods based on quintiles with methods based on scores.

(b) Principal component approach

If one agrees on the variables considered important, a reasonable way to address this issue is to examine the correlation structure of those variables over the various sub-populations in a manner that avoids low correlation due to disparities between activity classes.

The use of the first principal component gives a reasonably objective solution to the problem. As one may order all the units and define thresholds on the first principal component, these methods do not require individual data, but as cancellations have to be computed for a large number of sub-populations, individual data with categorical information are more convenient to use.

(c) Clustering approach

The clustering algorithms, like the k-means algorithm (Hartigan 1975), can be used to produce groups of firms that are supposed to correspond to various "size classes". Such an analysis can be carried out only with the availability of "individual" data.

6. FORMALIZATION

(a) Formalization of the general problem of small aggregates

Suppose that the total population is formed by N units; at each unit corresponds a vector X of p variables. The objective is to divide the set Ω into n, $(n=N/k)$, groups **of k points each,** $(G_1,..., G_n)$, in such a way that the n groups are as homogeneous as possible.

In order to define homogeneity of groups we need a notion of proximity or distance $d(\omega,\omega')$ of the points of Ω, which has to depend on the observed variables and a derived distance of groups, $D(G,G')$, of points of Ω. Formally the quantity that we have to minimize is a kind of "intra-group variance":

$$\Psi(G_1,...,G_N) = \sum_i \psi(G_i), \tag{6.1}$$

where

$$\psi(G_i) = \sum_{\omega\in G_i} P(\omega)\, D(\omega,G_i)^2 . \tag{6.2}$$

The nxk-grouping problem differs from the classical well known clustering problem (k-means algorithm (*cf* [3])) where one wants to divide the whole population in a fixed number of groups. In this case there is no cardinality condition.

(b) The *nxk*-grouping problem in R^p

The most usual and important case is where the space of values of $X=(X_1,..Xp)$ is the set of real numbers and the distance is the usual Euclidean distance in R^p.

We define the distances as follows:

$$d(\omega,\omega') = \| X(\omega) - X(\omega') \|$$

$$D(G,G') = \| m(G), m(G') \| \text{ where } m(G) \text{ is the average of } X \text{ over } G.$$

In this case the expression (6.2) becomes:

$$\psi(G) = \sum_{x\in G} p(x) D(x,G)^2 = \sum_{x\in G} p(x) \ \| x - m(G) \|^2, \tag{6.3}$$

which in terms of the conditional expectation of the vector X over the field generated by the partition $G = (G_1,...,G_n)$ can be written

$$\Psi(G_1,...,G_n) = E(\| X - E(X/G) \|^2),$$

where

$$E(X/G) = \sum m(G_i) I_{Gi}, \tag{6.4}$$

where I_{Gi} is the indicator function of the set G_i.

Suppose the vector X is centered, $(E(X) = 0)$, in $L_2(\Omega,A,P)$. The Pythagorean theorem yields the following decomposition:

$$\begin{aligned}\Psi(G_1,...,G_n) &= E(\| X - E(X/G) \|^2) \\ &= E(\| X \|^2) - E(\| E(X/G) \|^2)\end{aligned} \tag{6.5}$$

and the problem takes the equivalent form:

Maximize the expression

$$E(\| E(X/G) \|^2), \tag{6.6}$$

subject to the condition $P(G_i)=k/N$.

(c) The *2xk*-grouping problem in R^p

In the case where $n = 2$ the partition G is generated by a subset A, $G = \{A, A^c\}$, and a simple computation shows that the quantity to maximize is given by:

$$\Psi(A) = \| E(I_A X) \|^2, \tag{6.7}$$

under the condition $P(A) = 1/2$.

Suppose first that $p = 1$. Then the quantity in (6.7) becomes: $\Psi(A) = (\int I_A XdP)^2$.

From the classical lemma of Neumann-Pearson (see [4]) on the construction of most powerful test, it follows that the maximizing set is of the form:

$$A^* = \{ \omega \in \Omega \mid X(\omega) \geq \lambda \}, \tag{6.8}$$

for some constant λ.

This in practice means that A^* is composed by the $N/2$ elements of Ω with the highest values of X.

In the general case ($p>1$) we have a similar result:

Lemma A

There exists a vector $c \epsilon R^p$ and a constant λ such that the solution of (6.7) is of the form

$$A^* = \{ \omega \epsilon \Omega \mid \sum c_i X_i(\omega) \geq \lambda \}. \tag{6.9}$$

(d) The hyperplan in the 2-grouping problem

A little more can be said on the definition of the optimal set A^*.

The most difficult is to determine the vector c because after that the constant λ can be easily determined by the side condition $|A| = N/2$.

Lemma B

The vector $c \epsilon R^p$ which defines the optimal set $A = \{ \omega \epsilon \Omega \mid \sum c_i X_i(\omega) \geq \lambda \}$ in (6.9) satisfies

$$c = E(I_A X). \tag{6.10}$$

One immediate corollary of the proof is that, if we have a partition with a set A we can improve the value of Ψ by replacing A by the set $A^* = \{ Y(A) > \lambda \}$, where $Y(\omega) = < X(\omega), E(I_A X) >$. This operation can be repeated until $A = A^*$.

In fact it is shown in the proof of the previous lemma that

$$\Psi(A) = E[I_A Y(A)] \leq E[I_{A*} Y(A)] = E[I_A Y(A^*)] \leq E[I_{A**} Y(A^*)].$$

These results allow us to propose an algorithm for the general problem in R^p.

(e) The general problem in R^p

In the general case $(n>2)$, it is obvious that every pair of groups (G_i, G_j) has to be separated by an hyperplan as has been shown in the previous section. This hyperplan may be chosen to be perpendicular to the line joining the two barycenters of the two groups.

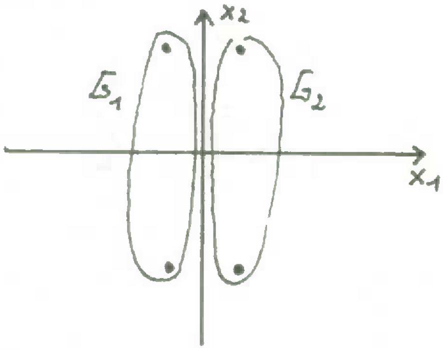
7. ALGORITHM PROPOSAL IN THE GENERAL CASE

The results established in the previous paragraph suggest the adoption of the following algorithm.

(a) Algorithm in the $n = 2$ case

1. Start with any grouping in 2 classes of the "same" size which satisfies the hyperplan separation condition.
2. Rank the projection of the points X on the line which joins the means of the two classes.
3. Define the new partition defined by the first $N/2$ points on that line and the last $N/2$ points.
4. If this partition is identical to the first one, stop: if not, go to 2.

It is easy to see that this algorithm will converge, but it can also be trapped in local minima, as shown in the example below, if the starting partition is as follows:



(b) Algorithm in the general case

In the general case ($n \geq 2$), the algorithm has to be transformed in the following way:

1. Start with any nxk-grouping which satisfies the hyperplan separation condition. This, for example, can be obtained by moving a hyperplan parallel to a fixed hyperplan and by taking groups of k points.

2. Take successively all pairs (G_i , G_j) of groups and optimise this partition in 2 groups as suggested in the preceding algorithm.

3. Restart step (2) until no further modifications are possible.

8. SOME SIMPLIFICATION

In our introduction, we mentioned the need to use a simple and cheap method. What has been proposed so far, is, maybe, still a bit complicated and could be expensive to perform. Some simplification is desirable.

We propose to increase the number of constraints in the definition of the group G. First, as already indicated, we specify that every class must contain the same number of elements. Second, to simplify the construction of G, we impose the condition that this partition can be derived from a simple ranking of units in terms of a unidimensional variable Y, by grouping contiguous units. More precisely, we stipulate the existence of a variable $Y:\Omega \rightarrow R$, and of boundaries a_i, such that

$$\omega \epsilon G_i \Leftrightarrow a_i < Y(\omega) < a_{i+1} .$$

Group G is thus simply defined in terms of $Q\ (i/n)$ of the distribution of a variable Y.

It must be noted that the condition we impose here boils down to assuming that the separator hyperplanes obtained in the general case are all parallel.

In fact, this hypothesis brings us back to a simple practical situation. It has already been suggested, in fact, in the case of business statistics, for example, that the units be classified on the basis of numbers employed and that G be defined by grouping enterprises k by k on the basis of the order obtained.

But why opt for numbers employed, rather than turnover? Is there not, among the linear combinations of variables X_i, a combination which minimizes the loss of information obtained by replacing all the elements in a class by their average value? Which combination of X_i values will minimize the loss of information brought about by grouping the elements on k by k basis defined by the order of this combination?

This problem is therefore a simplified version of our initial problem.

9. SIMULATIONS

Several simulations have shown us the significance of the choice of this unidimensional variable for ranking the units.

The data we analysed were taken from a sample of approximately 5,000 industrial enterprises, characterized by 11 economic variables: number of persons employed, turnover, export sales, physical investments, value added, remuneration of employees, gross operating margin, value of sub-contracts awarded to third parties, expenditure on publicity, number of establishments and number of economic activities of the enterprise.

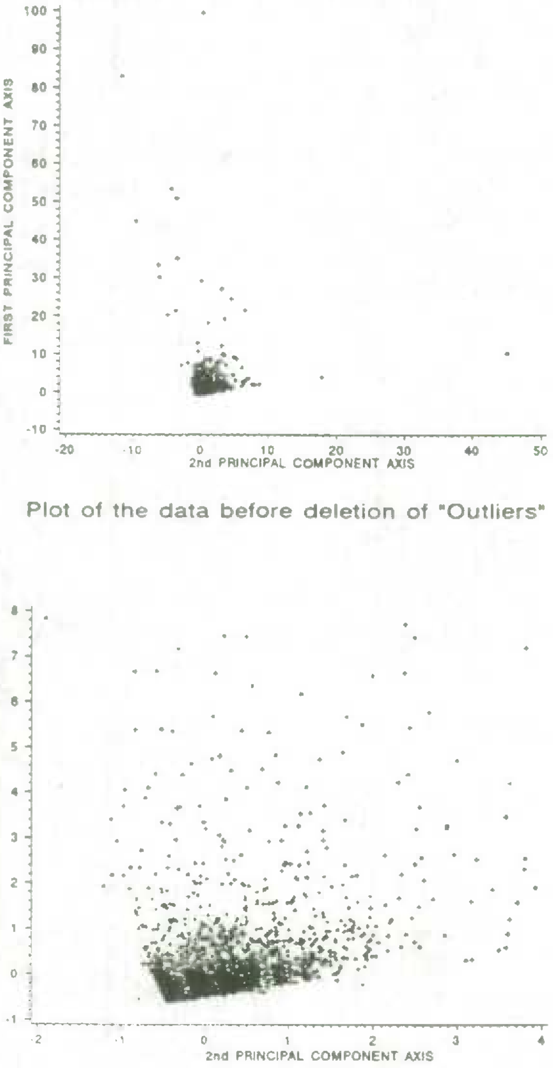
The following figures present the principle results of the simulations: structure of the data, percentage losses of information attributable to aggregation into classes of k individuals, for different values of k and different choices of variable Y, measured by the coefficient defined by

$$g(X/G) = E(\| X - E(X/G) \|^2) \ / \ E(\| X \|^2)$$

and some indication on how the grouping by 3 affect the internal structure of the data.

Figure 1 shows the data in the space of the first two principal components. About 40 "outliers" were withdrawn from the population in order to avoid effect linked to the excessive skewness of the distributions and all the variables have been standardized (mean = 0 and standard deviation = 1).

Figure 1: Plots of the First Principal Component against the Second Principal Component.

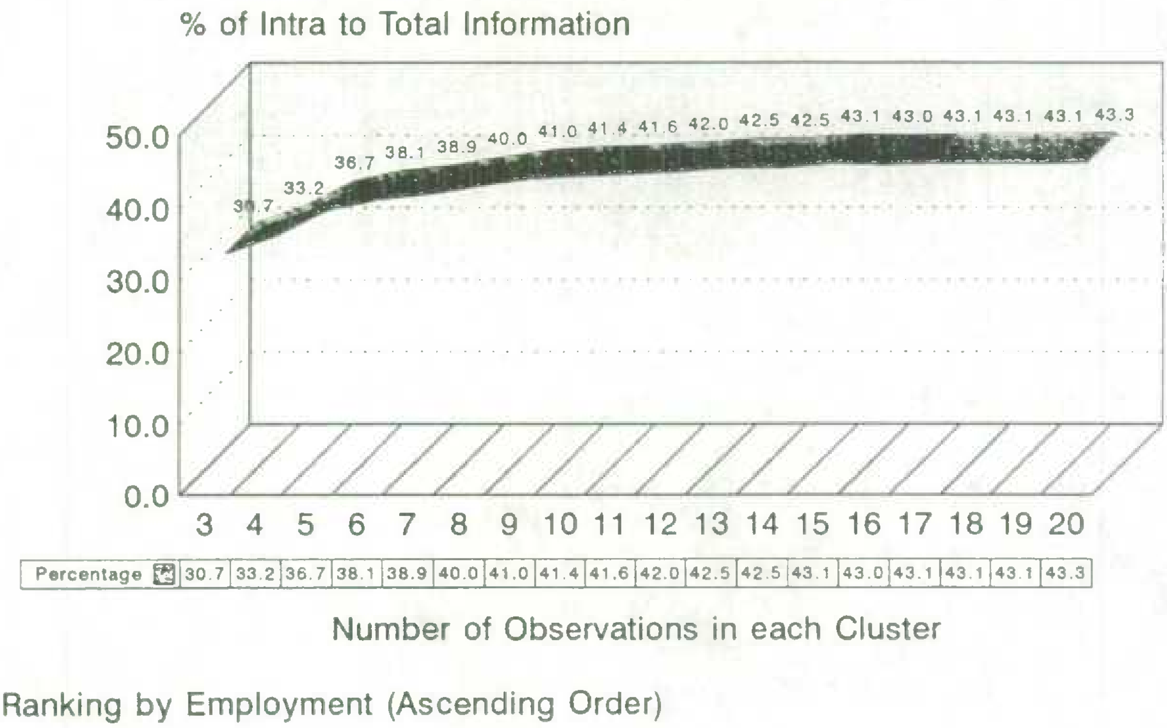


Plot of the data before deletion of "Outliers"

Plot of the data after deletion of "Outliers"

Figure 2 shows the increase in the losses of information measured by *g(X/G)* when the size of the groups in the *k*-grouping is increased. In that figure, the enterprises have been ranked according to their employment (which thus plays here the role of the variable *Y*) and grouped *k* by *k* along that dimension.

Figure 2: Percentage of Intra-group to Total Information (Outliers Deleted).



Figures 3 and 4 are similar to figure 2 but the dimension (variable *y*) which defines the grouping is respectively the value added and the first principal component. It is worth noting that as expected, the results obtained with the principal component are better than those obtained in figures 2 and 3.

Figure 3: Percentage of Intra-group to Total Information (Outliers Deleted)

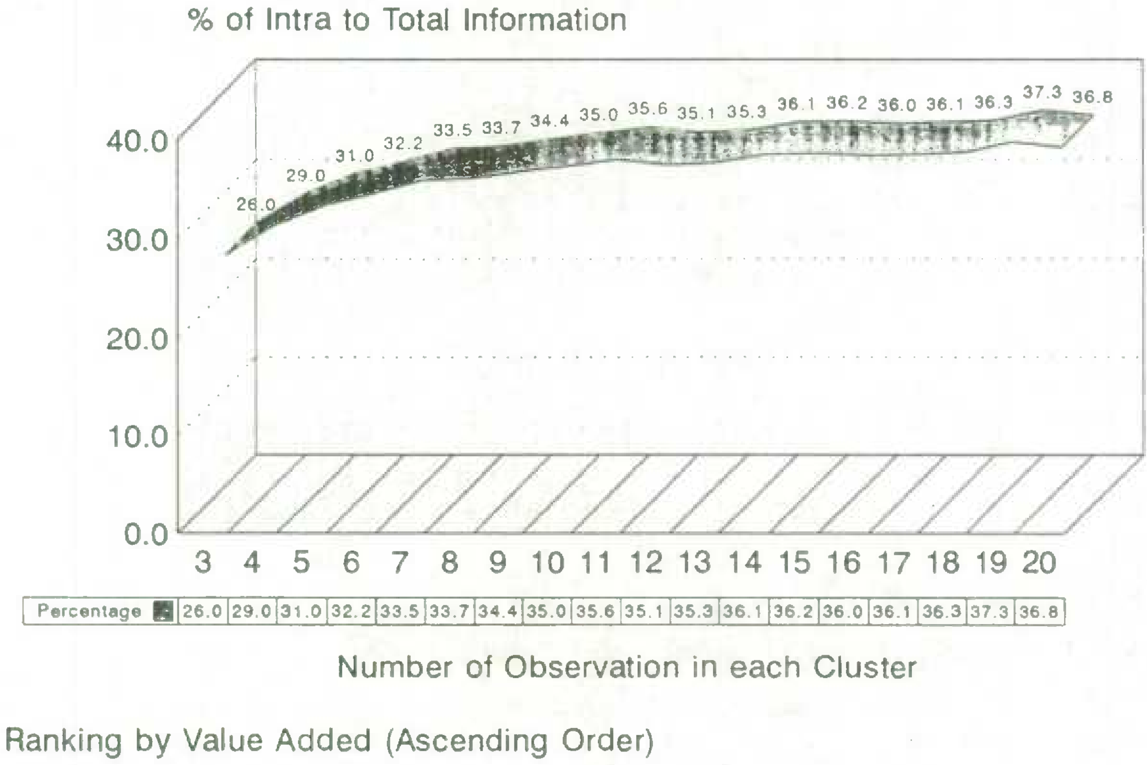


Figure 4: Percentage of Intra to Total Information (Outliers Deleted).

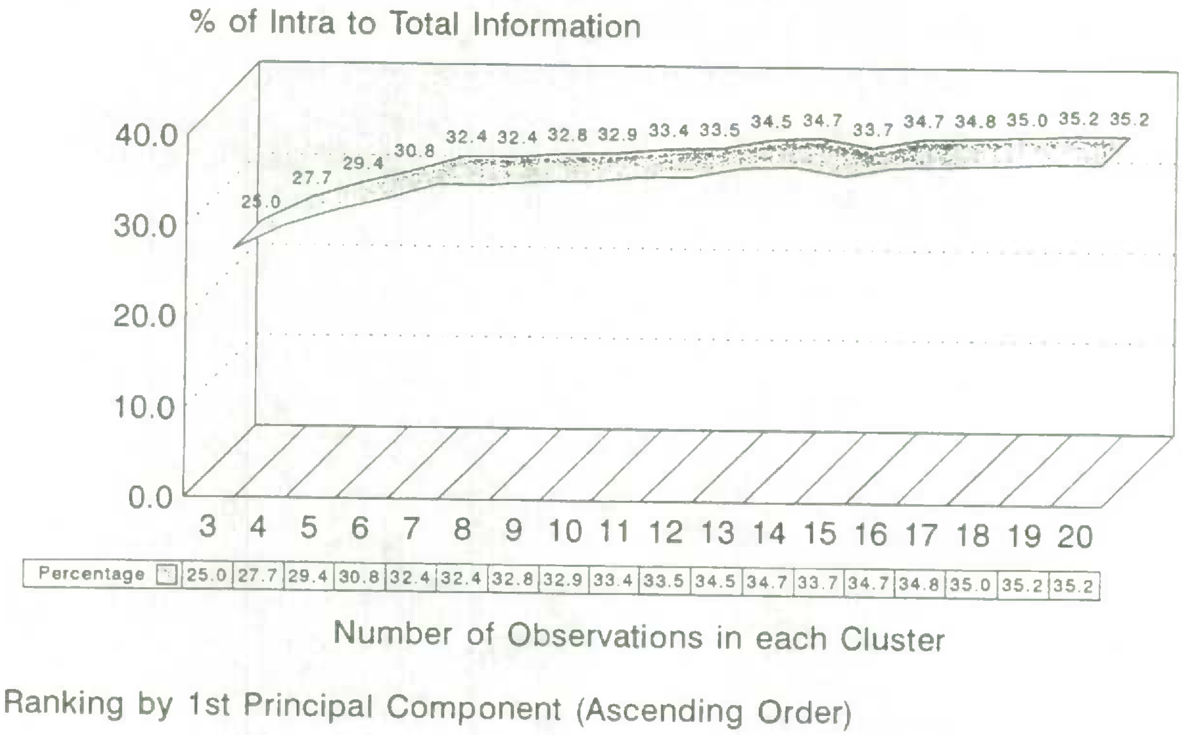


Figure 5 presents the correlations between the initial variables and their corresponding versions once the data have been aggregated 3 by 3 along the first principal component. The values indicate mainly how the aggregation process affects each variable.

Figure 5: Correlation between the *k*-grouped data and the original data.

Variable:	Correlation:
Persons employed	0.94357
Turnover	0.96582
Export sales	0.84760
Physical investments	0.81040
Value added	0.98181
Renumeration of employees	0.96182
Gross operating margin	0.86066
Value of sub-contracts awarded	0.70223
Expenditure on publicity	0.65148

REFERENCES

Adam, and Wortmann (1989). Security control methods for databases. A comparative study. *ACM Computing Surveys*, 21, 4.

Bragard, L., Roubens, M., Libert, J., and Gailly, B. (1988). Examen d'une méthode d'échantillonnage par la selection de prototypes. Eurostat internal report.

Hartigan, J.A. (1975). *Clustering Algorithms*, Wiley, New York.

Lehmann, E.L. (1959). *Testing Statistical Hypothesis*, John Wiley & Sons, New York.

Paass G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, 6, 4.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

THE CONTRIBUTION OF THE 'IFO INSTITUTE FOR ECONOMIC RESEARCH' TO PANEL DATA RESEARCH: THE STATE OF ART IN APPLIED AND METHODICAL RESEARCH

G. Nerb and H. Seitz[1]

ABSTRACT

For more than 40 years the IFO Institute is conducting its monthly business surveys (about 10,000 completed questionnaires in Industry trade and construction). Apart from the more traditional use of this type of information (indicator systems, incorporation into econometric models) several research projects have been launched aiming to exploit the informational content of micro data in longitudinal studies. It has been shown *e.g.* that industrial selling react only very weakly on unexpected changes of costs whereas unexpected demand changes lead to significant short term adjustments in production.

KEY WORDS: Longitudinal data; Panel estimation.

1. INTRODUCTION

Panel data or longitudinal data is a collection of observations on a (usually) large number of individuals, such as households or firms, over several periods of time (with the number of periods usually much smaller than the number of individuals). Economic analysis using this type of data has become increasingly popular in the last years. One reason for the growing interest in and importance of panel data research and analysis is the fact that in questionnaires, such as the IFO business survey, data on variables can be collected that are not covered by any official statistics, namely data on plans, appraisals, expectations, innovation activities, *etc.*, see for example Vogler (1977), Anderson and Strigel (1981), Oppenländer and Poser (1989).

Panel data provide a number of obvious advantages as compared to cross section or time series data, see for example Hsiao (1986) or Ronning (1991) for a more detailed discussion:

- Panel data sets usually contain a large number of observations which makes it possible to increase the efficiency of econometric parameter estimation considerably.

- Using individual data rather than aggregate data reduces much of the multicollinearity problems that are often encountered in econometric analysis.

- Compared to cross-section analysis, panel data make it possible to investigate dynamic patterns in the behavior of individual firms and households.

- Panel data make it possible to address and examine economic problems that cannot be approached with time-series data, at least without making rather severe a priori assumptions, especially because panel data collected in surveys provide information on variables, such as expectations, that are not covered in official statistics.

- Using panel data one can avoid the aggregation problem that plagues empirical work that exclusively relies on time series data, for an example see Seitz (1992).

[1] G. Nerb and H. Seitz, ifo Institut für Wirtschaftsforschung e.V., Postfach 860460, 8000 München 86.

- With respect to the interrelationship between economic theory and theory-based applied economic research, panel data make it possible to examine economic theories at the level at which these theories are formulated, that is, at the level of the individual household or firm, see Nerlove (1983).

2. MODELLING APPROACHES WITH PANEL DATA

Before we turn to a short discussion of the modelling approaches that have been developed for the handling of panel data we briefly comment on the character of the data collected in surveys. Virtually all of the data collected in the various IFO surveys, but also in other surveys, belong to one of the following categories:

i) *Continuous Variables:* The collection of continuous data in surveys is more an exception than the rule. Some examples of continuous data collected in IFO surveys are the rate of capacity utilization, the volume of inventories and order backlogs, *etc.*

iia) *Discrete Variables: Dichotomous (0,1) variables:* Dichotomous or binary data, such as the responses of firms whether or not they have realized process or product innovations, whether or not the firm plans to increase employment, *etc.* are typical examples of this type of data.

iib) *Discrete Variables: Trichotomous (+,=,-) variables:* Most data collected by the IFO Institute, especially in the IFO business survey, are trichotomous, that is, firms indicate whether a specific variable has or is expected to increase (+), stay about the same (=) or decrease (-).

As a matter of fact, discrete variables can have more than 3 categories, however, in most surveys one has to deal only with binary or trichotomous variables.

2.1 The aggregation approach

Within this approach researchers do not use the micro data for estimation but form aggregate time series out of the micro data. In the case of continuous data, one can for example derive estimates of the average capacity utilization rate, average stocks of inventories and order backlocks, *etc.* from the IFO business survey. Binary data can be used to get estimates of the percentage shares of firms that answer either 'yes' or 'no' to specific questions, such as the share of firms that have realized product or process innovations *etc.* For trichotomous variables, the most widely used concepts are the 'marginals' and the 'balance statistics'. The marginals are the percentages of firms that answer '+', '=', or '-' to a specific question; the balance is simply the difference between the shares of '+' and '-' responses.

Trichotomous survey data on expectations and plans of production activities, price changes *etc.* collected in many IFO surveys are used to generate quantitative proxy variables for time series of expectational variables. The appropriate technique, which is called 'quantification technique', for a survey see Seitz (1989a), has been developed by Anderson and Theil in the fifties, see Anderson (1951, 1953) and Theil (1966); modern refinements have recently been presented and applied to IFO panel data by Ronning (1986), Seitz (1987, 1988), and Tödter and Werfel (1988) to mention just a few. In addition, this kind of data is much used for the construction of leading business cycle indicators, see for example Dormayer and Lindlbauer (1984), Nerb (1989) and Entorf (1990, 1991).

2.2 Analysis of contingency tables and log-linear probability modelling

The first methodological approach that has been developed to exploit the availability of qualitative micro data has been the analysis of contingency tables and the estimation of log-linear probability models. A contingency table cross-classifies two or more discrete variables according to their joint occurrence. Statistical methods have been developed that can be used to calculate measures of associations between the qualitative variables involved, for an extensive discussion of the various measures of association see Reynolds (1977).

In a later stage of research the (descriptive) analysis of contingency tables has been improved by the estimation of log-linear probability models. This technique converts the entries of contingency tables into probabilities and estimates parametric models to explain the joint occurrence of specific variables, such as the probability that firms that have experienced an increase in demand, increase both their output price and the volume of production. For a methodical presentation of the log-linear probability model see the monographs by Nerlove and Press (1976), Fienberg (1977), and, for a short introductory survey with applications to the IFO business survey data, the paper by König, Nerlove and Oudiz (1982).

2.3 The panel estimation approach (micro level approach)

As has been set out above, the aggregation approach almost completely suppresses the micro-character of the panel data, and the log-linear-probability modelling approach is severely restricted with respect to the economic interpretation of the estimation results. The panel estimation approach is the only econometric technique that fully utilizes the panel character and the individuality of survey data. This approach makes it possible to consider individual-specific effects and the dynamics of individual behavior simultaneously. Recent developments in econometrics provide a wide variety of estimation techniques to handle panel data with both continuous as well as discrete variables.

3. APPLIED RESEARCH WITH IFO PANEL DATA

There is a substantial body of literature presenting applied economic research using data from the IFO business survey and the IFO innovation survey. The following overview classifies the papers and research approaches into the following categories:

- The modelling of expectations and plans,
- testing economic theories and investigating price, production, and inventory responses,
- research on R&D activities, market structure and the evaluation of labour market effects,
- ifo panel data and the construction of 'meso' indicators.

3.1 Modelling expectations and plans

The investigation of the formation of plans and expectations is crucial for our understanding of the working of the economic system because actions taken today are based on expectations with respect to the future development of the relevant variables. In a series of papers, König and Nerlove (1980, 1983) and König (1979, 1980) examined the formation of price expectations using micro data from the IFO business survey. The authors not only examined the question by which mechanism the formation of expectations can best be described but also investigated the joint formation of price and production expectations, thus expanding the univariate expectation formation model to a multivariate framework. Zimmermann (1986, 1988) also used micro data to examine the forecasting properties and the rationality of firms expectations, and compared French and German firms. His results strongly suggest that firms have biased expectations, in fact, expectations are downward biased, that is firms are too pessimistic.

3.2 Testing economic theories and investigating price, production and inventory responses

Many of the variables covered in the various IFO panel surveys provide information on expectations, plans and judgements of firms that are not covered by the data collection of official statistical offices. This creates the opportunity to test recently advanced modern economic theories, such as disequilibrium models or rational expectation models.

Modern theories of firm behavior that take disequilibrium phenomenon, such as market constraints and the stickiness of prices into account have been tested by König and Zimmermann (1983) and Kawasaki, McMilian and Zimmermann (1982, 1983). The examination of disequilibrium theories of the behavior of firms requests information on the status of firms in output and input markets. In the above quoted papers, König and Zimmermann and Kawasaki, McMilian and Zimmermann have demonstrated that this type of information can be derived from business survey data and successfully introduced in the estimation process.

Apart from testing assumptions on the behavior of individual firms a lot of applied research has been done on inventories. In the IFO business survey there are two questions related to inventories: Every month firms are asked whether they consider their current level of inventories as too large, about right or too small (inventory appraisal question). In addition, since 1981 the IFO Institute asks firms quarterly to indicate their volume of finished goods inventories measured in weeks of current production (quantitative inventory question). Similar questions are asked with respect to order backlogs. Both the qualitative as well as the quantitative inventory variables have been extensively used in empirical work. In an early stage of research, König and Nerlove incorporated both variables into log-linear probability models and treated both inventories and order backlogs as an additional response instrument of firms, see for example König and Nerlove (1984, 1986). König and Seitz (1989, 1991) estimated simultaneous fixed effects panel models in which production (measured by capacity utilization), inventories and order backlogs (both measured with the quarterly quantitative data) are taken into account and explained by a wide variety of exogenous variables as well as dynamic patterns.

3.3 Research on R&D activities, market structure and labour market effects

The IFO Institute regularly collects data on the innovation activities of firms in a special innovation test but also within the regular business survey test. These data have become increasingly popular in the area of R&D research. Most of the papers that are using IFO innovation test data also address the question of the impact of R&D activities on employment and market structure, for a more detailed survey on the usage of IFO survey data in innovation and employment research see Zimmermann (1990).

The theoretical and empirical analysis of the determinants of innovation activity and its relation to firm size and market power (measured by concentration ratios or the Herfindahl-Index) have been examined by König and Zimmermann (1986). The results of the König-Zimmermann study indicate that there exists a strong positive relationship between innovation activity and concentration ratios and that there is a negative impact of firm size on process innovation activities. However, the paper is inconclusive with respect to the relationship between product innovation and firm size because the estimation method used by König and Zimmermann did not allow the identification of parameters that describe this relation. The panel aspect is fully taken into account by Laisney, Lechner and Pohlmeier (1992a). These authors investigate a model of innovation activity using a five-wave balanced panel of 1,325 firms drawn from the IFO business survey for 1984 - 1988. This paper examines both cross section probit estimates and a panel probit estimator showing that the latter outperforms the cross section estimator.

3.4 Economic Indicators at the meso level derived from ifo panel data

Following the practise of the World Bank (1989) in the case of social indicators, one can also derive *meso indicators* out of survey data. Meso indicators refer to specific groups or segments of the society or the economy.

A theory-based segmentation of firms and the construction of appropriate meso indicators can be derived from special questions asked in the ifo business survey, for an example see Nerb (1992). The main idea of this approach adopts concepts provided by disequilibrium theory, see for example Malinvaud (1980). Disequilibrium theory classifies firms according to a combination of specific situations they face in the output and input markets and forms the following regime classification: the '*classical regime*' (that is, production is not demand-constrained; input prices - such as real wages - are too high to encourage additional production), the '*Keynesian regime*' (that is, production is constrained because of lack of demand), the '*repressed inflation regime*' (that is production is constrained because of tight input markets - *e.g.* not enough labour available - whereas there is still excess demand for goods), and the - rather implausible - '*underconsumption regime*' (that is, there are production constraints because of lack of demand and inputs).

Data collected in the ifo business survey can be used to design a typology of firms' characteristics that come rather close to these disequilibrium concepts. Using panel data one can track the evolvement of the various groups over the business cycle and thus one can get a deeper understanding of the nature of business cycles and can infer the kind of policy instruments that are appropriate to fight unemployment or high inflation rates.

Because at the present time, the economic development in the East of Germany is the central focus of German economic policy and because this development is closely watched by economists throughout the world, we briefly present an example of meso indicators by comparing the economic development in East and West Germany using disequilibrium meso indicators. To derive these indicators, we form the following four main groups, and, for West Germany only, 6 additional subgroups:[2]

Group 1: "Weakness in demand"

Subgroup 1.1 Transitory weakness in demand
Subgroup 1.2 Permanent weakness in demand

Group 2: "Balance", that is, no impediments on the supply and demand side.

Subgroup 2.1 In addition: business situation regarded as good or sufficient.
Subgroup 2.2 In addition: business situation regarded as bad.

Group 3: "Supply bottlenecks".

Subgroup 3.1 Transitory supply bottlenecks.
Subgroup 3.2 Pronounced supply bottlenecks.

Group 4: "Weakness in demand and supply bottlenecks".

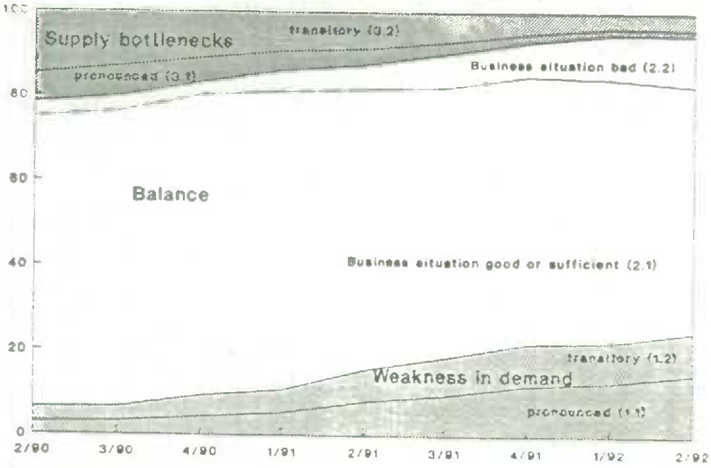
In figure 1 and 2 some results of this type of classification are demonstrated. Until now, the economic situation in the Neue Länder has improved only slightly. In July 1992, 28% of the industrial firms indicated that they had no impediments to production, neither from the demand nor from the supply side; in July 1990 this figure stood at 4%, rising to 17% in July 1991. The share of firms who see their current problems as exclusively demand constrained declined from 29% in July 1990 to 11% in July 1992. An increasing trend is shown by the group of firms who exclusively have problems on the supply side (20% in July 1990 and 33% in July 1992; above all insufficient product programme, financing difficulties). The largest group - nearly half of the firms - was at the beginning comprised of those industrial enterprises whose main problems derive equally from the supply and the demand side; in the meantime this share has decreased to 28% in July 1992. Such a constellation of business survey responses is otherwise only known for developing and newly industrializing economies. This shows that the present economic problems in the ex-GDR are more of a structural than a cyclical nature. In such a situation a traditional pump-priming programme should be ruled out. In West Germany, in spite of the current weakening in the economic situation, nearly three quarters of the firms are still in the balanced group (no impediments to production). The concurrent appearance of serious supply and demand problems is virtually non-existent (less than 1% of the West German companies). The percentage of West German firms that report supply-side problems has decreased from 22% in July 1990 to 5% in July 1992 indicating the weakening of the demand trend in West German economy.

4. CONCLUSIONS

Panel data, such as the IFO business survey data or the IFO innovation test data set provide a powerful instrument to carry out applied economic research. Economists throughout the world increasingly recognize the limits and drawbacks time series data modelling has. Economic actions are taken by individuals and not by any fictitious aggregate and consequently we have to introduce the individual - the firm or the household - in our models of economic behaviour. The rapid development of computer technology and the availability of easy to handle standard computer software makes the estimation of panel data model not more costly and complicated than the estimation of time series models 15 years ago. However, from the point of view of drawing policy implications from the analysis of micro data much work has to be done in the near future.

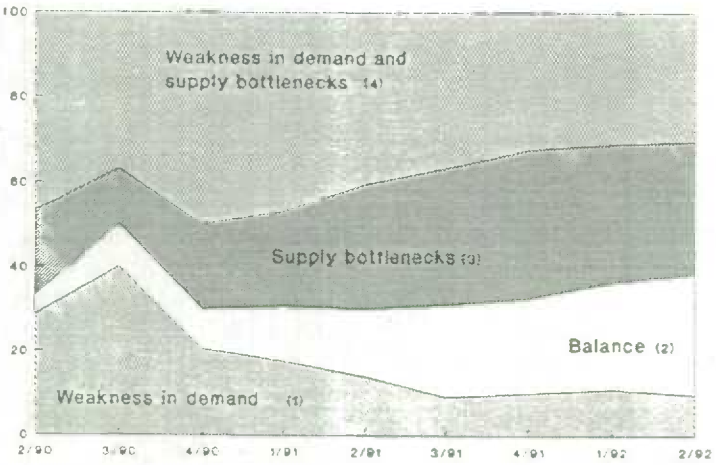
[2] For a detailed description of the following classification see the extended version of the present paper which is available upon request from the authors.

Figure 1: Regime classification of industrial companies West-Germany / Total industry



Share of companies in %; trimesters
Source: Ifo Business Survey

Figure 2: Regime classification of industrial companies East-Germany / Total industry



Share of companies in %; trimesters
Source: Ifo Business Survey

REFERENCES

Anderson, O. (1951). *Möglichkeiten und Grenzen einer Quantifizierung des Konjunkturtests des Münchener ifo Institutes für Wirtschaftsforschung*, in: Mitteilungsblatt für Mathematische Statistik, 3, 206-212.

Anderson, O. (1953). The business test of the IFO-Institut for economic research, Munich and its theoretical model. *Revue de L'institut International de Statistique*, 20, 1-17.

Anderson, O., and Strigel, W.H. (1981). Business surveys and economic research - A review of significant developments. H. Laumer and M. Ziegler (Eds.). *International Research on Business Cycle Surveys*, 25-54.

Dormayer, H.-J., and Lindlbauer, J.D. (1984). Sectoral indicators by use of survey data. K.H. Oppenländer and G. Poser (Eds.). *Leading Indicators and Business Cycle Surveys*, 467-498.

Entorf, H. (1990). *Multisektorale Konjunkturanalyse*, Campus Verlag.

Entorf, H. (1991). Das Ifo-Geschäftsklima, seine Komponenten und die Konjunkturprognose: Eine Regressionsstudie. *Ifo-Studien*, 37, 141-149.

Fienberg, S.E. (1977). *The Analysis of Cross-Classified Categorical Data*. Cambridge: The MIT Press.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge University Press.

Kawasaki, S., McMillan, J., and Zimmermann, K.F. (1982). Disequilibrium dynamics: An empirical study. *American Economic Review*, 72, 992-1004.

Kawasaki, S., McMillan, J., and Zimmermann, K.F. (1983). Inventories and price inflexibility. *Econometrica*, 51.

König, H. (1979). Zur Bildung von Preiserwartungen: Ein log-lineares multivariates Wahrscheinlichkeitsmodel. *Kyklos*, 32, 380-391.

König, H. (1980). Über den mikroökonomischen Zusammenhang zwischen Preiserwartungen und -realisationen. D. Duwendag and H. Siebert (Ed.). *Politik und Markt, Festschrift für Hans Karl Schneider, Stuttgart*.

König, H., and Nerlove, M. (1980). Micro-analysis of realisations, plans and expectations in the Ifo-business test by multivariate log-linear probability models. W.H. Strigel (Ed.). *Business Cycle Analysis, Westmead*.

König, H., and Nerlove M. (1983). Response of prices and production to unanticipated demand shocks: Some microeconomic evidence. K.H. Oppenländer and G. Poser (Eds.). *Leading Indicators and Business Cycle Surveys*, 349-384.

König, H., and Nerlove, M. (1984). A recursive log-linear probability model of production plans and price anticipation. An empirical investigation for French and German firms, D. Vitry and B. Marechal (Eds.). *Emploi-Chômage Modélisation et Analyses Quantitative, Collection de l'Institute de Mathématiques Économiques*, 28.

König, H., and Nerlove, M. (1986). Price flexibility, inventory behavior and production responses. W. Heller, R. Storr and D. Starrett. *Equilibrium Analysis, Essays in Honor of Kenneth J. Arrow*, II, 179-218.

König, H., Nerlove, M., and Oudiz, G. (1982). Die Analyse mikroökonomischer Konjunkturtest-Daten mit loglinearen Wahrscheinlichkeitsmodellen: Eine Einführung, *Ifo-Studien*, 28, 1, 155-191.

König, H., and Seitz, H. (1989). Zur Transmission von Nachfrage- und Kostenschocks auf Lagerhaltung, Preise und Produktion. *Jahrbücher für Nationalöknomie und Statistik*, 206, 421-433.

König, H., and Seitz, H. (1991). Production and price smoothing by inventory adjustment. *Empirical Economics*, 16, 233-252.

König, H., and Zimmermann, K.F. (1983). Mikroökonomische Preis- und Produktionsplanung im Ungleichgewicht. H. Enke, W. Köhler and W. Schulz (Eds.). *Struktur und Dynamik der Wirtschaft*, 147-160.

König, H., and Zimmermann, K.F. (1986). Innovations, market structure and market dynamics. *Zeitschrift für die gesamte Staatswissenschaft*, 142, 184-199.

Laisney, F., Lechner, M., and Pohlmeier, W. (1992a). Innovation activity and firm heterogeneity: Empirical evidence from Germany. Forthcoming in: *Economic Dynamics And Structural Change*.

Malinvaud, E. (1980). Macroeconomic rationing of employment. E. Malinvaud and J.P. Fitoussi, *Unemployment in Western Countries, MacMillan, 1980*.

Nerb, G. (1989). Zusammengesetzte Indikatoren und Indikatorsysteme, Chapter IV 2.1. K.H. Oppenländer and G. Poser (Eds.). *Handbuch der IFO-Umfragen*.

Nerb, G. (1992). Neuer Ansatz zur Analyse von Konjunkturtestdaten. *CIRET-Studie Nr. 44* (forthcoming).

Nerlove, M., and Press, J. (1976). Multivariate log-linear probability models for the analyses of qualitative data. Discussion Paper No. 1, Center for Statistics and Probability, Northwestern University, Evanston.

Nerlove, M. (1983). Expectations, plans, and realizations in theory and practice. *Econometrica*, 51, 1251-1279.

Oppenländer, K. H., and Poser, G. (1989). Handbuch der IFO-Umfragen. Duncker & Humblot, Berlin 1989.

Reynolds, H. T. (1977). The analyses of cross-classifications. New York: The Free Press.

Ronning, G. (1986). Econometric approaches to the estimation of indifference intervals in business tendency surveys. K.H. Oppenländer and G. Poser (Eds.). *Business Cycle Surveys in the Assessment of Economic Activity, Gower, Westmead (England)*, 175-209.

Ronning, G. (1991). *Mikroökonometrie*, Springer-Verlag Heidelberg.

Seitz, H. (1987). The estimation of inflation forecasts from business survey data. *Applied Economics*, 20, 427-438.

Seitz, H. (1988). An investigation into the reliability of business survey data. Discussion-Paper No. 358-387, Institut für Volkswirtschaftslehre und Statistik der Universität Mannheim.

Seitz, H. (1989a). Die Quantifizierung von Tendenbefragungs-daten: Ein Überblick, *Ifo-Studien*, 35, 1, 1-26.

Seitz, H. (1992). Still more on the speed of adjustment in inventory models: A lesson in aggregation. Discussion-Paper No. 377-388, Institut für Volkswirtschaftslehre und Statistik der Universität Mannheim. Forthcoming in: *Empirical Economics*.

Theil, H. (1966). Applied economic forecasting, North-Holland, Amsterdam.

Toedter, K.-H., and Werfel, M.C. (1989). Quantification of indifference responses from business surveys with mixed data. Forthcoming in: *IFO-Studien*.

Vogler, K. (1977). Content and determinants of judgemental and expectational variables in the Ifo business survey. W. H. Strigel (Ed.). *Problems and Instruments of Business Cycle Analysis*, 73-114.

Zimmermann, K. F. (1986). On rationality of business expectations: A micro-analysis of qualitative responses. *Empirical Economics 11*, 23-40.

Zimmermann, K. F. (1988). Prognosequalität von Surveydaten: Mikroökonomische Evidenz. W. Franz, W. Gaab and J. Wolters (Eds.). *Theoretische und angewandte Wirtschaftsforschung*, 261-274.

Zimmermann, K. F. (1990). Der IFO-Konjunkturtest in der arbeits- und industrieökonomischen Forschung. *IFO-Studien*, 36, 1-16.

SESSION 8

Data Analysis I

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

TESTING THE ROBUSTNESS OF ENTRY BARRIERS

J.R. Baldwin and M. Rafiquzzaman[1]

ABSTRACT

Longitudinal panel data from the Canadian Manufacturing sector are used to model the entry process and to investigate the existence of entry barriers. The paper investigates the robustness of previous findings by 1) using a variety of estimation procedures 2) testing different model specifications 3) varying the measures used to evaluate the importance of entry.

KEY WORDS: Entry barriers; Count data; Negative binomial.

1. INTRODUCTION

Since the seminal works of Bain (1956) and Modigliani (1958), the economic profession's attention has focused on the existence of entry barriers. The most widely-used models are the "Limit-Price" models of entry in which it is argued that the profit levels above which entry is attracted differ industry by industry and are a function of entry barriers. Incumbents can set higher prices in industries with high barriers without attracting entry. The level of price above which entry occurs is the limit-price.

Implicit in the "Limit-Price" model is the view that an entrant augments existing output. Alternately, the "Stochastic-Replacement" view of entry is based on the assumption that entry is a dynamic process involving both partial and complete replacement of existing firms by entrants (Baldwin and Gorecki 1983). The "Replacement" view of entry presumes that entry can be expected even when price equals long-run average cost and industry profits are zero.

Limit-price entry models purport to confirm the existence of entry barriers and, therefore, the existence of market imperfections.[2] The models that combine the stochastic-replacement phenomenon with the limit-price model, however, find that entry barriers are much less important (Baldwin and Gorecki 1987).

As is often the case in applied economics, interpretation of the significance of these differences is complicated by the fact that previous studies differ not only in terms of choice of model but also in terms of how entry is measured--units of observation, units of measurement, type of entrants, and time period.

This paper investigates the extent to which the findings of the importance of entry barriers depends on the measurement of entry. It does so by investigating the extent to which estimation techniques, model specification, and measurement alter the conclusion that entry is detrimentally affected by certain industry structural characteristics.

[1] J.R. Baldwin and M. Rafiquzzaman, Business and Labour Market Analysis, Analytical Studies Branch, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

[2] See Cable and Schwalbach 1991 for a survey of the Orr-type results.

2. THE MEASUREMENT OF ENTRY

Entry can be defined as the birth of a producing unit--a new plant or a new firm. In the first case, entry is defined as a new plant in a particular industry. In the second case, it is defined as a new firm with a new producing unit--a greenfield entrant. Entry can also be defined as the birth of a new legal entity. New legal entities may be associated with the birth of new plants; but they also include firms which enter an industry by acquiring existing firms. Entry can be defined either in gross or in net terms. In the former case, it is the total number of plants (firms) entering an industry in a given period. In the latter case, it is the difference between the number of plants (firms) in two different periods.

This paper uses a definition of entry implicit in the limit-price entry literature--greenfield entrants. This is a relatively homogeneous category. The definition uses greenfield entrants rather than both greenfield and merger entrants, since the latter do not, initially at least, augment industry output. It focuses on gross measures rather than net entry measures which combine the effect of both entry and exit. It measures entry as new firms that build new plants and not new plants per se. The latter include both greenfield entrants and also plants that are built by incumbent firms. Failure to distinguish between new firms and existing firm new plant activity confuses entry with the expansion decisions of continuing firms. Evidence (Baldwin and Gorecki 1983, 1987) shows substantial dissimilarities in the determinants of greenfield and merger entrants, of greenfield entrants and the plant creation process by continuing firms, of gross and net entry measures.

The importance of greenfield entrants is measured here both in terms of their number or their size. The strength of the competitive forces associated with entry probably depends both on the number of entrants and the share of a market that is captured by the entrants. The percentage of shipments in a market captured by entrants is equal to the count rate of entry times the average size of entrants relative to the population. In order to test the robustness of the results, all three were used--the count of entrants, the shipments of entrants, and the average size of entrants--as dependent variables in the regression analysis.

The importance of entry to the competitive process and market performance is also evaluated in this paper by using both short- and long-run data. Short-run estimates were derived by using two adjacent points of time; long-run estimates, by using two points removed from one another in time. Short-run rates of entry were estimated for each year between 1970 and 1979 and averaged. Long-run rates of entry were calculated as the number of firms in 1979 that had entered the industry since 1970. It is the total entry of all firms in each year since 1970 minus the deaths of entrants over this period. Short and long-run entry were estimated at the 4-digit SIC level for the Canadian manufacturing sector using a longitudinal data base that followed firms and plants between 1970 and 1979. A description of the file can be found in Baldwin and Gorecki (1990).

Finally, a fourth measure--the success rate of entrants--was calculated and regressed against the same set of explanatory variables used for entry counts, shipments and average size. The success rate is defined as the long-run entry count divided by the sum of the short-run entry count. This is the proportion of all entrants over a decade that are still alive at the end of the period and is a direct measure of population continuance or an inverse measure of infant mortality.

3. MODELS OF ENTRY

The most common entry models follow the earlier work of Orr(1974), which posits that entry will occur whenever profits are above their entry-precluding levels. Following Orr, the model of entry is

$$E_{it} = f(P_{it} - P^*_{it}), \qquad (1)$$

where E_{it} is the entry into industry i at time t, P_{it} is the entrant's perceived post-entry profit and P^*_{it} is the entry-precluding profit in industry i at time t.

The entry-precluding profit, P^*_{it}, depends on a vector of entry barriers, B, and a market risk variable, R. P^* (ignoring time and industry subscripts) can be specified as $P^* = h(B,R)$. As a consequence, the entry model in (1) may be written as

$$E = f_1(P,B,R). \tag{2}$$

E is expected directly to vary with perceived post-entry profit, P, and negatively with every component of B and R. It is, therefore, hypothesized that profit induces entry, whereas barriers to entry and risk reduce entry.

In our estimation, expected profitability (P_{it}) is represented by two variables. The first (PR) captures the average profitability of continuing firms. Since the average does not incorporate information about the trend that might be expected to influence expectations about future profits, the growth in profits (GP) over the period is also included. Entry is expected to be greater in those markets where profits are growing.

Barrier variables (B) are represented by economies of scale (MES), concentration (CON), advertising intensity (AD), and research and development intensity (RD). Market risk (R) is represented by volatility of market growth (VMG).

Equation #2 is an incomplete specification of entry since it does not consider stochastic aspects of entry. According to the "stochastic-replacement" view, a substantial amount of entry simply replaces existing firms, and occurs even if economic profits are zero. This flow of entry by replacement is posited to depend upon the size of the market. When firm counts are used to measure entry, size of the market (S) is taken to be the number of firms in the industry (N). When shipments of entrants is used as the dependent variable, size of the market is measured in terms of total industry shipments (TVS). The effect of the market-size variable can be interacted with the barrier variables to determine whether the magnitude of stochastic replacement is affected by the barrier variables.

The amount of entry should also depend on how easily entrants can enter and capture part of the market. A market with rapid growth will be associated with new consumers and, therefore, there is a greater likelihood that new firms will be gaining market share. Thus, industry growth, G, is added to equation #2. Following Baldwin and Gorecki (1983, 1987) an entry model which incorporates both stochastic-replacement and limit-price views of entry is specified as

$$E = g(S,G,P,B,R). \tag{3}$$

S, G, and P provide incentives to entry, whereas B and R provide disincentives.

4. ESTIMATION PROCEDURES APPROPRIATE TO COUNT DATA

Many previous empirical studies of entry have estimated a linear and/or a loglinear version of the entry equation (3), or something close to it, and the OLS method of estimation has mainly been used. Since entry data are integer-valued and deviate from classical regression assumptions, the statistical specification of entry calls for a discrete probability distribution. In order to meet this requirement, we specify and estimate an econometric model of entry based first, on the assumption that each observation is drawn from a Poisson distribution, and then, from a Negative Binomial distribution. Our methodology is in the spirit of Hausman, Hall and Griliches (1984), and Cameron and Trivedi (1986), who apply both the Poisson and Negative Binomial regression to count data on firms' patenting activity and consumer demand for health care services, respectively. It is also in the spirit of Chappell, *et al.* (1990), Mayer and Chappell (1992), and Papke (1991) who use Poisson regressions to study the entry behaviour of firms across industries in the U.S.

Under the assumption that the data on entry are drawn from a Poisson distribution, the probability of observing a count of entry E_i in industry i is

$$Pr(E_i) = Exp(-\lambda_i)\lambda_i^{E_i}/E_i!, \quad E_i = 0,1,2,..., \tag{4}$$

The mean and variance of E_i are equal to λ_i. To incorporate the explanatory variables, X_i which influence entry, the parameter λ_i is specified to be

$$\lambda_i = E(E_i\,|\,X_i) = Exp(X_i\beta), \tag{5}$$

where $X_i = (S,G,P,B,R)$ and β is a parameter vector to be estimated. The restriction of the equality of the mean and variance of the Poisson distribution can be overcome by following Gourieroux, Monfort and Trognon (1984a,b) and using a specific version of the Negative Binomial.[3]

Initially, OLS, Poisson, and Negative Binomial regressions were all estimated. The results are reported in Table 1. The estimated standard errors of the Poisson and Negative Binomial models are substantially lower than those of OLS estimates. These findings are consistent with those of Hausman, Hall and Griliches, and Cameron and Trivedi.

Although the Poisson point estimates and those in the Negative Binomial model are similar in sign and magnitude, the estimated standard errors under the Poisson model are substantially smaller.

In order to test the null hypothesis that the underlying model is Poisson against the alternative that the model is a Negative Binomial, both the Wald and the Likelihood Ratio test were used. Both test statistics were highly significant. The data also rejected the equality of the mean and the variance which is the key property of the Poisson model. Therefore, the Poisson model was rejected in favour of the Negative Binomial.

5. RESULTS

The first section investigates the extent to which the effect of entry-barrier variables on the number of entrants is robust to whether the limit-price as opposed to the stochastic-replacement model is used. The second section compares the effect of entry-barrier variables for alternative measures of entry by using four different dependent variables--number of entrants, shipments of entrants, average size of entrants, and population continuance rates of entrants.

5.1 The Limit-Price Versus the Stochastic-Replacement Models

Choice of the Negative Binomial over the OLS overcomes earlier observations (Baldwin and Gorecki 1987) that entry barriers are positive but insignificant determinants of the entry process. The choice of the integer-count model substantially increases the significance of the effect of concentration and plant scale economies variables.

The OLS procedure generates three significant variables in both the short and the long run. (See Table 1). The entry process is positively related to the existing number of firms (N), the growth of shipments (GS) and risk, measured in terms of volatility of growth (VMG). Other variables are not statistically significant.

Each of the variables that are significantly[4] different, from zero in the OLS estimate are also significant and have the same sign in the Negative Binomial regression. In addition, the Negative Binomial has a significant and negative coefficient for two entry-barrier variables in both the short and the long run that were not significant in the OLS equation. These are plant scale economies (MES) and concentration (CON). Research and

[3] We use a specific form of the Negative Binomial whose mean equals Exp($X\beta$) and whose variance is Exp($X\beta$)[1 + α Exp($X\beta$)]. See Cameron and Trivedi (1986) for details.

[4] In this paper, 5% is used for the level of significance.

development has a positive affect on entry in both the long and the short run but is only significant in the short run. Advertising is not significant either in the OLS or the Negative Binomial model.

When the model specification is varied these conclusions do not change. Three variations are reported in Table 2. As in Table 1, columns 1 to 3 represent the long-run results; columns 4 to 6 the short-run results. The first equation used just the variables that originate from a simple Orr-type model. These are profitability (PR), growth in profitability (GP), growth in sales (GS), concentration (CON), economies of scale (MES), research and development (RD), advertising intensity (AD), and demand variability (VMG). The second formulation added the industry size variable--number of firms (N) to the first. The third formulation added an interaction term involving industry size and the entry-barrier variables--concentration (CON), advertising intensity (AD), research and development intensity (RD), and economies of plant scale (MES).

The barrier variables that are significant in the simple limit-price model (cols. 1 and 4) are also significant in the two models that incorporate the stochastic-replacement phenomenon (columns 2 and 3; columns 5 and 6). Concentration (CON) and scale economies (MES) negatively affect entry in all formulations.

Nevertheless, the size of the coefficient on the concentration variable decreases by about 50% when the stochastic-replacement model in columns 2 and 5 is used. Moreover, in the third variant (columns 3 and 6), the fact that scale economies has a positive coefficient when interacted with number of firms means that the effect of scale economies declines as the number of firms in the industry gets larger. Indeed, for the long-run results, when the number of firms (N) is greater that 30, scale economies will not have a negative impact on entry. If the same exercise were conducted with concentration, the break-even value is about 47 firms. Barriers matter--but not where firm numbers are relatively high.

It is also noteworthy that advertising is weakly significant when interacted with the size of an industry. The rate of stochastic-replacement is lower in industries with high advertising-sales ratios. This is in marked contrast to the first column where it was not found to have a significant impact when included as a proxy for an entry barrier that affected the level of limit entry profits.

5.2 Alternative Measures of the Importance of Entry

In the previous section, entry is measured by the count of new firms. This section examines whether the determinants of entry remain the same when the unit of measurement that is used to define entry is changed. Four separate measures of entry are used to compare the robustness of our findings on the importance of entry barriers. Each of these is measured for both the long run and short run.

These are:

(1) the number counts of greenfield entrants (E),

(2) the amount of shipments by entrants (TVSE),

(3) the average size of entrants (ASE), and

(4) the ratio of the number of entrants in the long-run to the number of entrants in the short-run (RATIO).

Each dependent variable is regressed on the same set of explanatory variables, with one exception. The normalizing variable for the numbers count is the number of firms in the industry (N); for entry shipments (TVSE), it is total values of industry shipments (TVS); for the average size of entrants (ASE), it is the average size of existing firms (ASF); for the population continuation rate, RATIO, it is the average size of entrants relative to the industry average size (RELSIZE).

Table 1: Comparison of estimation procedures for entry model [1], [2]

	LONG RUN			SHORT RUN		
	OLS	POISSON	NEGATIVE BINOMIAL	OLS	POISSON	NEGATIVE BINOMIAL
Constant	- 4.995 (0.500) [7.380]	3.587 (0.000) [0.087]	2.848 (0.000) [0.327]	- 73.218 (0.338) [76.190]	6.508 (0.000) [0.022]	6.005 (0.000) [0.359]
N	0.294 (0.000) [0.015]	0.003 (0.000) [0.000]	0.005 (0.000) [0.043]	2.482 (0.000) [0.066]	0.001 (0.000) [0.000]	0.003 (0.000) [0.002]
PR	1.479 (0.598) [2.795]	0.121 (0.000) [0.025]	0.094 (0.520) [0.146]	9.946 (0.744) [30.390]	0.108 (0.000) [0.008]	0.145 (0.510) [0.221]
GP	0.166 (0.911) [1.480]	- 0.045 (0.023) [0.020]	- 0.062 (0.372) [0.070]	- 2.598 (0.872) [16.030]	- 0.054 (0.000) [0.005]	- 0.080 (0.263) [0.072]
GS	0.883 (0.046) [0.439]	0.083 (0.000) [0.006]	0.090 (0.000) [0.022]	14.653 (0.002) [4.675[	0.081 (0.000) [0.002]	0.089 (0.000) [0.019]
CON	- 0.074 (0.427) [0.093]	- 0.023 (0.000) [0.001]	- 0.018 (0.000) [0.004]	- 1.280 (0.196) [0.986]	- 0.027 (0.000) [0.000]	- 0.030 (0.000) [0.004]
MES	1.466 (0.952) [24.060]	- 2.971 (0.000) [0.490[	- 2.038 (0.033) [0.955]	- 77.137 (0.768) [260.718[	- 5.684 (0.000) [0.174]	- 2.158 (0.014) [0.876]
RD	0.00004 (0.999) [0.074]	0.006 (0.000) [0.001]	0.005 (0.362) [0.006]	1.191 (0.141) [0.806]	0.011 (0.000) [0.000]	0.013 (0.005) [0.005]
AD	- 77.680 (0.311) [76.360]	- 9.904 (0.000) [1.357]	- 3.080 (0.399) [3.649]	- 515.700 (0.533) [824.740]	- 6.297 (0.000) [0.335[	- 0.408 (0.902) [3.298]
VMG	0.084 (0.018) [0.035]	0.003 (0.000) [0.001]	0.004 (0.012) [0.002]	1.247 (0.001) [0.382]	0.003 (0.000) [0.000]	0.003 (0.046) [0.002]
Variance Parameter α			0.372 (0.000) [0.043]			0.487 (0.000) [0.057]
Adj R^2	0.81			0.93		
- Log L		1016.099	591.998		8711.917	997.528

[1] The significance levels of a two-tailed test for rejecting the null hypothesis that the coefficient is zero are given in parentheses.

[2] The associated standards errors of the estimates are reported in brackets.

Table 2: Comparison of different entry models: Negative binomial estimates[1],[2].

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	4.561 (0.000) [0.292]	2.848 (0.000) [0.327]	2.498 (0.000) [0.257]	7.944 (0.000) [0.284]	6.005 (0.000) [0.359]	5.409 (0.000) [0.303]
PR	0.083 (0.534) [0.133]	0.094 (0.520) [0.146]	0.064 (0.626) [0.131]	0.031 (0.846) [0.160]	0.145 (0.510) [0.221]	0.162 (0.389) [0.188]
GP	- 0.107 (0.195) [0.083]	- 0.062 (0.372) [0.070]	- 0.071 (0.256) [0.062]	- 0.103 (0.220) [0.084]	- 0.080 (0.263) [0.072]	- 0.073 (0.257) [0.065]
GS	0.131 (0.000) [0.025]	- 0.090 (0.000) [0.022]	- 0.056 (0.001) [0.017]	- 0.108 (0.000) [0.0022]	- 0.089 (0.000) [0.192]	0.052 (0.002) [0.016]
CON	- 0.036 (0.000) [0.004]	- 0.018 (0.000) [0.004]	- 0.014 (0.002) [0.005]	- 0.049 (0.000) [0.004]	- 0.030 (0.000) [0.004]	- 0.025 (0.000) [0.004]
MES	- 2.585 (0.032) [1.210]	- 2.038 (0.033) [0.955]	- 6.935 (0.000) [1.414]	- 2.160 (0.039) [1.047]	- 2.158 (0.014) [0.876]	- 5.642 (0.000) [1.082]
RD	- 0.009 (0.185) [0.006]	0.005 (0.362) [0.006]	0.005 (0.411) [0.006]	0.016 (0.001) [0.005]	0.013 (0.005) [0.005]	0.005 (0.147) [0.004]
AD	- 6.206 (0.116) [3.952]	- 3.080 (0.400) [3.649]	0.356 (0.941) [4.839]	- 4.552 (0.231) [3.803]	- 0.408 (0.902) [3.298]	- 0.696 (0.882) [4.695]
VMG	0.002 (0.370) [0.021]	0.004 (0.012) [0.002]	0.004 (0.002) [0.001]	- 0.001 (0.471) [0.002]	0.003 (0.046) [0.001]	- 0.004 (0.005) [0.001]
N		0.005 (0.000) [0.001]	0.004 (0.000) [0.001]		0.003 (0.000) [0.0002]	0.002 (0.002) [0.0005[
CON X N			0.00003 (0.487) [0.00005]			0.00002 (0.368) [0.00002]
AD X N			- 0.076 (0.064) [0.041]			- 0.015 (0.636) [0.031]
RD X N			- 0.0007 (0.313) [0.0007[			0.000003(0.967) [0.00008]
MES X N			0.231 (0.000) [0.042]			0.144 (0.000) [0.024]
Variance Parameter α	0.565 (0.000) [0.065]	0.372 (0.000) [0.043]	0.201 (0.000) [0.026]	0.713 (0.000) [0.086]	0.487 (0.000) [0.057]	0.311 (0.000) [0.039]
- Log L	620.304	591.998	556.235	1033.470	997.528	958.751

[1] The significance levels of a two-tailed test for rejecting the null hypothesis that the coefficient is zero are given in paratheses.

[2] The associated standards errors of the estimates are reported in brackets.

Table 3 presents the results for the long run. The short-run results are qualitatively the same. The model was estimated using the Negative Binomial for the entry count measure and OLS for other variables.[5]
A comparison of the entry count data and the industry shipments equations reveals that the former has less explanatory power. A simple OLS of the counts data (not reported here) has a considerably higher adjusted R^2 than the OLS for shipments data--.81 as opposed to .41 respectively. This is because the explanatory variables do a poor job of describing the average size of entrants. The adjusted R^2 for the equation using the average size of entrant as the dependent variable was only .32. Despite this difference, most of the significant coefficients in the count and shipments equations tell the same story. Entry depends positively on size, growth of shipments, and negatively on concentration.

While these variables tell basically the same story for the count and shipments data, they do not always affect the average size of an entrant (ASE) in the same way. Growth positively affects average size of the entrant (ASE) and the success rate (RATIO), but in neither case it is very significant. Growth thereby affects the importance of entrants because it affects the number of entrants and not because it facilitates entrants of a larger average size. Growth also positively affects the success rate (RATIO), but the coefficient is not significant.

Concentration also has a different effect on entry counts than on average size. Higher concentration leads to fewer entrants but it has a positive but insignificant effect on the average size of an entrant. There are, therefore, fewer entrants in concentrated industries but the entrants tend to be bigger--probably because the cost disadvantages of small scale entry are greater in these industries. Concentration is also associated with a significant positive effect on RATIO--the success rate of entrants.

While profits--either PR or GP--are rarely significant, there is a difference in the sign associated with this variable in the counts and shipments equations. Profits are positively related to the number of entrants but negatively related to shipments. This occurs because higher profits allow more entrants to penetrate an industry, but these entrants are smaller.

In conclusion, the entry-barrier variables that have been so often stressed in the literature are less important when we turn to measure the impact of entry in other than simple count terms. It may be that there are fewer entrants in concentrated industries, but the entrants in these industries tend to be larger and, therefore, concentration does not have as great an effect on shipments captured by entrants as it does on their numbers. Moreover, entrants in concentrated industries have greater staying power. The number of entrants that survive is greater in concentrated industries.

6. CONCLUSIONS

This paper investigates the robustness of conclusions that certain structural characteristics are barriers to entry. As often happens during robustness tests, we have learned not just whether a variable matters or whether a phenomenon exists, but instead the circumstances in which it matters. We have found that when we apply a more complex estimation procedure--regression for count data--that the effect of entry barriers is more easily separated from other variables. We have also seen that extending the model confirms the importance of barriers but qualifies the findings to hold only for industries in which there is a small number of firms. This exercise has therefore confirmed that barriers have a non-linear effect that may not extend across all industries. Finally, we have seen that barriers have a different effect on the number of entrants than on the average size of entrants and thus on the market share that entrants capture. Structural barriers reduce the number of entrants but in some cases this is partially offset by larger average size of the entrant.

[5] The negative binomial was also used for the shipments variable and a logistic transformation of the relative size and the population success rate was also employed; in all cases, the same qualitative results were obtained.

Table 3: Comparison of different entry measures: long run[1],[2].

	E[a]		TVSE		ASE		RATIO	
Constant	2.923 [0.280]	(0.000)	19.207 [4.529]	(0.000)	8.837 [5.828]	(0.132)	0.051 [0.027]	(0.060)
PR	0.125 [0.152]	(0.411)	- 0.863 [2.090]	(0.680)	- 2.514 [2.686]	(0.351)	- 0.010 [0.012]	(0.383)
GP	- 0.033 [0.063]	(0.602)	- 2.348 [1.152]	(0.043)	- 2.854 [1.487]	(0.057)	- 0.008 [0.007]	(0.246)
GS	0.082 [0.018]	(0.000)	1.024 [0.327]	(0.002)	0.621 [0.417]	(0.138)	0.003 [0.002]	(0.153)
CON	- 0.018 [0.004]	(0.000)	- 0.015 [0.062]	(0.015)	0.095 [0.082]	(0.249)	0.002 [0.000]	(0.000)
MES	- 1.806 [0.872]	(0.038)	- 27.703 [18.060]	(0.127)	- 45.760 [23.204]	(0.051)	- 0.086 [0.103]	(0.406)
RD	0.004 [0.005]	(0.377)	0.094 [0.056]	(0.095)	0.168 [0.071]	(0.020)	- 0.001 [0.000]	(0.032)
AD	- 3.103 [3.069]	(0.312)	- 78.861 [56.700]	(0.166)	- 55.002 [72.518]	(0.450)	- 0.308 [0.318]	(0.335)
VMG	0.003 [0.001]	(0.012)	0.006 [0.027]	(0.828)	0.007 [0.034]	(0.984)	0.0003 [0.000]	(0.096)
N	0.005 [0.0004]	(0.000)						
TVS			0.000004 [0.000]	(0.000)				
ASF					0.0001 [0.000]	(0.001)		
RELSIZE							- 0.00002 [0.000]	(0.195)
Adj R^2			0.41		0.32		0.25	
F			13.35		9.32		6.92	
Degrees of freedom			(9,148)		(9,148)		(9,148)	

[1] The significance levels of a two-tailed test for rejecting the null hypothesis that the coefficient is zero are given in parentheses.

[2] The associated standards errors of the estimates are reported in brackets.

[a] Excludes all zero values of the dependent variable.

APPENDIX: DESCRIPTION OF VARIABLES

PR The gross rate of return on capital defined as total activity value added less total activity value of wages and salaries, stock for 1970.

GP The ratio of the largest firm (top half of employment) weighted profit rate in 1979 to 1970, where profit rate is defined as the weighted margins/sales ratio.

GS The growth rate for real total activity value of shipments between 1970 and 1979.

CON 4-firm concentration ratio index.

MES The market share (in terms of shipments) of the smallest enterprise required to account for 50 percent of industry employment.

RD The ratio of research and development personnel to all wage and salary earners.

AD The advertising-sales ratio.

VMG The volatility of market growth, defined as the standard error of the residuals taken from a regression of the logarithm of shipments on time.

N The existing number of firms in an industry.

TVS The value of total activity shipment of all firms in an industry

ASF The average size of all firms in an industry measured in terms of shipment.

RELSIZE The average size of entrants relative to the industry average size measured in terms of shipments.

REFERENCES

Bain, J.S. (1965). *Barriers to New Competition*. Cambridge, MA: Harvard University Press.

Baldwin, J.R., and Gorecki, P.K. (1983). Entry and exit to the canadian manufacturing sector: 1970-1979. Discussion Paper # 225. Ottawa: Economic Council of Canada.

Baldwin, J.R., and Gorecki, P.K. (1987). Plant creation versus plant acquisition: The entry process in Canadian manufacturing. *International Journal of Industrial Organization*, 5, 27-41.

Baldwin, J.R., and Gorecki, P.K. (1990). Measuring firm entry and exit with panel data. *Analysis of Data in Time Symposium '89 proceedings*, (eds. A.C. Singh and P. Whitridge) sponsored by Statistics Canada, Ottawa and Carleton Universities, Ottawa, Ontario, 255-270.

Cable, J., and Schwalbach, J. (1991). International comparisons of entry and exit. *Entry and Market Contestability: An International Comparison*, (eds. P.A. Geroski and J. Schwalbach), Oxford: Blackwell, 1991, 257-281.

Cameron, A.C., and Trivedi, P.K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29-54.

Chappell, W.F., Kimenyi, M.S., and Mayer, W.J. (1990). A Poisson probability model of entry and market structure with an application to U.S. industries during 1972-77. *Southern Economic Journal*, 56, 918-927.

Geroski, P.K., and Schwalbach, J. (1991). *Entry and Market Contestability: An International Comparison*. Oxford: Blackwell.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52, 681-700.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica*, 52, 701-720.

Hausman, J., Hall, B.H., and Griliches, Z. (1984). Econometric models for count data with an application to the Patents-R&D relationship. *Econometrica*, 52, 909-938.

Mayer, W.J., and Chappell, W.F. (1992). Determinants of entry and exit: An application of the compounded bivariate Poisson distribution to U.S. industries, 1972-1977. *Southern Economic Journal*, 58, 770-778.

Modigliani, F. (1958). New developments on the oligopoly front. *Journal of Political Economy*, 66, 215-232.

Orr, D. (1974). The determinants of entry: A study of the manufacturing industries. *Review of Economics and Statistics*, 56, 58-66.

Papke, L.E. (1991). Interstate business tax differentials and new firm location: Evidence from panel data. *Journal of Public Economics*, 45, 47-68.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

FOLLOWING CHILDREN OVER TIME: CHILD DEVELOPMENT AND ITS LINKAGES WITH FAMILY SOCIAL AND ECONOMIC TRANSITIONS

P.C. Baker and F.L. Mott[1]

ABSTRACT

Using data from the NLSY Child surveys, this paper explores how family poverty and maternal employment are linked to changes in childrens' cognitive and behavioral outcomes. The analyses employ a change score approach to first assess short-term changes in child outcomes between two successive data points (1986 to 1988 or 1988 to 1990) and levels of family conditions in that interval, controlling for prior individual and family attributes. Results are then provided for a longer period over three survey points from 1986-1990 which suggest that cognitive and socioemotional change for children can vary, depending on the duration of time between base and end point as well as other factors such as the child's race and maturational level.

KEY WORDS: Child assessment; Family change; Change scores.

1. INTRODUCTION

In an effort to explore the relationship between family conditions and child outcomes, this paper examines how family socioeconomic factors are associated with changes in childrens' behavior and their performance on standard mathematics and reading achievement tests. Many studies have established the importance of parental attributes in the development of a child's intellect, both through the effects of differential parenting traits and varying ability to provide the requisite foundations for learning. Recent evidence suggests that a family's ability to offer economic security and a home environment characterized by appropriate levels of cognitive stimulation and emotional support differentially affect the way children perform in school and acquire social skills (Parcel and Menaghan 1990).

While other researchers have looked at the associations between family attributes and child outcome levels, few have explored the extent to which family conditions are linked with *changes* in child well-being. The National Longitudinal Surveys of Youth (NLSY) and its linked child assessment information offer unique opportunities to explore connections over time between family and maternal traits and child development. The NLSY contains repeated measures of child cognition and socioemotional well-being for a large U.S. sample of children for three points in time, 1986, 1988 and 1990. The analyses presented here first use a pooled sample to assess changes in child outcomes between two successive data points (1986 to 1988 or 1988 to 1990) and levels of socioeconomic well-being in that interval, controlling for prior family attributes. Subsequent results based on a more restricted sample over three assessment points in the 1986-1990 period suggest that cognitive and socioemotional change for children can vary, depending on the duration of time between base and end point as well as other factors such as the child's maturational level.

[1] P.C. Baker and F.L. Mott, Center for Human Resource Research, 921 Chatham Lane, Suite 200, The Ohio State University, Columbus, Ohio, U.S.A.

2. METHOD

2.1 The NLSY Child Data

The National Longitudinal Surveys of Youth (NLSY) have tracked a national sample of more than 12,000 individuals through extensive annual interviews since 1979. The cohort, about evenly divided between males and females, is a national probability sample of the noninstitutionalized population aged 14 to 21 in the base year. It also contains oversamples of military, Hispanic, black, and economically disadvantaged white youth[2]. The NLSY has gathered comprehensive information about the employment, education, training, and family-related experiences of the respondents as they have moved from adolescence into adulthood.

Starting in 1986, the children of the female NLSY respondents have been interviewed at two-year intervals to assess their cognitive ability, socioemotional development, and home environments. The child measures vary with the ages of the children, who range in age from newborn to midteens as of the 1990 interview. Response rates have remained over 90 percent for all waves.

2.2 The Analysis Sample

The analyses incorporate two types of child samples, a pooled sample for investigating outcome transitions over a two-year period and a more restricted sample for examining child outcomes over a four-year span. The pooled sample of 2,010 consists of children at least age five at the time of the 1986 or 1988 assessment and who had valid scores on all three outcome measures in 1988 or 1990. Turning to the longer time span, the restricted sample is limited to 930 children at least age five in 1986 who were interviewed at all three points in time. Since children of the disadvantaged white oversample were not assessed in 1990, they were eliminated from analyses using the restricted four-year sample.

As of 1986, the mothers in the sample were between the ages of 21 and 28, and as of 1990 they were 25 to 32 years of age. Since the children in the sample were 5 and over in 1986, their mothers were relatively young when the children were born - averaging about age 19 at birth for the black mothers and closer to 20 for their white counterparts. The mothers (and their children) do *not* represent a full cross-section of mothers and children, but more appropriately may be viewed as typifying a national sample of relatively younger mothers and their children[3]. However, these mothers are far from being population outliers; they represent very well a cross-section of mothers and their children at a point several years - at least seven at the second outcome point - past their birth.

2.3 The Measures

Three assessments that tap the cognitive and behavioral dimensions were selected from the 1986, 1988, and 1990 NLSY child data: a behavior problems scale completed by all mothers of children age four and over and two achievement tests administered to children age 5 or older. Behavior problems are measured by a 28 item mother-report scale designed to gauge the frequency and type of recent problem behaviors exhibited by the child in the three months prior to the interview. An increase in scores on this index indicates greater behavior problems. The Math and Reading Recognition subtests of the Peabody Individual Achievement Test (PIAT) were used to assess mathematics and oral reading ability. The difference in percentile scores on these tests between the two assessment points was used to measure change in cognitive achievement. Increases on the two PIAT's indicate improvement in cognition. Noncompletion rates for all three outcomes ranged between 10 and 15 percent, depending on the child's age and race/ethnicity (Baker and Mott 1989).

[2] Individual case weights are constructed for each survey year to make the sample conform to independently derived population totals for individuals aged 14 through 21 on January 1, 1979. The weights take into account the probability of selection at the baseline interview, differential nonresponse at the initial baseline phase, and random variation associated with sampling. The weights produce group population estimates when used in tabulations.

[3] Child weights are based on mother weights with an adjustment factor used to account for different interview rates for children in various age, race, and sex groups. These factors use counts of children known to exist as well as estimates of fertility for women who have attrited. Child attrition does not, however, adjust for differential child assessment completion rates.

The NLSY Child dataset contains several measures useful for operationalizing the kinds of socioeconomic and maternal employment factors that are postulated to affect children's outcomes. Table 1 lists these key determinants as well as the antecedents and basepoint factors that are used as controls to account for child characteristics, maternal traits, and other contemporaneous family attributes.

Change in family poverty is measured by averaging poverty levels over the span of time between the assessment points. The variable is a ratio of the family unit's total income to the officially established poverty level, defined as minimum need for a family, based on the family's size and age of household head. The extent of maternal employment is measured by taking the average of total weeks worked in the years intervening between the basepoint and the end point assessment date. For example, reports of weeks worked from the 1987 and 1988 mother interviews were used in creating the average for children assessed first in 1986 and again in 1988. Low work (less than 20 weeks) and moderate work (20 to 39 weeks) were distinguished from high levels of work (40 to 52 weeks). The full time schedule was the reference category in the equations.

Mother's intellectual resources are represented by her AFQT (Armed Forces Qualification Test) score, a measure of developed abilities (Profile of American Youth, 1982). Mother's level of schooling, measured by highest grade completed as of the 1986 interview, is used as an indicator of prior socioeconomic status of the family. Dummy variables indicating whether the mother drank or smoked in the twelve months preceding the child's birth represent a number of what might be considered "mothering" traits. Mother's age in years at the time of each child's birth is used to account for other prior unobservable traits. Child birthweight, measured in ounces, is an indicator of potentially compromised development reported by the mother in the first interview following the child's birth.

Two measures of household composition were included: (1) the number of years in the assessment time span in which a grandparent was present in the child's household and (2) the number of years in this period that the mother's husband or partner was present.

2.4 Model

For all three dependent variable, straightforward models are estimated in which change in child outcome is treated as a function of levels of family income and maternal employment, in the intervening period, controlling for the base point child and maternal characteristics noted earlier (such as birthweight, nonmaternal childcare, education, AFQT, age at birth, prenatal practices, and household composition). Pre-existing family characteristics are incorporated into the analysis to control for possible selection effects that may result if one only considers family circumstances that are contemporaneous to the assessment period.

Selecting the most appropriate method for modeling this change process is problematic. The characteristics of the model do not entirely meet the criteria normally associated with the change score model, in which $Y_2 - Y_1$ is regressed on X, or with the regressor variable model in which Y_2 is regressed on both Y_1 and X (Allison 1990). In an effort to avoid overestimating any possible change effects or underadjusting for prior differences, the change score was used as the dependent variable instead of the regressor variable model. While there may be some causal processes operative between Y_1 and Y_2, there is no compelling reason to assume that the period specific components of Y_1 are correlated with X. Placing intervening changes in the family conditions as temporally subsequent to the basepoint assessment should reduce the problems of measurement error in Y_1. Because so much research suggests that family processes and their linkages with child outcomes can vary in fundamental ways between black and white children, analyses were performed separately, stratified by race.

3. FINDINGS

Mean characteristics for all explanatory and outcome variables used in the analysis appear in Table 1. For both the mathematics and behavior problems scores, there is little net change in average score over the two-year period, although of course small net changes can and do mask some substantial gross flow, both up and down, at the individual level. In contrast, the overall sample, particularly the blacks, showed a substantial decrease in reading recognition scores over the period. Given that these assessment scores have been normed against a U.S.

national sample, this decline implies that the NLSY children lost some ground in their reading skills during these early school years when contrasted with a full national cross-section of U.S. children comparable in age.

Table 1: Mean Statistics for Explanatory and Outcome Variables by Race: Pooled Sample.

	BLACK	NONBLACK
Difference in PIAT Reading Rec. (Percentile), Base to Endpoint	-8.7	-3.1
Difference in PIAT Math Percentile, Base to Endpoint	+0.4	-0.9
Difference in Behavior Problems Percentile, Base to Endpoint	+0.9	+1.9
% Received Nonmaternal Child Care, First 3 Years of Life	0.5	0.5
Birth Weight (Ounces)	108.7	116.6
% of Mothers with Less than 12 Years of School	39.6	46.2
% of Mothers with 12 Years of Schooling	41.7	41.5
% of Mothers Drinking Alcohol During Pregnancy	31.5	41.2
% of Mothers Who Smoked Cigarettes During Pregnancy	32.1	42.1
Mean Age of Mother at Birth of Child	18.8	19.6
% of Mothers with AFQT Score Below 50+ Percentile	91.7	70.2
Mean Survey Points Grandparents in Home, Base-Endpoint	0.5	0.2
Mean Survey Points Mother's Spouse in Home, Base-Endpoint	1.0	2.0
% of Families with Prov. Ratio < 1, Base-Endpoint	48.1	25.1
% of Families with Prov. Ration of 1-1.99, Base Endpoint	25.3	30.8
% of Mothers Worked < 20 Weeks Per Year, Base-Endpoint	44.6	40.6
% of Mothers Worked 20-39 Weeks Per Year, Base-Endpoint	16.0	20.4
SAMPLE SIZE	744	1,266

Given their relatively youthful age, these mothers and their families have characteristics that leave them somewhat disadvantaged when compared with a full cross-section of U.S. women; in particular, they are less likely to have completed high school and to have attended college, and more likely to be living in poverty. The black mothers and their families in the sample are substantially more disadvantaged than their white counterparts.

3.1 Levels versus Changes in Levels of Outcomes

The focus here is on examining changes in child outcomes in relationship to the various explanatory variables, in contrast to most research which has typically examined how *levels* of child outcomes are linked with various family and socioeconomic inputs. Table 2 indicates how these two perspectives can give very different results. The "level" equation uses the end point percentile achievement or behavioral score as the outcome variable. In the "change" equation, the outcome is the *difference* in the percentile scores (end point less base year percentile score).

Since evidence suggests that the association between the extent of maternal employment and child success may be fairly complex (Piotrkowski, Rapoport, and Rapoport 1987), dummy variables were used to measure both inputs as a means to test for lack of linearity in the association between the input(s) and outcomes. The omitted reference groups are categories measuring higher income (at least twice the poverty level) and more maternal employment (at least 40 weeks a year), respectively. Rather than present all of the coefficients from the full equations, the explanatory coefficients of interest are highlighted by presenting only the poverty and maternal employment coefficients, "controlled" for the full range of explanatory variables discussed above.

Turning first to the linkages between poverty status and the child outcomes, for black children, it may be seen that after controlling for the full set of explanatory variables, there is no statistically significant association between poverty status and the levels of scores for any of the outcomes over the two year period. Paralleling this finding, it does *not* appear that either improvements or deterioration in these child outcomes over the two-year period are linked with poverty status. In contrast, for white children, there are powerful associations between poverty and both levels of child cognition and behavior problems. For all three assessments there are statistically and substantively important relationships between being in poverty (or near poverty) and scoring poorly in mathematics or reading or having an above average level of behavior problems. As with the black children, there is no evidence of deterioration in scores associated with poverty status. It appears that for both black and white children, short-term transitions in cognitive or behavioral well-being are linked with factors other than those proxied for by the variables in our equations[4]. These could include ecological factors such as neighborhood characteristics, child peer networks and school characteristics as well as other unobserved maternal or family traits.

Table 2: Comparing Poverty and Maternal Employment Effects on Levels and Changes in Levels of Child (Percentile) Outcomes (O.L.S. Coefficients): Pooled Sample.

	PIAT MATHEMATICS				PIAT READING RECOGNITION				BEHAVIOR PROBLEMS			
	Level		Change		Level		Change		Level		Change	
Black												
Poverty												
In Poverty	-3.6	(2.9)	1.4	(3.2)	-5.6^{c}	(3.3)	-3.4	(2.8)	3.6	(3.1)	4.2	(3.2)
"Near Poverty"	0.2	(2.7)	2.4	(3.0)	-2.7	(3.1)	-1.2	(2.6)	0.1	(2.8)	-3.6	(2.9)
Maternal Employment												
Mom Works < 20 Weeks Yearly	-3.1	(2.3)	2.1	(2.5)	-1.3	(2.6)	4.8^{b}	(2.3)	-2.1	(2.4)	-9.0^{a}	(2.5)
Mom Works 20-39 Weeks Yearly	2.1	(2.6)	7.4^{a}	(2.8)	3.4	(2.9)	7.2^{a}	(2.5)	-3.9	(2.7)	-3.5	(2.8)
Non Black												
Poverty												
In Poverty	-7.3^{a}	(2.4)	0.6	(2.4)	-8.9^{a}	(2.7)	-2.0	(2.3)	8.7^{a}	(2.5)	1.2	(2.2)
"Near Poverty"	-6.3^{a}	(1.7)	-0.1	(1.7)	-7.0^{a}	(1.8)	-1.3	(1.6)	4.8^{a}	(1.8)	-0.8	(1.6)
Maternal Employment												
Mother Works < 20 Weeks Yearly	-2.0	(1.6)	-1.3	(1.6)	-0.6	(1.8)	-1.3	(1.6)	-3.2^{c}	(1.7)	-1.1	(1.5)
Mother Works 20-39 Weeks Yearly	-5.4^{c}	(1.9)	-5.5^{a}	(1.9)	-0.4	(2.1)	3.2^{c}	(1.8)	-4.1^{b}	(2.)	1.3	(1.7)

Note: The poverty and employment variables are controlled for all of the explanatory variables listed in Table 1. The poverty and employment variables are average statistics for the assessment period. Omitted reference categories are: Poverty Ratio ≥ 2.0 and Mother Works 40-52 weeks yearly respectively. The poverty ratio is a ratio of the family unit's total income to the officially established poverty level (minimal family need level for a family, based on the family's size and age of household head). Statistics are from separate black and non-black equations.

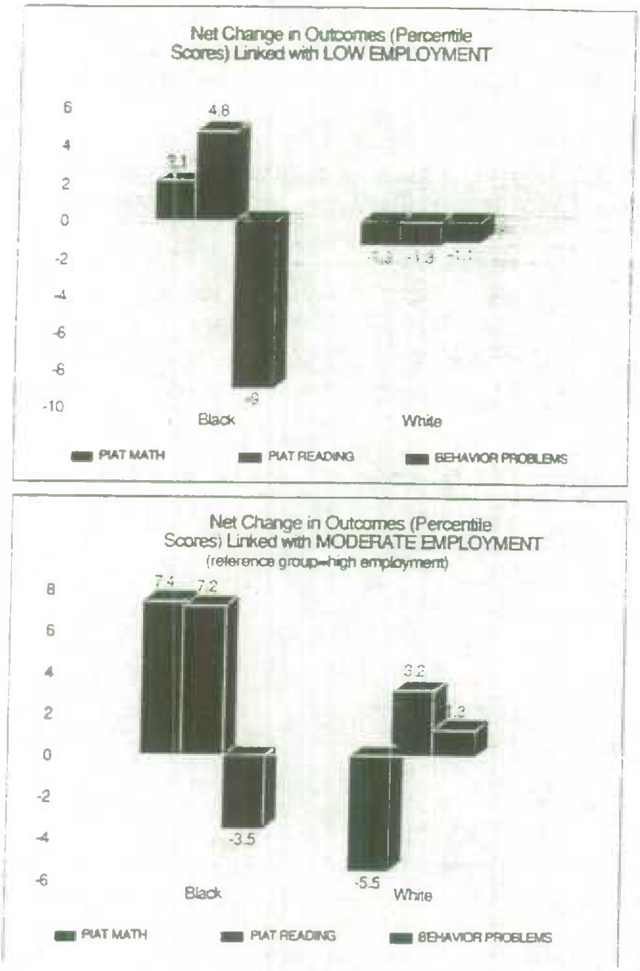
a = coefficient significant at $p \leq .01$; b = $\leq .05$; c = $\leq .10$.

Shifting from poverty to maternal employment, a somewhat different story emerges. For black children, there is no association between the extensiveness of maternal employment and the *level* of a child's well-being. However, as highlighted in the bottom graph of Figure 1, for the two cognitive outcomes, moderate employment (working 20 to 39 weeks yearly) is associated with improvements in cognitive scores over the two-year period. What might be happening here is a "tradeoff" between the value of maternal time spent with a child and maternal quality time. The time a mother spends with a child can be linked with a child's intellectual capability, the more so if the mother brings intellectual capability to the interaction. The skills a mother learns on the job, be they reading or mathematics linked, can enhance her ability to teach her child. The more time she spends with a child, the better she is able to provide this training. For black families, moderate employment may represent an optimum balance in this regard[5]. A contrasting picture appears with respect to the behavior problems outcomes, as illustrated in the top half of Figure 1. It appears that for black children, a greater maternal time commitment in the home (working less than 20 weeks a year) is associated with a substantial improvement (declining score) in behavior problems. Controlling for other related factors (such as education and the presence of other family members), it appears that black children whose mothers do not work extensively are substantially advantaged emotionally compared with black children whose mothers work most of the year.

[4] It should be noted that the equations include two other proximate family variables, measuring the presence of a man in the home and the presence of a grandparent. For black children, these household composition variables did not attain significance in any of the change equations; for white children, the presence of grandparents was associated with higher mathematics scores.

[5] Interacting occupation or educational level with intensity of employment might clarify the separate quality-intensity dimensions.

Figure 1.



None of these employment effects appear for white children. In the mathematics domain, white children are disadvantaged (both level of score and change in score) if their mother works a moderate amount during the year compared with a more intensive maternal work involvement. This may have to do with differences in selection biases implied by white and black women working. Also, whereas moderate employment was associated with improvements in black childrens' reading ability, no such pattern appeared for white children. Additionally, white children are affected only moderately in their behaviors if they have a mother who works less - and this effect shows up as a modest level effect with no linkage between maternal employment intensity and changes in child's behavior.

Separate equations not shown were examined to assess the degree with which the coefficients of the key explanatory variables are affected by the set of control variables. In almost all cases, the magnitude of the coefficients is not altered when the control variables are added to (or subtracted from) the equations. In no instance is the statistical significance of a coefficient materially affected when controls are added. Thus, the changes in cognition or behavior for black children associated with moderate employment appear to be independent of the other factors in the equation (at least in the form specified). Apparently the lack of association between poverty status and changes in cognition or behavior in the controlled equations for both the black and white children do not mask effects which would appear in uncontrolled bivariate associations.

3.2 Age of Child

A substantial amount of research in child development suggests that childrens' ability to acquire particular skills or respond to different stimuli can be closely linked with their stage of psychological or intellectual development--which is closely linked with physiological or calendar age. Since child outcomes are often shaped by age, the equations were run separately by age to gauge the extent to which associations between family status and changes in child outcomes might be sensitive to the maturational level of the child. Results not shown here revealed that just as the overall associations between poverty (or near poverty) status and changes in the child outcomes tended not be significant, similarly age-specific associations also were not in evidence. For both blacks and whites there was virtually no systematic variation by age in the likelihood of poverty status affecting changes in cognition or behavior problems.

In contrast, Table 3 suggests that there may indeed be some important variations by age and race in how maternal employment may affect changes in child outcomes[6]. For black children, having a mother who works very little outside the home is associated with improvements in behavior problems for children at all ages[7]. In

[6] This conjecture is somewhat tenuous since we have not yet tested for statistically significant differences in effects between the age-race categories.

[7] Additional analysis is needed to clarify the direction of causality between these factors. Mothers of children who have significant behavior problems might well be less likely to work (or more likely to reduce their work) than other mothers.

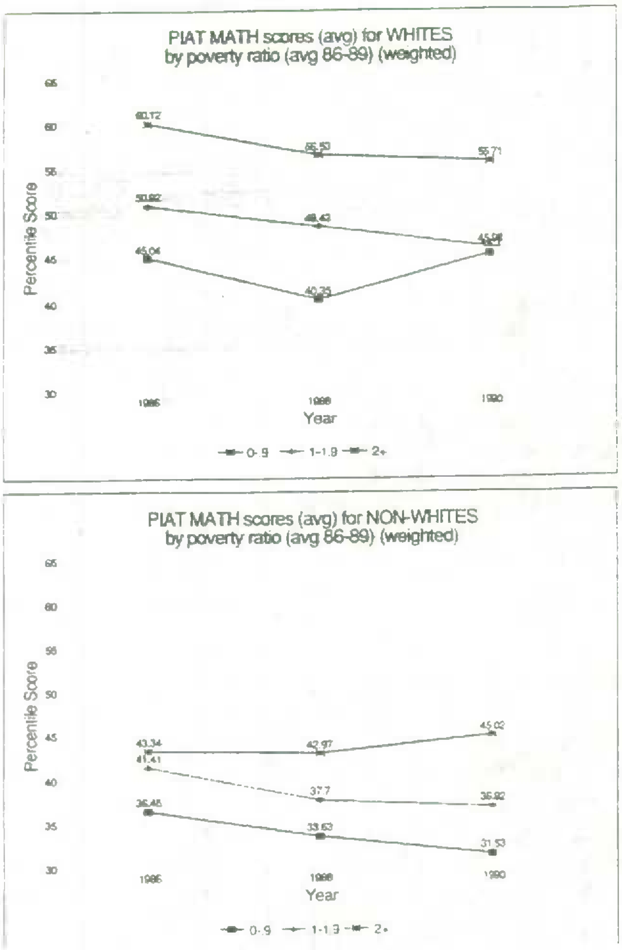
addition, for all except the oldest black children, mother staying home is associated with more reading improvement. In contrast, the overall association between moderate (20-39 weeks) maternal employment and

Table 3: Maternal Employment Linkages with Changes in Cognition and Behavior Problem Scores By Race and Age (O.L.S. Coefficients): Pooled Sample.

	MOTHER WORKS LESS THAN 20 WEEKS PER YEAR						MOTHER WORKS 20-39 WEEKS PER YEAR					
	PIAT Math.		PIAT Reading Rec.		Behavior Problems		PIAT Math.		PIAT Reading Rec.		Behavior Problems	
Black												
7-8 Years	6.2	(4.1)	7.9^{b}	(3.8)	-8.8^{b}	(3.7)	11.5^{a}	(4.5)	8.5^{b}	(4.1)	-2.7	(4.1)
9-10 Years	-0.3	(4.3)	7.2^{c}	(3.7)	-11.6^{b}	(5.4)	5.9	(5.2)	12.1^{a}	(4.5)	-6.4	(6.6)
11 and Over	-5.5	(5.0)	2.4	(4.2)	-12.0^{b}	(5.0)	-1.8	(5.1)	4.5	(4.3)	-6.4	(5.1)
Non Black												
7-8 Years	0.5	(2.3)	-2.2	(2.4)	1.6	(2.2)	-3.9	(2.6)	3.1	(2.7)	2.4	(2.5)
9-10 Years	-6.9^{b}	(3.1)	-2.1	(3.0)	-5.6^{c}	(3.0)	-5.6	(3.6)	4.8	(3.5)	-4.6	(3.5)
11 and Over	0.8	(3.4)	1.2	(2.3)	-4.2	(2.9)	-8.7	(4.1)^{b}	3.5	(2.8)	6.5^{c}	(3.5)

Note: The poverty and employment variables are controlled for all of the explanatory variables listed in Table 1. The poverty and employment variables are average statistics for the assessment period. Omitted reference categories are: Poverty Ratio ≥ 2.0 and Mother Works 40-52 weeks yearly respectively. The poverty ratio is a ratio of the family unit's total income to the officially established poverty level (minimal family need level for a family, based on the family's size and age of household head). Statistics are from separate black and non-black equations.
a = coefficient significant at p ≤ .01; b = ≤ .05; c = ≤ .10.

Figure 2.



improvement in mathematics noted earlier for black children largely reflects a very large positive coefficient for the youngest black children. This pattern is consistent with the supposition that maternal nurturance may have its greatest effect for younger children. Shifting to the white children, it may be seen that the modest negative intermediate level effects of employment noted for the overall sample are more in evidence for older children.

3.3 A Four-Year Perspective

It is important to emphasize the short-term time constraints of the analysis thus far. It might be that child well-being could be closely associated with the duration of time a family or child experiences particular levels of poverty or maternal employment. Figure 2 illustrates this possibility by comparing four-year timeliness in mathematics scores for children in different poverty situations. If one follows the black children from 1986 to 88, there is some modest widening in mathematics percentile scores by whether or not the children are living in or near poverty. However, this gap by povertystatus widens substantially between 1988 and 1990. Thus, it may be that the amount of time spent in poverty or the intensity of that poverty existence increasingly erodes achievement, and apparently is not ameliorated by external forces such as compensatory education. In contrast, for white children (who clearly start with higher levels of mathematics knowledge) a different picture emerges. The children in poverty show an initial decline (between 1986 and 1988) but then a substantial recovery, perhaps reflecting advantages being gained from better schools or better home environments.

To better understand the patterns suggested by the descriptive long-term trajectories, a more restricted sample of children assessed in 1986, 1988, and 1990 was used to help identify the processes at work that may be sensitive to the length of time a child spends in a particular status. Turning to the longer time period in a multivariate context, one sees from Table 4 that some new patterns emerge not seen in the two year equations. Black children in poverty and near poverty exhibit significant decreases in both reading recognition and math achievement relative to children in more economically affluent families. The latter finding is consistent with the pattern described in Figure 2. While moderate gains in math scores are evident for black children of mothers working low and moderate amounts, maternal employment now appears to have little effect on reading scores. The improvement in behavior for black children of mothers who work few weeks during the year, already seen in the two-year equations, appears to be increased when a four-year span is considered. A striking effect not seen in the two-year equations is the negative effect of poverty on behavior problems for black children. This result may suggest that duration of time in poverty ultimately takes its toll on black children, as shown by the dramatic increase in behavior problems associated with poverty for black children in the four-year equation. The change processes already seen for white children in the two-year equations appear relatively unchanged in the longer time line.

Table 4: Employment and Poverty Linkages with Changes and Behavior Problem (Percentile) Scores by Race: Children Assessed in 1986, 1988, 1990 (O.L.S. Coefficients).

	PIAT MATHEMATICS		PIAT READING RECOGNITION		BEHAVIOR PROBLEMS	
Black						
Poverty						
In Poverty	-11.5^{a}	(4.6)	-11.7^{a}	(4.5)	12.3^{a}	(4.7)
"Near" Poverty	-6.7^{c}	(3.9)	-7.2^{c}	(3.8)	.2	(4.0)
Maternal Employment						
Mom Works < 20 Weeks Yearly	7.9^{b}	(3.8)	.88	(3.7)	-11.7^{a}	(3.8)
Mom Works 20-39 Weeks Yearly	6.8^{c}	(3.9)	-3.2	(3.8)	-1.5	(4.0)
Non Black						
Poverty						
In Poverty	8.0^{c}	(4.4)	.2	(4.2)	-3.8	(4.0)
"Near" Poverty	.4	(2.8)	-4.5^{c}	(2.7)	-2.3	(2.9)
Maternal Employment						
Mom Works < 20 Weeks Yearly	-4.9^{c}	(3.1)	-4.4^{c}	(3.0)	4.5^{c}	(2.9)
Mom Works 20-39 Weeks Yearly	-5.5^{c}	(3.1)	-.9	(2.9)	3.3	(2.8)

Note: The poverty and employment variables are controlled for all of the explanatory variables listed in Table 1. The poverty and employment variables are average statistics for the assessment period. Omitted reference categories are: Poverty Ratio ≥ 2.0 and Mother Works 40-52 weeks yearly respectively. The poverty ratio is a ratio of the family unit's total income to the officially established poverty level (minimal family need level for a family, based on the family's size and age of household head). Statistics are from separate black and non-black equations.

a = coefficient significant at $p \leq .01$; b = $\leq .05$; c = $\leq .10$.

3.4 Home Environment

The NLSY dataset includes several items and psychometric scales linked with the children which may help interpret some of the poverty and employment results presented here. The NLSY HOME scale is an abbreviated and modified version of a widely used scale designed to measure the nature of mother-child interactions and the quality of home environments (Baker and Mott 1989). The HOME taps a cognitive dimension (including language stimulation, variety of experiences, encouragement of child achievement) and an emotional support dimension (including responsiveness, warmth, encouragement of maturity). As shown in Table 5, how well children score on this assessment is very sensitive to a family's poverty status. Higher scores on the overall scale as well as the cognitive stimulation and emotional support subscales are closely linked with family income for both black and white children. Although the association is less pronounced, there are also positive associations between higher HOME scores and more extensive maternal employment. This latter association is particularly pronounced for black children. Apparently, factors within the black home which tend to be linked with greater acquisition of cognitive skills are also associated with more maternal employment[1].

Table 5: Average HOME Percentile Scores by Poverty Status and by Extent of Maternal Employment: Children Assessed in 1986, 1988, 1990.

	Total HOME Score	Sample Size	Cognitive Stimulation Subscore	Sample Size	Emotional Support Subscore	Sample Size
Black	36.9	404	43.1	393	37.0	350
Poverty Status						
0 - .9 Pov Ratio	29.2	176	33.4	169	34.0	146
1 - 1.9 Pov Ratio	36.9	109	41.4	107	40.5	95
2 + Pov Ratio	49.8	85	59.1	185	39.5	77
Maternal Employment						
< 20 Weeks	33.1	161	38.7	154	33.3	139
20 -39 Weeks	37.6	79	40.8	79	38.3	70
40 - 52 Weeks	40.3	164	48.4	160	40.0	141
Non Black	57.4	492	55.7	484	57.2	463
Poverty Status						
0 - .9 Pov Ratio	34.6	79	35.7	79	39.1	75
1 - 1.9 Pov Ratio	50.4	147	49.1	144	53.2	137
2 + Pov Ratio	67.0	232	63.8	228	64.7	220
Maternal Employment						
< 20 Weeks	52.5	177	51.0	174	54.9	172
20 -39 Weeks	56.6	115	52.0	112	59.1	106
40 - 52 Weeks	61.8	200	61.3	198	58.1	185

As a preliminary step in considering the independent importance of variations in the child's home environment, the overall HOME score was entered into the equations as an additional explanatory variable (results not shown). For black children, the HOME very slightly moderates the effect of poverty status on changes in reading scores and causes a negligible strengthening of the effects on math scores. There is virtually no change in the poverty and employment coefficients for white children when the HOME is added to the model. Overall the variable has little effect on the magnitude of the employment variables. It plays a significant role in predicting absolute outcome levels but *not* in change in outcomes.

4. CONCLUSIONS

From a substantive perspective, these findings suggest that the relationships between family conditions and child outcomes can be sensitive to a variety of factors, including race and, to some extent, the time span being considered. When one moves from a two-year to a four-year time span, certain patterns are accentuated, particularly for black children. Over a four-year period, many children of the ages considered here pass through more than one maturational stage - a development process that could well be linked with how they respond to stimuli from both within and outside the family. A lengthening time span also implies a greater probability that a child may be living in varying environments. To some extent some of the more "domestic" environmental changes have been incorporated into the analyses, but with an increasing time span the problem of "unobservables" in the outside environment becomes more significant. Characteristics of neighborhoods, schools, and peer groups could potentially affect the associations between child outcomes and the inputs considered here. Such influences can be particularly strong for black children, whose family poverty status may more closely mirror contextual community factors than is true for white children.

When viewed over a four-year period, the inputs central to this study, poverty status and, to a lesser degree, maternal employment, seem to exhibit more variability in the longer run. The influence of such changes can be further explored by examining the temporal ordering of poverty and employment status within the period being considered. One might also pursue how the association between the ordering of these inputs and the outcomes may be moderated by the child's maturational status or race. While this analysis has treated family socioeconomic success and maternal employment as equivalent in terms of their meaning as status variables, there are perhaps important differences between them which may affect the results presented here.

Despite many unanswered questions, this analysis has helped clarify some important issues; black children do respond differently in several important respects to within-family status and behaviors. The importance of a mother's employment behavior on how a younger child develops should not be discounted; more importantly, effects are not uniform across outcomes. It is still unclear how the benefits associated with low and moderate levels of maternal employment are linked with the quality of the mother's employment. Most surprisingly for black children, is the pattern of weak direct associations between poverty *per se* and both short-term levels or changes in levels of cognition or behavior problems. This pattern is significantly altered when one shifts from a two to a four-year time line. While the patterns of employment effects on child outcomes appear generally consistent regardless of the time span considered, an exclusively short-term approach to the data may underestimate the effects of other family conditions on child outcomes that may only be revealed by taking a longer view.

REFERENCES

Allison, P.D. (1991). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20, 93-114.

Baker, P.C., and Mott, F.L. (1989). NLSY child handbook 1989: A guide and resource document for the national longitudinal study of youth 1986 child data. Columbus: Center for Human Resource Research, Ohio State University.

Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery. (1982). Washington, D.C.: Office of the Assistant Secretary of Defense.

Parcel, T.L., and Menaghan, E.G. (1990). Maternal working conditions and child verbal facility: Studying the intergenerational transmission of inequality from mothers to young children. *Social Psychology Quarterly*, 53, 132-147.

Piotrkowski, C.S., Rapoport, R.N., and Rapoport, R. (1987). Families and work. In M.B. Sussman and S.K. Steinmetz, Eds. *Handbook of Marriage and the Family*, 251-283. New York: Plenum.

[1] This scale includes some items which have a higher face validity than others in being predictive of mathematical and verbal skills. We plan to examine the extent to which these items, which have considerable cognitive content, help explain the relationships between maternal employment and, in particular, mathematics skill acquisition for black children.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

USING LMAS TO ESTIMATE THE WAGE DIFFERENTIAL BETWEEN LARGE AND SMALL FIRMS IN CANADA

R. Morissette[1]

ABSTRACT

Even after controlling for workers' observable characteristics and unobserved constant-over-time abilities of male job changers, a substantial wage differential remains between large and small firms. What underlies the remaining wage differential is unclear. The wage differential observed in a first-difference wage equation may be subject to important self-selection problems; it could merely reflect differences in sector-specific (*i.e.* firm size-specific) productive abilities. Alternatively, as suggested by efficiency wage models, large firms may pay higher wages to increase worker effort or to reduce turnover.

KEY WORDS: Wages; Firm size; Earnings; Labour market; Employment.

1. INTRODUCTION

Human capital theory and the theory of compensating differentials assert that wages are determined solely by workers' human capital and by non pecuniary aspects of the jobs. Once one controls for these factors, wage differentials should disappear. Recent work by Krueger and Summers (1988) on interindustry wage differentials shows that this is not the case; differences in wages across industries persist even after controlling for these factors. As well as industry structure, employer size seems to affect wages. Recent U.S. studies (Brown and Medoff 1989; Idson and Feaster 1990) suggest that larger employers tend to pay higher wages. The main purpose of this paper is to assess whether such a wage-size relationship also holds for Canada.

Using data from the 1986 Labour Market Activity Survey (LMAS), we find that larger firms generally pay higher wages for observationally equivalent workers. This suggests that part of the difference in wages among Canadian workers depends on factors unrelated to workers' attributes.

The paper is organized as follows. Section 2 presents the theoretical model. Section 3 shows that even after controlling for observable workers' characteristics, as well as for occupation and industry-specific effects, large firms still pay approximately 20% more than small firms. Part of this wage differential may be due to differences in workers' unobserved abilities. We use longitudinal data to control for differences in unobserved constant-over-time abilities and we still get a substantial wage differential for male job changers. This highlights the need for alternative explanations of the wage determination process. These are examined briefly in section 4. Concluding comments follow section 4.

2. THE MODEL

Wages are assumed to depend on workers' human capital, on non pecuniary aspects of jobs, on firm size and on other factors:

$$W_{it} = W(HC_{it}, CD_{it}, SIZE_{it}, OTHER_{it}, u_{it}) \qquad (1)$$

where W_{it} is the wage of worker i at period t, HC_{it} is a vector of worker i's observable characteristics which enhance his productivity, CD_{it} is a vector of jobs characteristics which affect worker i's level of

[1] R. Morissette, Business and Labour Market Analysis Group, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

utility, $SIZE_{it}$ is a vector of variables measuring firm size, $OTHER_{it}$ is a vector of other factors potentially affecting wages (*e.g.* union status, marital status, race) and u_{it} is a random disturbance. Human capital theory and the theory of compensating differentials assert that $SIZE_{it}$ has no effect on W_{it}. Hence, equation (1) is a specification in which a model based on human capital theory and on compensating differentials is nested within a model where firm size matters. Firm size may matter because of heterogeneity in employers' characteristics such as power in the product market, training costs and monitoring costs.

3. EMPIRICAL RESULTS

3.1 The Data

In this paper, the firm is defined as the set of all establishments owned in Canada by a given employer. Small firms are defined as those having less than 20 employees, medium-sized firms are those which employ between 20 and 499 workers and large firms have 500 or more workers. The sample used in this paper includes all full-time jobs held in 1986 by paid workers of the commercial sector[2]. The data is taken from the 1986 Labour Market Activity Survey (LMAS)[3].

At the aggregate level, large firms pay approximately 50% higher wages than small firms (Table 1). This obviously raises the following question: what underlies this wage differential?

Table 1: Average hourly wages by firm size, full-time jobs.

Firm size (number of employees)				
(1) 1-19	(2) 20-99	(3) 100-499	(4) 500+	(5) [(4) - (1) / (1)]
8.91	10.66	11.98	13.50	0.52

Source: Labour Market Activity Survey (1986).

3.2 Controlling for workers' observable characteristics

Human capital theory suggests that individuals who have more education, more general and firm-specific skills receive higher wages because they are more productive. The theory of compensating differentials suggests that firms have to offer higher wages to attract workers of a given quality in jobs which involve poor working conditions. To test these arguments, we specify equation (1) as follows:

$$ln\ W_{it} = B_0 + B_1*HC_{it} + B_2*CD_{it} + B_3*SIZE_{it} + B_4*OTHER_{it} + u_{it} \qquad (2)$$

[2] In this paper, the commercial sector includes all industries except : 1) agriculture, 2) fishing and trapping, 3) education and related services, 4) health and welfare services, 5) religious organizations, 6) federal administration, 7) provincial administration, 8) local administration and, 9) other government offices. Since it is restricted to paid workers, the resulting sample excludes unpaid family workers and self-employed workers. In 1986, hours worked in full-time jobs in the commercial sector were distributed as follows : 25 % in small firms, 35 % in medium-sized firms and 40 % in large firms.

[3] Morissette (1992) compares the employment data from LMAS to that from the Survey of Employment, Payrolls and Hours (SEPH). He shows that the distribution of employment by establishment size resulting from LMAS substantially underestimates the establishment size as derived from SEPH and thus, cannot be used to assess whether large establishments pay higher wages than small establishments. However, the distributions of employment by firm size derived from the two surveys are similar. This suggests that LMAS can be used to investigate whether large firms pay higher wages than small firms.

where: HC_{it} includes: education, age, age squared, tenure, tenure squared;
CD_{it} includes: 2-digit industries, 2-digit occupations;
$SIZE_{it}$ includes: firm size;
$OTHER_{it}$ includes: marital status, visible minority status, union status, central metropolitan area, region.

As in Mincer's (1974) human capital earnings function, equation (2) contains variables reflecting workers's various levels of schooling (*i.e.* education) and labour market experience (proxied by age). Job tenure is added to capture differences in workers' experience relevant to the present job[4]. To take into account differences in non pecuniary aspects of the jobs, one would ideally like to have an index of quality of working conditions which could be included on the right-hand side of equation (2). No such information is available on LMAS. Following Brown and Medoff (1989), we add controls for industry and occupation; this may help capture part of the variation in working conditions which occurs across industries and occupations. Because they may have a different quit behaviour or because they may face discrimination, unmarried workers and workers from visible minorities may receive differing wages; two intercept shift terms are included to take this into account. Union status is included to capture the impact of unionization on wages. Four regional dummy variables and one census metropolitan area dummy variable are included to allow for the possibility of having distinct local labour markets as a result of workers' imperfect geographical mobility.

Table 2 presents the wage differential between small and larger firms derived from equation (2). Results are shown separately for men and women. In both cases, large firms pay higher wages than small firms. The wage differential between large and small firms varies from 20% to 24%[5].

Table 2: Wage differential between small and larger firms, full-time jobs[1].

	Men	Women
Firm size		
FIRM2[3]	0.0815	0.1007
	(0.0093)[2]	(0.0121)
FIRM3	0.1450	0.1650
	(0.0105)	(0.0137)
FIRM4	0.1853	0.2166
	(0.0095)	(0.0119)
Adj. R^2	0.4256	0.4388
Sample size	15,506	8,382

[1] The dependent variable is the logarithm of the hourly wage rate. Regressions are run using weighted least squares; each job is weighted by the number of working hours.

[2] Standard errors of size coefficients are between parentheses.

[3] FIRM2 (FIRM4) refers to firms with 20-99 (500 or more) employees. Firms with 1-19 employees are the reference group. The percentage wage differential between small and larger firms is equal to the antilog of the regression coefficient minus 1, and expressed as a percentage.

* not significant at the 5% level.

Source: Labour Market Activity Survey (1986).

[4] Tenure squared and age squared are included to allow non-linearity in the age/wage or tenure/wage relationship.

[5] The percentage wage differential is the antilog of the regression coefficient minus 1, and expressed as a percentage. Thus, in full-time jobs held by men, the 20% wage differential between large and small firms results from : exp(0.1853) - 1.0.

3.3 Controlling for workers' unobserved constant-over-time abilities

As is commonplace in studies looking at the effect of unionization (Freeman 1984), industry (Krueger and Summers 1988) or firm size (Evans and Leighton 1989, Brown and Medoff 1989) on wages, one may argue that part of the variation in wages is due to the fact that workers have differing unobserved abilities. More precisely, if workers in large firms have more of these unobserved abilities, then it is possible that the wage differential found so far merely reflects an "unobservable workers' quality differential". First-differencing equation (2) allows us to take into account the portion of these unobserved abilities that is constant over time. To see this, consider the following wage equation:

$$ln\ W_{it} = B_1 * X_{it} + B_2 * a_i + u_{it} \qquad (3)$$

where $ln\ W_{it}$, the logarithm of the wage of worker i at time t, depends on a vector X_{it} of observable variables, on unobserved constant-over-time abilities a_i and on a random term u_{it}. First-differencing the above equation leads to the following equation:

$$ln\ W_{it} - ln\ W_{it-1} = B_1 * (X_{it} - X_{it-1}) + (u_{it} - u_{it-1}) \qquad (4)$$

in which unobserved constant-over-time abilities no longer appear and are thus implicitly taken into account. Equation (4) has been estimated in the United States for the period 1973-1977 (Brown and Medoff 1989) using the Quality of Employment Survey and for the period 1976-1981 (Evans and Leighton 1979) using the National Longitudinal Survey of Young Men. While Brown and Medoff (1989) find that the size effect remains substantial even after controlling for differences in unobserved abilities, Evans and Leighton (1979) conclude that "about 60 percent of the wage-size effect is due to unobserved heterogeneity when all firms are considered and about 100 percent when firms with 25 or more employees are considered" (p. 299).

As mentioned above, we rely on the 1986 version of the LMAS file. This file groups information on up to five jobs held by a given individual in 1986. We concentrate on the first and second job held that year by all job changers[6]. This leads to a total of 1,539 and 897 observations on wage differences for male and female workers, respectively.

We estimate the first-difference version of equation (2)[7]. The dependent variable is the difference between the (natural logarithm of the) hourly wage rate in the second job held in 1986 and that in the first job held in 1986. We also add a dummy variable to distinguish job changers who stay in the same 2-digit occupation from those who change occupations when going from their first to their second job. Because workers belonging to the first group are more likely to carry to their second job a substantial portion of the knowledge acquired in the previous job, they are expected to experience higher net wage increases than other workers. The third column of Table 3 presents the wage differentials resulting from this first-difference wage equation. The first two columns show size estimates of equation (2) run on: 1) on all first jobs held by job changers and 2) on all second jobs held by job changers.

The first two columns of Table 3 show that the wage differential between large and small firms (FIRM4) varies between 8% and 15% for male job changers. If this differential were simply due to differences in workers' unobserved abilities, then it should vanish when using the first-difference wage equation. Clearly, this is not the case; the wage differential resulting from the first-difference wage equation equals 9% (column 3). One may argue that workers who quit their jobs should be expected to receive higher net wage increases than those who are laid-off. If so, the reason for changing jobs should be included as a regressor. Adding two dummy variables for quits and layoffs to the first-difference wage equation (column 4) does not alter the size coefficients

[6] The number of individuals holding more than two jobs within the same year is too small for meaningful statistical analysis.

[7] Variables that take constant values within a year (education, age, sex, visible minority status) disappear when we go from equation (2) to its first-difference version.

substantially; the wage differential between large and small firms remains at 9%[8]. Moreover, although the wage-firm size effect disappears (when going from equation (2) to its first-difference version) for firms with between 100 and 499 employees (FIRM3), it remains significant and fairly constant for firms with between 20 and 99 employees (FIRM2). Thus, it seems fair to conclude that unobserved abilities cannot explain the whole wage differential observed for male job changers.

Table 3: Wage differential between small and larger firms for job changers, full-time jobs.

	Wage equation[1]: first job held in 1986	Wage equation[1]: second job held in 1986	First-difference wage equation[2]	
			without	with
			dummies for quits and layoffs	
	(1)	(2)	(3)	(4)
Men				
FIRM2	0.0790	0.0546	0.0629	0.0604
	(0.0263)	(0.0261)	(0.0237)	(0.0234)
FIRM3	0.1315	0.1146	0.0337*	0.0237*
	(0.0345)	(0.0310)	(0.0313)	(0.0310)
FIRM4	0.0803	0.1423	0.0895	0.0866
	(0.0286)	(0.0295)	(0.0287)	(0.0284)
Adj. R^2	0.4771	0.5063	0.1799	0.1974
Sample size	1,539	1,539	1,539	1,539
Women				
FIRM2	0.1107	0.0326*	0.0113*	0.0093*
	(0.0423)	(0.0325)	(0.0372)	(0.0370)
FIRM3	0.1107	0.2478	0.1405	0.1420
	(0.0481)	(0.0361)	(0.0430)	(0.0428)
FIRM4	0.0365*	0.1235	0.0879	0.0928
	(0.0422)	(0.0297)	(0.0377)	(0.0376)
Adj. R^2	0.4325	0.5640	0.1085	0.1189
Sample size	897	897	897	897

[1] The dependent variable is the logarithm of the hourly wage rate. Regressions are run using weighted least squares. Each job is weighted by the number of working hours.

[2] The dependent variable is the first-difference of the logarithm of the hourly wage rate. Regressions are run using weighted least squares. Each observation is weighted by its sample weight. For other details, see Table 2.

Source: Labour Market Activity Survey (1986).

Results for female job changers are somewhat puzzling. While the wage-firm size effect remains substantial for firms with between 100 and 499 employees (FIRM3), it disappears for firms with between 20 and 99 employees

[8] The dummy variable for quits is significant at the 5% level and implies that, compared to the reference group [i.e. workers leaving their job for other reasons (illness, personal or family responsibilities, bad weather, labour dispute, unpaid vacation, seasonal nature of job, sale of the business or farm, other)], male workers who quit receive 10% higher net wage increases. The dummy variable for layoffs is significant at the 6% level and implies that male workers who are laid-off receive 5 % lower net wage increases than those leaving their job for other reasons.

(FIRM2). Moreover, the size coefficient for large firms (FIRM4) is not significant for the first job held by female workers in 1986. We have no simple explanation to offer for this pattern.

Despite this, the evidence presented in Table 3 suggests that, at least for male job changers, a sizeable wage differential remains between large and small firms, whether or not we control for workers' unobserved abilities. This implies that observationally equivalent workers may receive differing wages depending on the size of the firm they work for. Then why would larger firms pay higher wages?

4. WHY WOULD LARGER FIRMS PAY HIGHER WAGES?

Economics offers many explanations as to why larger firms would pay higher wages. Previous sections have dealt with some of these explanations. Following Brown and Medoff (1989), one can argue that larger firms would pay higher wages because:

1) they have a higher quality workforce;
2) they must compensate workers for bad working conditions;
3) they want to avoid unionization;
4) they face fewer applicants per job and have to raise wages to attract a given quality of applicants (Weiss and Landau 1984).
5) they have more market power (*i.e.* more inelastic demand curves) and share part of their excess profits with workers;

While the labour-quality argument assumes that firms of different sizes pay workers with identical characteristics an identical wage, the four other hypotheses imply that identical workers can be paid differing wages. Efficiency wage models can also be used to explain why employers would pay identical workers differing wages (Yellen 1984). As applied to the wage-firm size relationship, they could be used to argue that larger firms would pay higher wages because:

6) they have more difficulty than small employers detecting shirking and use higher wages as a worker discipline device (Shapiro and Stiglitz 1984);
7) they have higher training costs and use higher wages as a way to reduce turnover (Salop 1979);
8) they rely more on teamwork than small employers and want to raise the work norms of their workers above the minimum required by paying them wages in excess of the minimum required (Akerlof 1982)[9].

5. CONCLUSIONS

The evidence presented in this paper suggests that, at least for male job changers, differences in workers' observable characteristics or in unobserved constant-over-time abilities cannot fully explain why large firms pay higher wages than small firms. Why this is so is still unclear. The wage differential observed in the first-difference wage equation may be subject to potentially important selection problems. Following Heckman and Sedlacek (1985), one may argue that it merely reflects differences in sector-specific (*i.e.* firm size specific) abilities. Workers moving voluntarily from small to large firms would have productive abilities whose unit price would be **much** higher in large than in small firms whereas workers moving voluntarily from large to small firms would have skills whose unit price would be **slightly** higher in small than in large firms. As was done by Krueger and Summers (1988) in the case of interindustry wage differentials, one may take this argument into account by: a) checking whether those who move from small to large firms experience a wage increase similar to the wage decrease (supposedly) experienced by those who move from large to small firms, b) looking at the wage changes of displaced workers, *i.e.* workers changing jobs involuntarily after being laid-off.

[9] Another version of efficiency wage models (adverse selection models : see Weiss (1980)) suggests that firms cannot infer workers' ability (which is assumed to be unobservable) and have to pay higher wages to attract a better pool of applicants. As applied to the wage-size relationship, these models would imply that larger firms pay higher wages because they want to have high-ability workers. Since a substantial wage differential remains even after controlling for unobservable constant-over-time abilities as well as for observable characteristics of male job changers, these models cannot be used to explain the remaining wage differential.

Alternatively, the remaining wage differential may be due to firms' heterogeneity. As suggested by efficiency wage models, if firms of different sizes differ in ease of monitoring workers, in training costs or in their reliance on teamwork, they may find it profitable to pay differing wages to identical workers. Because different explanations of the wage differential may lead to different policy implications, determining the source of the wage-firm size effect is an important question for future research.

The author wishes to thank Ted Wannell for his aid and valuable comments, as well as Garnett Picot and John Baldwin for helpful comments on earlier drafts of this paper.

REFERENCES

Akerlof, G.A. (1982). Labor contracts as a partial gift exchange. *Quarterly Journal of Economics*, 97, 543-569.

Brown, C., and Medoff, J. (1989). The employer size-wage effect. *Journal of Political Economy*, 97, 1027-1059.

Evans D.S., and Leighton, L.S. (1989). Why do smaller firms pay less? *Journal of Human Resources*, 24, 299-318.

Freeman, R. (1984). Longitudinal analyses of the effects of trade union. *Journal of Labor Economics*, 2, 1-26.

Heckman, J., and Sedlacek, G. (1985). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of Political Economy*, 93, 6, 1077-1125.

Idson, T.L., and Feaster, D.J. (1990). A selectivity model of employer-size wage differentials. *Journal of Labor Economics*, 8, 99-122.

Krueger, A.B., and Summers, L.H. (1988). Efficiency wages and the inter-industry wage structure. *Econometrica*, 56, 259-293.

Mincer, J. (1974). Schooling, experience, and earnings. (New York: Columbia University Press).

Morissette, R. (1992). Canadian jobs and firm size: do smaller firms pay less? *Canadian Journal of Economics*, forthcoming.

Salop, S.C. (1979). A model of the natural rate of unemployment. *American Economic Review*, 69, 117-125.

Shapiro, C., and Stiglitz, J.E. (1984). Equilibrium unemployment as a worker discipline device. *American Economic Review*, 74, 433-444.

Weiss, A. (1980). Job queues and layoffs in labor markets with flexible wages. *Journal of Political Economy*, 88, 526-538.

Weiss, A., and Landau, H. (1984). Wages, hiring standards, and firm size. *Journal of Labor Economics*, 2, 477-499.

Yellen, J.L. (1984). Efficiency wage models of unemployment. *American Economics Association Papers and Proceedings*, 74, 200-205.

SESSION 9

Data Analysis II

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

BUILDING UP AN INTERNATIONALCOMPARATIVE PANEL DATABASE: THE PACO - PROJECT

G. Schaber, G. Schmaus and G.G. Wagner [1]

ABSTRACT

International research with data sets from national panels is difficult because each of the national dataset is organized in a different manner and uses a different format. In order to overcome these problems, CEPS/INSTEAD is creating in Luxembourg - in partnership with DIW Berlin - an international comparative database with various national household panels. The PACO Project's aim is to develop instruments for analyzing, programming and simulating socio-economic policies. It is intended to facilitate comparative cross-national research on policy issues such as labour force participation, income distribution, poverty, problems of the elderly and so on.

KEY WORDS: Harmonized variables; Relational structure.

1. INTRODUCTION

For capturing and analyzing phenomena in the various fields of, for example, labour market, income distribution, poverty or social protection and also for developing blueprints for shaping public policy, the social scientists and economists of the eighties and nineties make heavy use of micro-data.

The micro-data being used in most of their studies, however, stem from cross-sectional surveys. LIS (see Smeeding, T.M.,and Schmaus, G. 1988; Smeeding, T.M.,and Schmaus, G. 1990) makes a strong attempt to bring together cross-sectional files with income information. For the purposes mentioned above, cross-sectional data are obvious superior to aggregates; nevertheless for dealing with economic or social PROCESSES and the underlying DYNAMICS, they fare rather poorly.

This is the reason why some research teams - first in the US, then in Europe - started to set up panel samples and collect longitudinal micro-data that allow detailed analyses OVERTIME - time being essential in any attempt to come to grips with changes, processes and dynamics.

Panel analyses put a heavy demand on researchers, who will have to spend a large amount of time becoming familiar with the panel's data organization and with the procedures for its exploitation.

One single panel could be sufficient and most of the panel analysts do in fact do their work on one given panel, which in all known instances is that of their own particular country.

So up to now, as a rule, only single country data and problems have been treated. Little is known about differences and similarities between countries.

Social scientists will have to turn more regularly to cross-national and truly comparative use of panel data, in order to learn more about the various national systems which organize taxation or social security and protection

[1] Prof. Dr. Dr.h.c. G.Schaber, President of CEPS/INSTEAD; University of Liege, Belgium; Clark University, Massachusetts, U.S.A. G. Schmaus, Senior Researcher at CEPS/INSTEAD; Luxembourg. Dr. G.G. Wagner, German Institute for Economic Research (DIW,Berlin); since Sept. 1992 Professor at Ruhr-University Bochum, Germany.

or the labour market domestically, about the way they operate and how they affect the various groups and categories of people within the total population.

Some research teams have already joined together in order to proceed in common to clearly defined international comparisons based on data from various national panels, the data being treated according to commonly accepted standards. In one case, a working group was set up involving researchers from Canada, France, Germany, Ireland, Luxembourg, the Netherlands, Sweden and the United States, the co-ordinator being Greg Duncan (see Duncan, G.J. *et al.* 1991) from the PSID.

In a second case, working groups have been set up within the framework of the ESF Network on Household Panels, on topics like labour market, mobility and poverty. But every corresponding set of data has been treated separately.

More demanding are the following approaches: a team directed by Richard Burkhauser and Tim Smeeding at the University of Syracuse in USA is now starting to bring together the German (see Burkhauser, R.V.) and the American PSID panel data set.

With PACO we will try to bring together several European panels and the US PSID in Luxembourg.

2. THE PROBLEM WITH PANEL DATA

Cross-national research with data sets from national panels is difficult because each of the national datasets is organized in a different manner and uses a different format. The situation is that, in summary, there are:

- no common variable names,
- no common format,
- no common software,
- no management by an identical database management system,
- no common storage system, *e.g.*, as SPSSX/SAS system files.

There is no central database for hosting the various national datasets. Furthermore: any database management software creates problems because most researchers are not willing to learn the intricacies of different databank systems. They want to work with statistical packages with which they are familiar.

At the present stage, internationally comparative studies on panel data are feasible only by teams which actively involve people from the respective domestic panels.

Single researchers are not in a position to progress in comparative analysis without help or without close contact with the respective national panel teams.

Without a central or common databank, it is practically impossible to cope **SYSTEMATICALLY** with the tasks to be undertaken to standardize each of the variables of each of the panels, in order to work out in detail the concepts and the definitions needed for harmonized analyses.

3. THE PACO APPROACH
(PANEL COMPARABILITY)

In order to overcome these problems, CEPS/INSTEAD is creating in Luxembourg - firstly in partnership with DIW Berlin (Deutsches Institut für Wirtschaftsforschung - German Institute for Economic Research) - an international comparative database with various national household panels.

The PACO Project's aim is to develop instruments for analyzing, programming and simulating socio-economic policies. It is intended to facilitate comparative cross-national research on policy issues, **POLICY ISSUES** like labour force participation, income distribution, poverty, problems of the elderly, and so on.

At the bottom level, PACO will progressively set up a data archive of household panels existing in Europe and the USA. PACO will start by bringing together the panel data from PSID (see Hill, M.S. 1992) (USA), SOEP (see Wagmer, G.G. *et al.* 1991) (Germany) and PSELL (see Hausman, P. 1987; Schmaus, G. 1987) (Luxembourg). New Household Panels which are starting now in Europe will be added when the datasets are available.

At the second and more important level, PACO will add value to the original panel data by creating **COMPATIBILITY** and **COMPARABILITY**. This means that the PACO database will contain harmonized and standardized variables both at the cross-sectional **AND** at the longitudinal level: with identical variable names, corresponding to a common plan established for defining and recoding variables. The strict comparative approach makes it up to now unique[2] at present.

Characteristics of the PACO Database:

- access to harmonized panel variables,
- access to LIS variables,
- possibility to access original variables,
- standardized variables names,
- common format,
- common software,
- storage in a **relational** database **structure**,
- storage as SPSSX system files,
- possibility of raw data output.

The database will be expanded in a second step by a documentation system (**META-DATABANK**) and hopefully in a third step by an **INSTITUTIONAL DATABASE**.

META-DATABANK: It is planned to integrate all necessary information on original and standardized variables into the documentation system (on PC) which CEPS/INSTEAD has developed for its own Household Panel. Additional documentation about the newly created comparable variables, in machine readable and in written form, will be prepared. But in our first step we are only able to collect the original user manuals of each panel study and make them available to PACO users.

INSTITUTIONAL DATABASE[3]: The interpretation of results from cross-national research with panel surveys requires adequate information on the countries' systems of social security, taxation, schooling, *etc.* When setting up the micro-database, we will have to develop the institutional database in close conjunction. In this field, highly valuable documents and techniques have been gathered and developed by the LIS PROJECT.

The PACO database should contain **AS MANY COMPARABLE VARIABLES AS POSSIBLE**. Each panel carries a set of questions which are identical from wave to wave. These **CORE QUESTIONS** are the first candidates for variables to be standardized. By screening the questionnaires of the various country panels for core questions available in each, it will be possible to set up the first:

LIST OF CORE VARIABLES:

- Demographic variables,

[2] For example, the Syracuse project " Cross national Studies in Aging" brings together the entire German Panel SOEP and the entire US PSID on the same computer and within a common software environment. But the two databases are **not** harmonized on the level of variables.

[3] In Collaboration with LIS and the ASEG Project at the University, Frankfurt (Director Prof. Richard Hauser).

- Income variables,
- Labour Force variables,
- Unemployment variables,
- Education variables,
- Housing variables,
- Calendar variables.

In order to include PSID into PACO, the core of questions for all included panel studies is relatively small because the PSID is not really a panel study of **persons** but of head of households. The PSID gives a lot of proxy information for their spouses, but only very little information about other adult members in their households. The European panels contain much more information about those adults in households who are neither the head nor the spouse.

A SECOND LIST will contain the variables with the **HISTORY OF INDIVIDUALS** before they entered the panel study. The following topics are available in most files and can be harmonized:

- Family Background,
- Education History,
- Employment History,
- Marriage History,
- Fertility History.

Any national panel has or may have - in addition to the core issues and questions - specific components which do not appear in many of the other panels or which do appear only in one or in a limited number of its own waves. Such components are very poor candidates for harmonization and they will be stored only in their original form.

AS A RULE: The **PACO RESULT FILES** should contain all variables that can be standardized. **MOREOVER** the user of the result file will have the possibility of **accessing** those **ORIGINAL VARIABLES** in the panel studies which have not been made comparable for some reason. This procedure allows researchers to simultaneously access original and harmonized variables.

PACO will create files on the **HOUSEHOLD-** and on the **INDIVIDUAL- LEVEL**. Each file will contain variables for one year and one panel data set. Additional identifiers will guarantee that **MATCHES** and **AGGREGATIONS** between the single files are possible.

All files will be **not** held in a database management system.
The PACO Files at CEPS are stored as SPSS Files. An export to SAS is very easy.

Although it is against the conventional wisdom of computer scientists, it is possible to create a relational data structure without using a full database management system such as SQL products like ORACLE and other software like INGRES. Well known statistical packages (see Schmaus, G. 1992; Witte, J. 1992) like SPSS and SAS also allow one to store and access panel data in a relational manner.

Additionally the packages have the advantage that a researcher who knows "his" package must not learn the intricacies of a database management system. This allows statisticians to create standard work files for researchers and to create specific files for very particular analyses.

We plan to standardize the variables, the file structures and the access system in such a form that the analysis of different panel studies in a cross-national and longitudinal context will be possible with a **MINIMUM NEED FOR MODIFICATION OF PROGRAMS** which have been written for one country. This will be the case at least for standard tabulations and standard analyses. More complex analyses could probably not be standardized in such a way, but will be efficiently supported by the data organization.

In the first phases of this project not all required harmonized variables are already available. Therefore, a researcher may need to create some harmonized variables from the non-standardized data archive files and to match these variables with those from the harmonized panel database. This will not be difficult because unique identifiers will allow the matching of both types of files.

4. A VERY FIRST PACO-EXAMPLE

Labour market dynamics are a good example for demonstrating the **possibilities** of the PACO-files and the **limitations** caused by different concepts of panel studies.

The core variables of PACO presently available already allow us to meaningfully deal with issues of labour market and income dynamics as well as with demographic variables. We cannot do much more right now because of the limited number of variables available at the personal level in PSID.

Comparative analyses undertaken in the PACO project once we include PSID data will have to use a smaller number of variables than otherwise (since PSID deals in this respect only with heads and spouses of heads)[4].

One of the key questions of labour market dynamics in the United States and in European countries is the importance of "marginal jobs" which are unstable and low paid jobs with very few working hours. For example in Germany all jobs of less than 20 weekly working hours are "marginal", because they are not covered by the social security system. It is a common critique from the European point of view that the American "job machine" works by means of marginal jobs which make it impossible to reach a reasonable welfare position for the "marginal employees". Moreover those jobs could be an important reason for the productivity slowdown in the United States. On the other hand, marginal jobs are typical jobs for spouses who are not interested in full-time jobs. Marginal jobs raise the welfare of this particular group.

A comparison of the United States, Luxembourg and (West) Germany may be fruitful, because the overall labour market structures are very different. The labour market of the USA is less regulated (in comparison with Europe), the unemployment rate is an average one and female labour supply is high. The labour market in Germany is strictly regulated, the unemployment rate is high (compared to other high-wage countries) and female labour supply is low. From an economic point of view, Luxembourg appears in the eighties as a very wealthy country with strict labour regulation, full employment (creating jobs picked up to over 60% by transborder commuters) and low labour supply of females.

Only panel data allow the researcher to address the question of whether marginal jobs are permanent jobs for heads of households or jobs held by spouses over a short period. Furthermore panel data allow the research to estimate very tricky labour supply functions. So far we can present some preliminary descriptive results. They show that a comparative analysis of panel data is possible, but demonstrate also that even for some simple descriptive figures a lot of problems arise in making variables comparable.

In order to analyze labour market dynamics for some categories of jobs, we define one labour status variable:

- non-active	0	hours of labour force participation,
- marginal employed	1-19	weekly working hours,
- part-time employed	20-29	weekly working hours,
- full-time employed	30	and more weekly working hours.

From the German panel, for example, we know that approximately 50% of the marginally employed consider that they are not in the labour force. That means that they mark the category "unemployed". It is necessary to ask as well for "second jobs" ("Nebenerwerbstätigkeit") to capture their marginal jobs. In the PSID the interviewers must go back to the employment section if the respondent has a second job. Thus it is not necessary to bring together a first and a second job during analysis because this information is already on the PSID record.

[4] As mentioned above the PSID is a panel based on the responses of head of household. For households with two and more people, there is a lot of (proxy-) information about the spouse of the head but only very little about other household members - whereas in the European panels all adult members of the households are interviewed.

In the Luxembourg Panel "PSELL" we have no question asking about a marginal job addition to the main employment status. This may cause artefact. But so far we have discovered no better solution to make the three panels comparable.

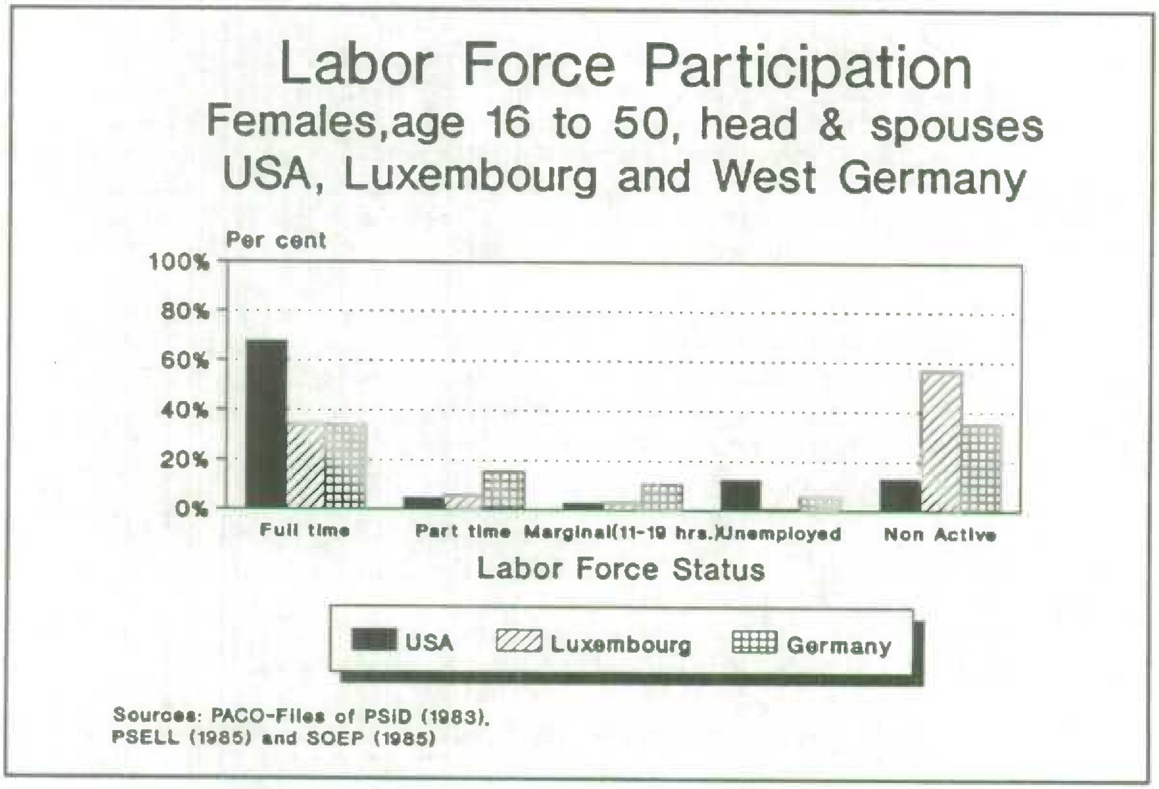
In order to avoid particular effects of the retirement transition process in our longitudinal results, we analyze people within the age bracket of 16 up to 50 only.

Even a look at the **cross-sectional** results for women is surprising: the share of female full-time labour force participation is highest in the USA, the share in Luxembourg and Germany is much lower. Part-time work is most important in Germany and not so important in USA and Luxembourg. In Germany, the share of marginal jobs is most important and in the USA and Luxembourg, it is very small. Unemployment is only a problem for Germany and USA. Low labour force participation and very low unemployment in Luxembourg have the effect that the share of non active women is highest for that country.

A look at the **longitudinal results** gives the overall impression that the labour market dynamics in the three countries are very similar. But it is remarkable that in the USA the probability that a marginal job will lead to a full job is high. In West Germany the probability of staying in a marginal job is highest. In Luxembourg the probability of progressing into a regular part-time job is the highest.

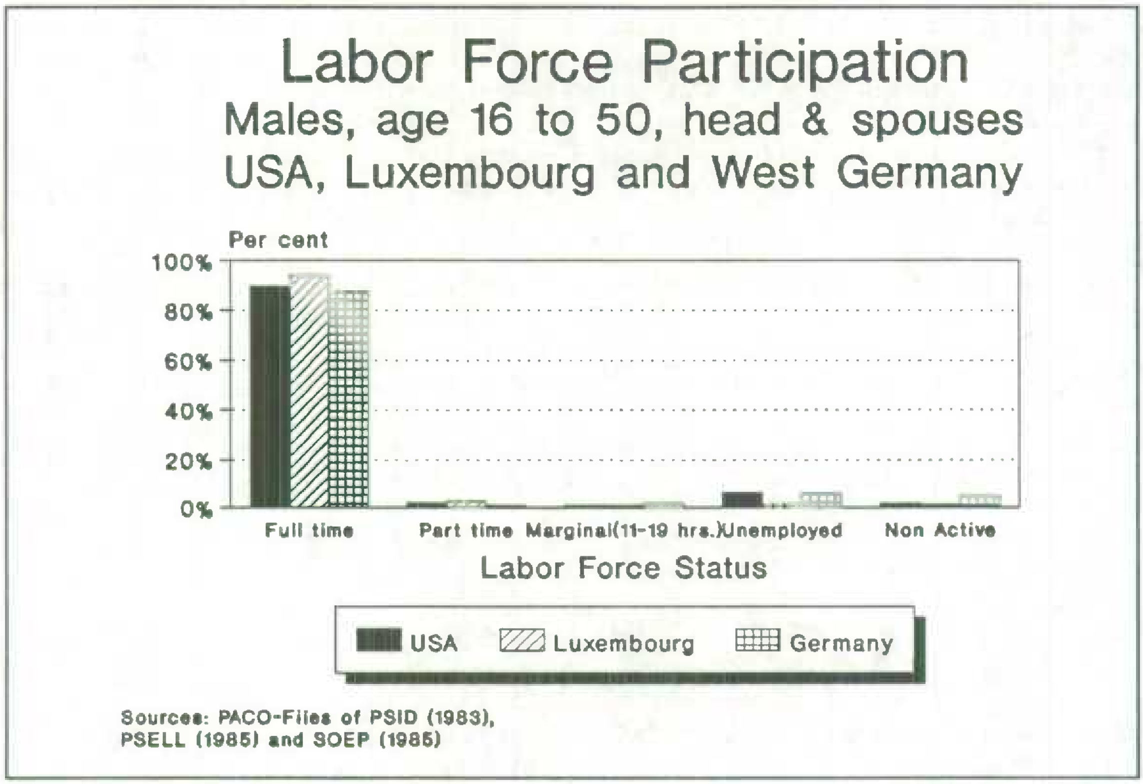
These results do not fit our preconceptions of the Labour markets in Europe and the USA. Thus they stimulate further research. This is an encouraging first step for PACO.

Table 1.



Sources: PACO-Files of PSID (1983). PSELL (1985) and SOEP (1985)

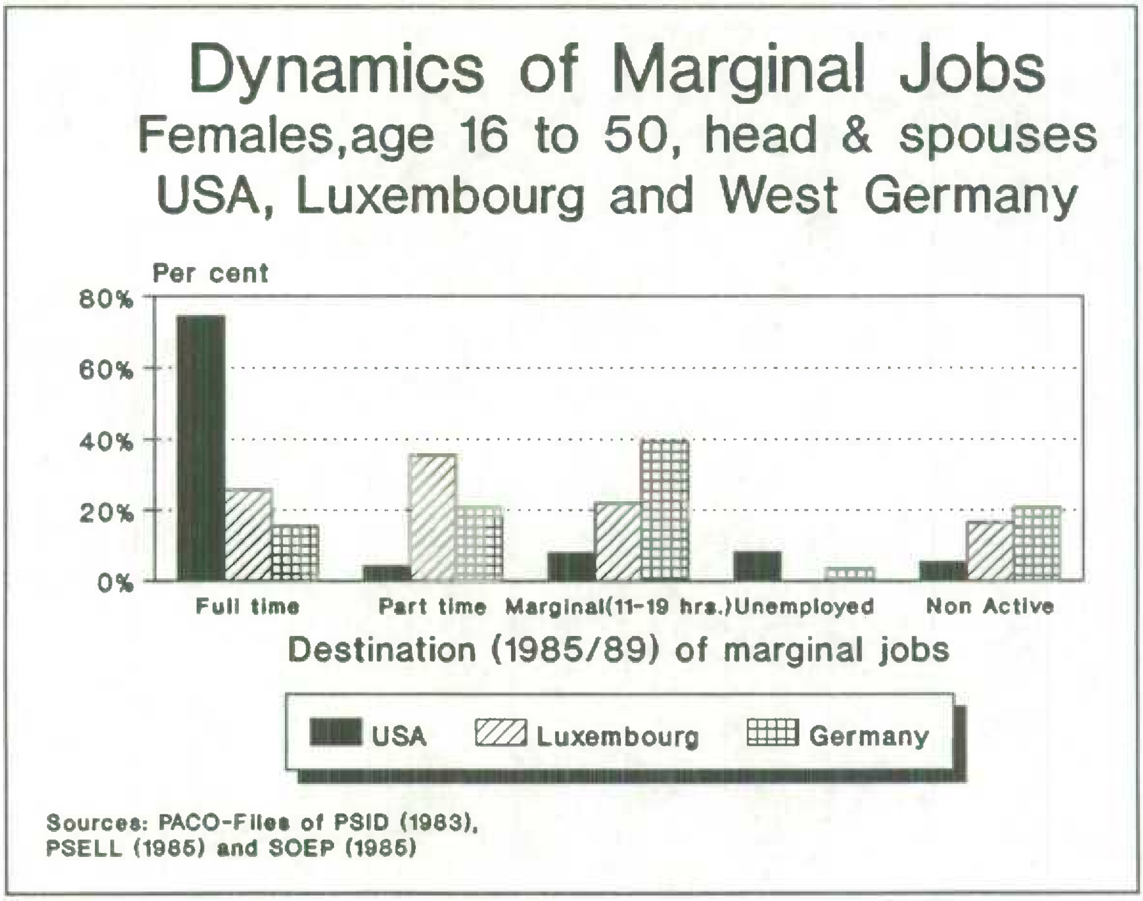
Table 2.



Labor Force Participation
Males, age 16 to 50, head & spouses
USA, Luxembourg and West Germany

Sources: PACO-Files of PSID (1983),
PSELL (1985) and SOEP (1985)

Table 3.



Dynamics of Marginal Jobs
Females,age 16 to 50, head & spouses
USA, Luxembourg and West Germany

Sources: PACO-Files of PSID (1983),
PSELL (1985) and SOEP (1985)

REFERENCES

Burkhauser, R.V. An introduction to the German Socio-economic panel for English speaking researchers. Cross-National Studies in Aging, Program Project Paper No. 1. All-University Gerontology Center, Maxwell School of Citizenship and Public Affairs, Syracuse University New York.

Duncan, G.J., Gustafsson, B., Hauser, R., Schmaus, G., Laren, D., Messinger, H., Muffels, R., Nolan, B., and Ray, J.-C. (1991). Poverty dynamics in eight countries, (mimeo). Survey Research Center Ann Arbor, MI, U.S.A. (October).

Hausman, P. (1987). Niveaux de vie et de bien-être économique des ménages en 1985: principaux résultats. PSELL Working Paper Number 4. CEPS/INSTEAD, Walferdange, Luxembourg.

Hill, M.S. (1992). The Panel Study of Income Dynamics - A Users's Guide Newbury Park, London, New Delhi.

Schmaus, G. (1987). Organization of the database for the Luxembourger household panel (input, storage and analysis). PSELL Working Paper Number 17a. CEPS/INSTEAD, Walferdange, Luxembourg.

Schmaus, G. (1992). Storage and retrieval of data from the Luxembourg. Household Panel (PSELL). ESF Working Paper Number 9. CEPS/INSTEAD and University of Essex.

Smeeding, T.M., and Schmaus, G. (1988). LIS Information Guide (Revised, November). LIS Working Paper Number 7, CEPS/INSTEAD, Walferdange.

Smeeding, T.M., and Schmaus, G. (1990). The LIS Database: Technical and Methodological Aspects. In (Eds.) T.M. Smeeding, M. O'Higgins and L. Rainwater. Poverty, Inequality and Income distribution in Comparative Perspective Hemel Hempstead Herfordshire.

Wagner, G.G. *et al.* (1991). The socio-economic panel (SOEP) of Germany - methods of production and management of longitudinal data. DIW Discussion Paper Number 31a, Berlin.

Witte, J. (1992). Data management and analysis of the German socio-economic panel using SPSS, DIW documentation. Berlin.

Proceeding of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

ANALYSIS OF LONGITUDINAL BUSINESS TEST DATA WITH ORDERED CATEGORICAL VARIABLES

G. Arminger[1]

ABSTRACT

Business test data from a panel of firms usually come from questionnaires. Hence, many of the dependent variables are non-metric, say dichotomous, censored or ordered categorical. Heckman's (1981) model for dichotomous panel data is extended in two ways. First, the dichotomous threshold model is generalized to probit-type threshold models for metrically classified, censored and ordered categorical variables. Second, the single equation model is expanded to simultaneous equation models and to factor analytic measurement models. Special attention is given to panel specific problems such as initial states and restrictions on the covariance matrix of errors. As application, a model of simultaneous decisions about price, production and inventory is specified and estimated for German business test data. The model parameters are estimated using the MECOSA program.

KEY WORDS: Dichotomous threshold model; Polychloric covariance coefficient; censured variable.

1. INTRODUCTION

The analysis of business test data is typically concerned with the analysis of non-metric questionnaire data from a sample of a population of firms. To discover longitudinal developments, the firms in a business test survey are often interviewed consecutively yielding a set of panel data. In the following section the construction principles of the seminal work of Heckman (1981a, 1981b) on the specification and estimation of dichotomous outcomes in panel data are extended to include ordered categorical and censored dependent variables as well as simultaneous equation systems of non-metric dependent variables. The parameters of such models are estimated by assuming multivariate normality of the error terms in threshold models and using conditional polychoric and polyserial covariance coefficients in the framework of mean- and covariance structure models for non-metric dependent variables. These models and estimation techniques have been introduced by Muthén (1984) and extended by Küsters (1987) and Schepers and Arminger (1992). Special attention is given to the problems of unobserved heterogeneity, initial states and identification of scale. The discussion and illustration of models and methods is restricted to the most common case (state dependence and random unobserved heterogeneity). As an example, the trichotomous output variable of a four wave panel of 656 firms from the German business test regularly conducted by the IFO Institute, Munich, is analyzed.

2. MODEL SPECIFICATION

2.1 Heckman's model for dichotomous variables

Heckman (1981a, ch. 3.3) considers the following model for an unobserved variable y_{it}^*, $i = 1,...,n$, $t = 1,...,T$ where i denotes the individual and t denotes a sequence of equispaced time points:

$$y_{it}^* = \mu_{it} + \epsilon_{it}^* \qquad (1)$$

[1] G. Arminger, Department of Economics (FB 6), Bergische Universität, Gaußstr. 20, D-5600 Wuppertal, Germany.

$$\mu_{it} = x_{it}\beta + \sum_{j=1}^{\infty} \gamma_{t-j,t} y_{i,t-j} + \sum_{j=1}^{\infty} \lambda_{j,t-j} \prod_{l=1}^{j} y_{i,t-l} + \sum_{k=1}^{K} \delta_k y_{i,t-k}^{*} \tag{2}$$

$$\epsilon_i^{*} = (\epsilon_{i1}, ..., \epsilon_{iT}^{*})', \quad \epsilon_i^{*} \sim N\ (0, \Sigma\,), \tag{3}$$

$$y_{it} = \begin{cases} 1 \text{ if } y_{it}^{*} > 0 \\ 0 \text{ if } y_{it}^{*} \leq 0 \end{cases} . \tag{4}$$

The unobserved variable y_{it}^{*} is considered as a disposition or utility which is connected to the observed dependent variable y_{it} through a dichotomous threshold model with threshold 0. The values y_{it}, y_{it}^{*} and x_{it} are collected in $T \times 1$ vectors y_i, y_i^{*} and the $R \times 1$ vector $x_i = (x_{i1}, ..., x_{iT})'$. The random variables $\{y_i, x_i\}$ are identically and independently distributed which corresponds to a simple random sample from a population. The model specification consists in determining the structure of the systematic part μ_{it} and of the stochastic part ϵ_{it}^{*} of the model. A discussion of the different parts of the model specification for y_{it}^{*} is found in Heckman (1981a) and in Hamerle and Ronning (1993). We repeat here only the most important elements of the specification.

The first component of μ_{it} is the variation induced by possibly time varying explanatory variables x_{it}. The parameter vector β is time constant in this model which may changed to parameter vectors β_t, $t = 1, ..., T$ that vary over time.

The second component of μ_{it} captures the influence of former states of the observed dependent variable $y_{i,t-j}$, $j \geq 1$ which is called true state dependence. If $\gamma_{t-1,t} = \gamma_1 \neq 0$ and $\gamma_{t-j,t} = 0$ for all $j > 1$ and $t = 1, ..., T$ we have a simple Markov model. Note that the inclusion of former states requires the knowledge of initial states $y_{i0}, y_{i,-1}, ...$ depending on the specification of the parameters $\gamma_{t-j,t}$. If the initial states are known and non-stochastic they can be included in the vectors x_{it} as additional explanatory variables. If the initial states are themselves outcomes of the process that generates y_{it}^{*}, the distribution of the initial states must be taken into account as discussed by Heckman (1981b) and in section 3 of this paper. Note that the effects of the former states $y_{i,t-j}$ may change for each time point. This is captured by $\gamma_{t-j,t}$. In most applications, $\gamma_{t-j,t}$ is set to γ_{t-j} and almost all parameters are set to 0.

The third component of μ_{it} models the dependence of y_{it}^{*} on the duration of the state $y_{i,t} = 1$. Again, the effects of duration may be different for each time length. These different effects are parameterized in $\lambda_{j,t-j}$. If $\lambda_{j,t-j} = \lambda$ we have the simple case of a linear duration effect.

The fourth component of μ_{it} models the dependence of y_{it}^{*} on former values $y_{i,t-j}^{*}$ of the unobserved endogenous variables. Structures that incorporate this component are called models with habit formation or habit persistence. The idea behind such models is that y_{it}^{*} is not dependent on the actual former state of the observed variable but on the former disposition or habit identified rather with $y_{i,t-j}^{*}$ than $y_{i,t-j}$. If the initial dispositions $y_{i,0}^{*}, y_{i,-1}^{*}, ..., y_{i,-(K-1)}^{*}$ are known and non-stochastic they may be included in the list of explanatory variables. Otherwise, assumptions about the distribution of the initial variables $y_{i,0}^{*}, y_{i,-1}^{*}, ..., y_{i,-(K-1)}^{*}$ must be made.

Now we turn to the specification of the error term ϵ_{it}^{*}. Throughout this paper it is assumed that the error terms are uncorrelated with past, present and future explanatory variables (strong exogeneity). The method of Keane and Runkle (1992) to deal with weak exogeneity cannot be extended to limited dependent variables in a straightforward manner. The usual decomposition of ϵ_{it}^{*} is

$$\epsilon_{it}^{*} = \alpha_i + \epsilon_{it}, \tag{5}$$

where α_i denotes the error term which varies across individuals but not over time and may be considered as unobserved heterogeneity just as in the metric case (*cf.* Hsiao 1986). The values of α_i may be considered as fixed effects for each i or may be considered as random effects such that $\alpha_i \sim N(0, \sigma_\alpha^2)$. In the first case, α_i is an individual specific parameter. If y_{it}^{*} is modelled by $y_{it}^{*} = x_{it}\beta + \alpha_i + \epsilon_{it}^{*}$ and is actually observed as in the metric case then α_i may be eliminated by taking the first differences $y_{it}^* - y_{i,t-1}^* = (x_{it} - x_{i,t-1})\beta + \epsilon_{i,t}^* - \epsilon_{i,t-1}^*$. This technique does not work for non-metric models such as the probit model. In the case of a dichtomous logit model, the α_is may be eliminated by conditioning on a sufficient statistic as shown in Hsiao (1986) and Hamerle and Ronning (1993). If α_i is a random variable, then it is assumed to be uncorrelated with z_{it} and ϵ_{it}. If ϵ_{it} has constant variance σ_ϵ^2 and is serially uncorrelated, then ϵ_i^* has the typical covariance structure:

$$V(\epsilon_i^*) = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\epsilon^2 & & & \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{bmatrix} = \sigma_\alpha^2 11' + \sigma_\epsilon^2 I,$$

where 1 is a $T x 1$ vector of ones and I is the $T x T$ identity matrix. More generally, a serial or a factor analytic structure may be assumed for $V(\epsilon_i^*)$. Details are found in Heckman (1981a) or Arminger (1992). Heckman (1981a) discusses the ML estimation of these models under the assumption that ϵ_i^* is normally distributed. Furthermore, he assumes that only dichotomous outcomes are observed. Since $\epsilon_i^* \sim N(0, \Sigma)$ the parameter vector of the explanatory variables can be estimated from a univariate probit model for each y_{it} with the usual identification restriction $\sigma_{tt} = 1, t = 1, ..., T$ which implies that β may only be estimated up to a scale factor. Note that this identification restriction is not required for all panel waves. If β is constant for all waves then the variance of ϵ_{it}^* need only be restricted for one wave, usually the first (*cf.* Arminger 1987).

2.2 Extension to general threshold models

A thorough discussion and some extensions of Heckman's models to fixed effects logit and tobit models as well as to random effects probit models are found in Madalla (1987). We extend now Heckman's (1981a) dichotomous models in a systematic way to censored, metrically classified and ordered categorical dependent variables and simultaneous equation systems that allow as dependent variables any mixture of metric and/or limited dependent variables. Only random effect models are considered. Like Heckman (1981a) we consider the case that the initial states are known and non stochastic or that information about the distribution of the initial states is known so that the initial states can either be modelled simultaneously with the dependent variables or can be included in the explanatory variables. In this case the $T x 1$ vector y_i^* may be written as

$$y_i^* = \gamma + \Pi x_i + \epsilon_i^*, \tag{6}$$

where ϵ_i^* follows a normal distribution with $\epsilon_i^* \sim N(0, \Sigma)$. The $R x 1$ vector x_i contains the explanatory variables $x_{it}, t = 1, ..., T$. The $T x 1$ vector γ is the vector of regression constants. The $T x R$ matrix Π is the matrix of regression coefficients. If $\mu_{it} = x_{it}\beta_t$, then Π is structured as follows:

$$\Pi + \begin{bmatrix} \beta_1 & 0 & \ldots & 0 \\ 0 & \beta_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \beta_T \end{bmatrix}. \tag{7}$$

In the case of time constant parameters, the vector β_t is replaced by β.

Together with the specification of Σ through a model for unobserved heterogeneity and for serial correlation, the above specification of $y_i^* = \gamma + \Pi x_i + \epsilon_i^*$ yields a conditional mean and covariance structure in the latent variable vector y_i^* with $y_i^* \sim N(\gamma + \Pi x_i, \Sigma)$.

The model is now extended by allowing not only the dichotomous threshold model of equation (4), but any one of the following threshold models that maps y_{it}^* onto the observed variable y_{it} (*cf.* Schepers, Arminger and Küsters 1991). For convenience, the case index $i = 1,...,n$ is omitted.

- y_t is metric (identity relation).

$$y_t = y_t^* . \tag{8}$$

- y_t is ordered categorical with unknown thresholds $\tau_{t,1} < \tau_{t,2} < ... < \tau_{t,K_t}$ and categories $y_t = 1,...,K_t + 1$ (ordinal probit relation, McKelvey and Zavoina 1975).

$$\begin{gathered} y_t = k \Longleftrightarrow y_t^* \in [\tau_{t,k-1}, \tau_{t,k}) \quad \text{with} \\ [\tau_{t,0}, \tau_{t,1}) = (-\infty, \tau_{t,1}) \quad \text{and} \quad \tau_{t,K_t+1} = +\infty . \end{gathered} \tag{9}$$

Note that for reasons of identification the threshold $\tau_{t,1}$ is set to 0 and the variance of the reduced form error term σ_t^2 is set to 1. The parameters in β_t are only identified up to scale. If one considers simultaneous equation models or analyzes two or more panel waves simultaneously, only hypotheses of proportionality of regression coefficients across equations can be tested in general. Hypotheses of equality of regression coefficients across equations can only be tested under additional, sometimes non-testable, assumptions (Sobel and Arminger 1992).

- Classified metric variables may be treated analogously to the ordinal probit case with the difference that the class limits are now used as known thresholds (Stewart 1983). No identification restrictions are necessary.

- y_t is one-sided censored with a threshold value $\tau_{t,1}$ known a priori (tobit relation, Tobin 1958).

$$y_t = \begin{cases} y_t^* & \text{if } y_t^* > \tau_{t,1} \\ \tau_{t,1} & \text{if } y_t^* \leq \tau_{t,1} \end{cases} . \tag{10}$$

- y_t is double-sided censored with threshold values $\tau_{t,1} < \tau_{t,2}$ known a priori (two-limit-probit relation, Rosett and Nelson 1975).

$$y_t = \begin{cases} \tau_{t,1} & \text{if } y_t^* \leq \tau_{t,1} \\ y_t^* & \text{if } \tau_{t,1} < y_t^* < \tau_{t,2} \\ \tau_{t,2} & \text{if } \tau_{t,2} \leq y_t^* \end{cases} . \tag{11}$$

In the case of general threshold models several modifications of the dichtomous models for panel data have to be considered. First, the state dependence must be modified for non-dichotomous dependent variables. In the case of a censored dependent variable there should be a dummy variable $d_{i,t-j}$ which takes on the value of 1 if the variable $y_{i,t-j}^*$ has been observed and a value of 0 if $y_{i,t-j}$ takes on the threshold value. For metrically classified variables and ordered categorical variables a $K \times 1$ vector of dummy variables $d_{i,t-j}^{(k)}$ must be defined. The dummy variable $d_{i,t-j}^{(k)}$ equals 1 if $y_{i,t-j}^*$ falls in category $k, k = 2,...,K + 1$ and is 0 otherwise.

Second, the identification restrictions for the variances $\sigma_t^2, t = 1,...,T$ of ϵ_{it}^* and for the thresholds collected in the vector $\tau^{(t)}$ for each panel wave t must be chosen with great care. In my opinion, the threshold vectors $\tau^{(t)}$ should be set equal across panels waves, otherwise the meaning of the categories of an ordered categorical variable is assumed to vary across time. This restriction immediately implies that, except for the first wave, the variances σ_t^2 need not be restricted but can vary across the panel waves. For censored dependent variables, no restrictions for σ_t^2 are necessary.

A further extension concerns the specification of models for a system of variables. Instead of considering only one variable over T waves one may consider a vector y_{it}^* of H dependent variables over time. Each component of y_{it}^* is denoted by $y_{ith}^*, h = 1,...,H$. The vector of dependent variables y_i^* of the i th observation is then a $H \cdot T \times 1$ vector of dependent variables observed at T time points. Each latent variable y_{ith}^* at each time point is then mapped onto the observation y_{ith} through a threshold model of the form given above (different thresholds may be used for each component). The covariance matrix Σ of ϵ_i^* contains in this case not only the conditional serial covariance structure for each variable y_{ih}^* but also the conditional covariance structure between the variables across all time points. An example for such a model is found in Arminger and Ronning (1991).

3. ESTIMATION METHOD

3.1 Conditional polychoric correlation coefficients

For general mean and covariance structures we assume that a $P \times 1$ vector y_i^* of latent dependent variables follows a multivariate normal distribution with conditional mean and covariance:

$$\begin{aligned} E(y_i^* | x_i) &= \gamma(\vartheta) + \Pi(\vartheta)x_i, \\ V(y_i^* | x_i) &= \Sigma(\vartheta). \end{aligned} \tag{12}$$

In the analysis of panel data, P equals T if a univariate dependent variable is analyzed or P equals $H \cdot T$ if a multivariate dependent variable is analyzed $\gamma(\vartheta)$ is a $P \times 1$ vector of regression constants and $\Pi(\vartheta)$ is a $P \times R$ matrix of reduced form regression coefficients. x_i is a $R \times 1$ vector of explanatory variables. $\Sigma(\vartheta)$ is the $P \times P$ covariance matrix of the errors of the reduced form. ϑ is the $\bar{q} \times 1$ vector of structural parameters to be estimated. The reduced form parameters $\gamma(\vartheta)$, $\Pi(\vartheta)$ and $\Sigma(\vartheta)$ are continuously differentiable functions of a common vector ϑ. Typical examples are simultaneous equation systems:

$$y_i^* = By_i^* + \Gamma x_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0,\Omega), \tag{13}$$

with the reduced form parameters

$$\Pi(\vartheta) = (I - B)^{-1}\Gamma \qquad \text{and} \quad \Sigma(\vartheta) = (I - B)^{-1}\Omega(I - B)^{-1'} \tag{14}$$

and confirmatory factor analysis

$$y_i^* = \Lambda\eta_i + \epsilon_i \quad \text{with} \quad \eta_i \sim N(0,\Phi) \quad \text{and} \quad \epsilon_i \sim N(0,\Theta) \tag{15}$$

and the reduced form parameters

$$\Pi(\vartheta) = 0, \quad \Sigma(\vartheta) = \Lambda\Phi\Lambda' + \Theta. \tag{16}$$

In the first example ϑ consists of the structural parameter matrices B, Γ and Ω. In the second example, ϑ consists of Λ, Φ and Θ.

Note that the typical structure of panel models, that is $\mu_{it} = x_{it}\beta$ cannot be embedded in the reduced form of equation (13) directly. If one uses the model of (13) with $x_i = (x_{i1},...,x_{iT})'$ then the dependent variables y_{it}^* are regressed not only on x_{it} but also on all other variables $x_{is}, s \neq t$. The parameter restrictions that apply to the reduced form have to be introduced at the third stage of the estimation procedure.

The estimation of the structural parameter vector from the observed data vector y_i proceeds in three stages. This section is based on Schepers, Arminger and Küsters (1991). Computation of the estimates with the MECOSA program is described in Schepers and Arminger (1992).

1. In the first stage the threshold parameters τ, the reduced form coefficients γ and Π of the regression equation, and the reduced form error variance σ_t^2 of the t th equation are estimated using marginal maximum likelihood. Note that this first stage is the estimation of the mean structure without restrictions as in equation (13). The parameters to be estimated in the t th equation are the thresholds denoted by the vector τ_t, the regression constant denoted by γ_t, the regression coefficients, *i.e.* the t th row of Π denoted by $\Pi_{t.}$ and the variance denoted by σ_t^2.

2. In the second stage the problem is to estimate the covariances of the error terms in the reduced form equations. Note that in this stage the covariances are estimated without parametric restrictions. Since the errors are assumed to be normally distributed and strongly consistent estimators of the reduced form coefficients have already been obtained in the first stage the estimation problem reduces to maximizing the loglikelihood function

$$l_{tj}(\sigma_{tj}) = \sum_{i=1}^{n} \ln P(y_{it}, y_{ij} \mid x_i, \hat{\tau}_t, \hat{\gamma}_t, \hat{\Pi}_{t.}, \hat{\sigma}_t^2, \hat{\tau}_j, \hat{\gamma}_j, \hat{\Pi}_{j.}, \hat{\sigma}_j^2, \sigma_{tj}) \ , \qquad (17)$$

in which $P(y_{it}, y_{ij} \mid x_i, \hat{\tau}_t, \hat{\gamma}_t, \hat{\Pi}_{t.}, \hat{\sigma}_t^2, \hat{\tau}_j, \hat{y}_j, \hat{\Pi}_{j.}, \hat{\sigma}_j^2, \sigma_{ti})$ is the bivariate probability of y_{it} and y_{ij} given x_i and the reduced form coefficients. A typical example of this bivariate probability is the case when y_t and y_j are both ordinal. Then the probability that $y_{it} = k$ and $y_{ij} = l$ is given by:

$$P(y_{it} = k, y_{ij} = l \mid x_i) = \int_{\hat{\tau}_{t,(k-1)}}^{\hat{\tau}_{t,(k)}} \int_{\hat{\tau}_{j,(l-1)}}^{\hat{\tau}_{j,(l)}} \varphi(y_t^*, y_j^* \mid \hat{\mu}_{it}, \hat{\sigma}_t^2, \hat{\mu}_{ij}, \hat{\sigma}_j^2, \sigma_{tj}) \, dy_j^* \, dy_t^* \qquad (18)$$

in which $\hat{\mu}_{it} = \hat{\gamma}_t + \hat{\Pi}_{t.} x_i$, $\hat{\mu}_{ij} = \hat{\gamma}_j + \hat{\Pi}_{j.} x_i$ and $\varphi(y_t^*, y_j^* \mid \mu_t, \sigma_t^2, \mu_j, \sigma_j^2, \sigma_{tj})$ is the bivariate normal density function. Note that in the ordinal case $\sigma_t^2 = \hat{\sigma}_j^2 = 1$. Hence, σ_{tj} is a correlation coefficient which is called the polychoric correlation coefficient. The loglikelihood function $l_{tj}(\sigma_{tj})$ has to be modified accordingly if variables with other measurement levels are used. Note that the covariances σ_{tj} are the covariances of the error terms in the equations for $y_t^*, t = 1, \ldots, P$ conditional on x_i. In contrast to LISREL 7, it is not assumed that the variables y_{it}^* and y_{ij}^* are jointly normal. It is only assumed that the errors are normal.

   The estimated thresholds $\hat{\tau}_t$, the reduced form coefficients $\hat{\gamma}_t$ and $\hat{\Pi}_{t.}$, the variances $\hat{\sigma}_t^2$ and the covariances $\hat{\sigma}_{tj}$ from all equations are then collected in a vector $\hat{\kappa}_n$ which depends on the sample size n. For the final estimation stage, a strongly consistent estimate of the asymptotic covariance matrix W of $\hat{\kappa}_n$ is computed. This estimate is denoted by $\hat{W}_n$. The asymptotic covariance matrix W is difficult to derive since the estimates of $\hat{\sigma}_{tj}$ of the second stage depend on the estimated coefficients $\hat{\tau}_f, \hat{\gamma}_f, \hat{\Pi}_{f.}, \hat{\sigma}_f^2, f = t, j$ of the first stage. The various elements of the asymptotic covariance matrix W are given in Küsters (1987). The estimate $\hat{W}_n$ is computed in MECOSA by using analytical first order and numerical second order derivatives of the first and second stage loglikelihood function.

3. In the third stage the vector κ of thresholds, the reduced form regression coefficients and the reduced form covariance matrix is written as a function of the structural parameters of interest, collected in the parameter vector ϑ. The parameter vector ϑ is then estimated by minimizing the quadratic form

$$Q_n(\vartheta) = (\hat{\kappa}_n - \kappa(\vartheta))' \hat{W}_n^{-1} (\hat{\kappa}_n - \kappa(\vartheta)), \qquad (19)$$

   which is a minimum distance approach based on the asymptotic normality of the estimators of the reduced form coefficients. The vector $\hat{\kappa}_n$ is asymptotically normal with expected value $\kappa(\vartheta)$ and covariance matrix

W. Since $\hat{W}_n$ is a strongly consistent estimate of W the quadratic form $Q_n(\vartheta)$ is centrally χ^2 distributed with $p - \bar{q}$ degrees of freedom if the model is specified correctly and the sample size is sufficiently large. The number p indicates the number of elements in $\hat{\kappa}_n$ while $\bar{q}$ is the number of elements in ϑ. The computation of $\hat{W}_n$ is quite cumbersome for models with many parameters. As an alternative, one can use the weight matrix of the GLS approach in LISREL which may be based on the estimated covariance matrix of y_i^* and x_i. This covariance matrix may be computed from the estimated matrix $\hat{\Sigma}$ of the error terms, the matrix $\hat{\Pi}$ of the reduced form regression coefficients and the estimated covariance matrix of the explanatory variables. The function $Q_n(\vartheta)$ is minimized using the Davidon-Fletcher-Powell algorithm with numerical first derivatives.

The program MECOSA follows these three estimation stages. In the third stage, the facilities of GAUSS are fully exploited to be able to estimate parameters under arbitrary restrictions. The parameter vector $\kappa(\vartheta)$ may be defined using the matrix language and the procedure facility of GAUSS. Consequently, arbitrary restrictions can be placed on $\tau(\vartheta), \gamma(\vartheta), \Pi(\vartheta)$ and $\Sigma(\vartheta)$ including the restrictions placed on $V(\epsilon_i^*) = \Sigma(\vartheta)$ in the analysis of panel data.

The MECOSA program provides estimates of the reduced form parameters $\Pi(\vartheta)$ in the first stage by regressing all dependent variables in $y_i^* = (y_{i1}^*, ..., y_{iT}^*)'$ on all regressors collected in x_i without the usual restriction that the effect of x_{it} on $y_{i,t-j}, j \geq 1$ is zero. This restriction has to be put into the program in the third stage. If the model includes lagged dependent variables, it may be necessary to restrict the parameter estimates already in the first stage. In this case, one can trick the program into restricted estimation already in the first stage by running the first stage of MECOSA T times separately for each y_{it}^* which is then regressed on x_{it}. In the second stage, the standard batch input of MECOSA has to be corrected to deal with the T estimation results from the first stage. The polychoric covariances and/or correlation coefficients of the reduced form are then computed under the restrictions of the first stage.

3.2 The problem of initial states

For the estimation of the general Heckman model it was assumed that the initial states for a model with state dependence or with habit persistence are known and non-stochastic or that information about the initial states could be used in the list of explanatory variables. If this is not the case, one must assume that either a new process begins at the first wave of the panel or that the error terms of the process for the initial states are uncorrelated with the error terms of the T panel waves. (*cf.* Heckman 1981a, b). Note that the first assumption holds for the attrition process in a panel where the dependent variable is the participation of a person in panel wave t given that this person has participated in wave 1. At the first wave, all persons participate. Hence, the initial states are known and a new process starts at time 1.

Heckman (1981b) considers the following special case of the initial states problem:

$$\epsilon_{it}^* = \beta + \gamma y_{i,t-1} + \alpha_i + \epsilon_{it} \quad \text{with} \tag{20}$$

$$y_{it} = \begin{cases} 1 \text{ if } y_{it}^* > 0 \\ 0 \text{ if } y_{it}^* \leq 0 \end{cases}.$$

The random effects are $\alpha_i \sim N(0,\sigma_\alpha^2), E(\alpha_i \epsilon_{it}) = 0, \epsilon_{it} \sim N(0,\sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 1$.

In this model, the only initial value is y_{i0}. What happens if y_{i0} is not known and non-stochastic, but is determined by the same process as y_{it} for $t = 1, ..., T$? In this case, y_{i0} is dependent on α_i and the sample conditional likelihood function for $y_{it}, t = 1, ..., T$ given y_{i0} is given by integration over the unobserved variable α which may be written as $\alpha = \sigma_\alpha \eta$ with $\eta \sim N(0,1)$.

$$L(\beta_0, \gamma, \sigma_\alpha^2) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{t=1}^{T} (\Phi(\beta + \gamma y_{i,t-1} + \sigma_\alpha \eta)^{y_{it}} (1 - \Phi(\beta + \gamma y_{i,t-1} + \sigma_\alpha \eta))^{1-y_{it}} \tag{21}$$

$$\times P(y_{i0} | \sigma_\alpha \eta) \varphi(\eta) d\eta \ ,$$

where $\varphi(.)$ is the standard normal density. The term $P(y_{i0}|\alpha)$ is the probability that the random variable y_{i0}^* is either > 0 or ≤ 0. Possible specifications of $P(y_{i0}|\alpha)$ by introducing restrictions on the process generating the exogenous variables x_{it} and application of the estimation method of Kiefer and Wolfowitz (1956) are discussed by Heckman (1981b). However, for all practical purposes this estimation method is too cumbersome.

As a first simple solution for this estimation problem, Heckman proposes the estimation of α_i as a parameter by ML methods. This method, however, yields only consistent estimates of the structural parameters β and γ if $T \rightarrow \infty$ (*cf.* Neyman and Scott 1948 and Andersen 1973) which cannot be assumed for panel data. The limited Monte Carlo evidence given by Heckman (1981b) for $T = 8$ panel waves in the model of equation (21) is certainly not sufficient to recommend the ML estimation of the α_i's in conjunction with β and γ as a general solution.

The second simple solution proposed by Heckman (1981b) is the substitution of $P(y_{i0}|\alpha)$ by $F(x_{i0}\delta)$, where x_{i0} is a vector of explanatory variables for y_{i0}^* and $F(.)$ is the distribution function of the random variable

$$y_{i0}^* = x_{i0}\delta + \epsilon_{i0}^* \ . \tag{22}$$

This ad hoc solution is simply the attempt to replace the unobserved heterogeneity in y_{i0}^* by observed heterogeneity in x_{i0}. The error term is allowed to be correlated with α_i and ϵ_{it} for $t = 1,...,T$ without restrictions to capture the serial correlation between ϵ_{i0}^* and $\epsilon_{it}^* = \alpha_i + \epsilon_{it}$ induced by α_i. In practice, this procedure amounts to the inclusion of y_{i0}^* in the vector of dependent variables and additional parameters in the augmented covariance matrix for $\tilde{\epsilon}_i^* = (\epsilon_{i0}^*, \epsilon_{i1}^*, ..., \epsilon_{iT}^*)'$. If $\epsilon_{it}^* = \alpha_i + \epsilon_{it}$ with $V(\alpha_i) = \sigma_\alpha^2$ and $V(\epsilon_{it}) = \sigma_\epsilon^2$, then the augmented covariance matrix is given by

$$V(\tilde{\epsilon}_i^*) = \begin{pmatrix} \sigma_{00}^2 & & & & \\ \sigma_{10} & \sigma_\alpha^2 + \sigma_\epsilon^2 & & & \\ \sigma_{20} & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\epsilon^2 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \sigma_{T0} & \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\epsilon^2 \end{pmatrix} . \tag{23}$$

The limited Monte Carlo simulation of Heckman (1981b) shows that this approach works better than the simultaneous ML method to yield consistent estimates of β and γ. If $F(.) = \Phi(.)$, the above method can immediately be incorporated in the MECOSA framework by specifying

$$y_{i0}^* = x_{i0}\delta + \epsilon_{i0}^* \ . \tag{24}$$

$$y_{it}^* = x_{it}\beta + \epsilon_{it}^* \ . \tag{25}$$

The joint covariance matrix of ϵ_{i0}^* and ϵ_i^* may be structured as in equation (23).

This solution for the initial states problem depends crucially on the specification of a model for the initial states. The key idea is to specify a model for the initial states that gives as good predictions as possible. The model for the initial states need not be the same model as for y_{it}, $t = 1,...,T$. The errors $\epsilon_{i0}^*, \epsilon_{i,-1}^*, \epsilon_{i,-K}^*$ are allowed to correlate freely with ϵ_{it}^{*}. If the model for the initial states is misspecified, the simple solution will work badly. If the model for the initial states is correct, the ML estimators of B, Γ and $V(\epsilon_i^{*})$ will be consistent for fixed T as $n \rightarrow \infty$.

4. ANALYSIS OF PRODUCTION OUTPUT FROM GERMAN BUSINESS TEST DATA

To illustrate the model in section 2 and the estimation methods of section 3 I analyze business test data from the IFO Institute, Munich. Four waves obtained in August and November 1987 and February and May 1988 from 656 German firms are considered. The variables and codes used in the analysis are given in table 1.

Table 1: Questions and variables from the IFO business test.

Question	measurement level	variable in model	variable name
Our domestic production activity concerning XY has been with regard to the month before brisker(3), unchanged(2), weaker(1).	ordinal	$\Delta y_t = y_t - y_{t-1}$	output O
At the moment our stock of unsold finished products of XY corresponds, to no stock, less than 0.5/4, 1/4,...,6/4, more than 6/4, or ... months of production.	metric	$\ell n \frac{f_t}{f_{t-1}}$	stock of finished products LFP
At the moment our stock of raw materials for the production of XY corresponds to, no stock, less than 0.5/4, 1/4,...,6/4, more than 6/4, or ... months of production.	metric	$\ell n \frac{r_t}{r_{t-1}}$	stock of raw material LRP
At the moment we feel that our stock of orders of XY is relatively big (*e.g.* extended delivery period) (3), stayed the same(2), be rather unfavourable(1).	ordinal	$a_t - a_t^{*}$	order stock AB
Under elimination of mere seasonal fluctuations our business situation for XY during the coming 6 months will be rather favourable(3), stay more or less the same(2), be rather unfavourable(1).	ordinal	$d_{t,t+1} - d_t$	business expectation GL
With regard to the month before for us the demand situation (at home and abroad) has improved(3), stayed the same(2), decreased(1).	ordinal	$\Delta d_t = d_t - d_{t-1}$	demand at t D
Number of employees in the business (grouped).	metric	ℓn k	LNE
Production activity at time $t-1$ with 1 if $\Delta y_{t-1} = 1$ and 0 otherwise.	dummy	u_t	output OL1
Production activity at time $t-1$ with 1 if $\Delta y_{t-1} = 3$ and 0 otherwise.	dummy	v_t	output OL2

The variables are indiced by the number of the panel wave, hence O_0, LFP_0, LRP_0,... refers to the variables at time 0, O_1, LFP_1, LRP_1,... refers to the variables at time 1. The variables O (output), AB (order stack), GL (business expectation) and D (demand) are only measured as ordered categorical variables, the variables LFP, LRP, LNE are metric, OL1 and OL2 are dummy variables.

The model that is specified now is motivated in Arminger and Ronning (1991). The process described by the model is conditional on wave 0, hence the process is not modelled for this wave.

$$\Delta y_t = \mu_t + \beta_{t1}(a_{t-1} - a^*_{t-1}) + \beta_{t2}(d_{t,t+1} - d_t) + \beta_{t3}s_t + \gamma_{t1}(\ell\mathrm{n}k) + \gamma_{t2}\left[\ell\mathrm{n}\frac{r_t}{r_{t-1}}\right]$$

$$+\gamma_{t3}\left[\ell\mathrm{n}\frac{f_t}{f_{t-1}}\right] + \gamma_{t4}\mu_t + \gamma_{t5}v_t + \alpha + \epsilon_t, t = 1,2,3. \tag{26}$$

This model describes the change in output at time t as a function of the stock of orders at time $t-1$, business expectation at time t, the number of employees, the relative change of stock in raw material from $t-1$ to t, the relative change in the stock of finished products from $t-1$ to t and the states of the output variable at time $t-1$. The variable ln k, that is the number of employees does not change over time.

The variable s_t is defined as

$$s_t = [(d_t - d_{t-1}) - (d_{t-1,t} - d_{t-1})], \tag{27}$$

which may be interpreted as a shock or surprise effect. If $(d_t - d_{t-1}) > (d_{t-1,t} - d_{t-1})$ then the shock is positive, that is the demand is higher than the past expectation, otherwise the shock is zero or negative. The effect ofs_t may be estimated by setting the parameter of the demand $(d_t - d_{t-1})$ to β_{t3} and the parameter of the past business expectation $(d_{t,t-1} - d_{t-1})$ to $-\beta_{t3}$.

Note that the variables Δy_t, $(a_{t-1} - a^*_{t-1})$, $(d_{t,t-1} - d_t)$ and $(d_t - d_{t-1})$ are only observed at an ordinal scale where the following observation rule is supposed to be hold:

$$O_t = \begin{cases} 1 \text{ if } \Delta y_t \le \tau_1^{(1)} \\ 2 \text{ if } \tau_1 < \Delta y_t \le \tau_2^{(1)} \\ 3 \text{ if } \Delta y_t > \tau_2^{(1)} \ . \end{cases} \tag{28}$$

The observation rules for AB$_t$, GL$_t$ and D$_t$ are analogous. For identification, the first threshold is set to 0.

The random variable $\alpha \sim N(0,\sigma_\alpha^2)$ captures the unobserved heterogeneity, $\epsilon_t \sim N(0,\sigma_t^2)$ is assumed to be serially uncorrelated. Since AB$_t$, GL$_t$ and D$_t$ are only observed at an ordinal scale we assume that $(a_{t-1} - a^*_{t-1})$, $(d_{t,t-1} - d_t)$ and $(d_t - d_{t-1})$ are endogenous variables with means conditioned by the exogenous variables ℓn k, $\left(\ell\mathrm{n}\frac{r_t}{r_{t-1}}\right)$, $\left(\ell\mathrm{n}\frac{f_t}{f_{t-1}}\right)$, μ_t and v_t and a multivariate normal error vector that is uncorrelated with α and ϵ_t. The model for all endogenous variables in the first wave is therefore given by

$$\begin{bmatrix} a_0 - a_0^* \\ d_{2,1} - d_1 \\ d_1 - d_0 \\ d_{1,0} - d_0 \\ y_1 - y_0 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \beta_{51} & \beta_{52} & \beta_{53} & -\beta_{53} & 0 \end{bmatrix} \begin{bmatrix} a_0 - a_0^* \\ d_{2,1} - d_1 \\ d_1 - d_0 \\ d_{1,0} - d_0 \\ y_1 - y_0 \end{bmatrix}$$

$$+ \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \gamma_{24} & \gamma_{25} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \gamma_{34} & \gamma_{35} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} & \gamma_{44} & \gamma_{45} \\ \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & \gamma_{55} \end{bmatrix} \begin{bmatrix} \ell n\, k \\ \ell n\, \frac{r_1}{r_0} \\ \ell n\, \frac{f_1}{f_0} \\ u_1 \\ v_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \alpha \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ \epsilon_5 \end{bmatrix} . \qquad (29)$$

The models for wave 2 and wave 3 are constructed in similar ways. Note that the variable GL occurs two times in the first wave as GL_0 and GL_1. In the second wave GL_1 must be taken from the first wave. Hence the whole model consists of 13 equations. Our interest however focuses only on equations 5, 9 and 13 that is $y_1 - y_0$, $y_2 - y_1$ and $y_3 - y_2$. The first model does not take into account restrictions by proportionality of coefficients for each wave and unobserved heterogeneity. The parameter estimates obtained from MECOSA are shown in table 2.

Table 2: Unrestricted parameter estimates for IFO output model (t - values in parenthesis).

explanatory variables	wave 1	wave 2	wave 3
τ_1	0	0	0
τ_2	2.175 (28.989)	2.175 (28.989)	2.175 (28.989)
μ	0.604 (1.613)	1.147 (3.281)	0.599 (2.793)
$(a_{t-1} - a_{t-1}^*)$	0.439 (5.685)	0.305 (4.161)	0.293 (6.236)
$d_{t-1,t} - d_t$	0.053 (1.046)	0.089 (2.319)	0.125 (2.822)
s_t	0.413 (10.993)	0.312 (9.844)	0.385 (13.532)
$\ell n\, k$	-0.049 (-1.433)	-0.023 (-0.707)	0.015 (0.468)
$\ell n\, \frac{r_t}{r_{t-1}}$	0.097 (0.679)	-0.077 (-0.481)	0.086 (0.632)
$\ell n\, \frac{f_t}{f_{t-1}}$	0.238 (2.973)	-0.031 (-0.369)	-0.058 (-0.796)
u_t	-0.543 (-3.040)	-0.834 (-5.916)	-0.227 (-2.386)
v_t	0.129 (1.039)	-0.204 (-1.697)	0.429 (3.217)
R^2_{MZ}	0.089	0.184	0.116
covariances			
wave 1	0.514		
wave 2	0.127	0.715	
wave 3	0.045	-0.010	0.616

The pseudo R^2s of McKelvey and Zavoina (1975) show that only a small portion of the variance of the output is explained. The output increases in the second wave in comparison to the first and third wave. Judged by the

t -values, the variables stock order (AB) and surprise effect are more important than business expectation (GL). The firms react primarily to shocks of the recent past. If the shock has been positive, more output is produced. The variables stock of raw materials and stock of unsold finished products are of lesser importance than the dependence on the state of the period before. Here, however, only the decrease of output in the past period matters. The covariances between the errors are rather small indicating that unobserved heterogeneity may be unimportant.

In the next table, the results of the restricted parameter estimation under the hypothesis of proportionality of the regression coefficients except for the constants and the effect of the number of employees are shown.

Table 3: Restricted parameter estimates for IFO output model (t - values in parenthesis).

explanatory variables	wave 1	wave 2	wave 3
τ_1	0	0	0
τ_2	2.135 (30.379)	2.135 (30.379)	2.135 (30.379)
μ	0.842 (2.497)	1.203 (3.025)	0.736 (3.106)
$(a_{t-1} - a^*_{t-1})$	0.372 (8.547)	0.372 (8.547)	0.372 (8.547)
$d_{t+1,t} - d_t$	0.103 (3.539)	0.103 (3.539)	0.103 (3.539)
s_t	0.417 (13.305)	0.417 (13.305)	0.417 (13.305)
$\ell n\, k$	-0.064 (-1.976)	-0.032 (-0.775)	0.008 (0.230)
$\ell n \frac{r_t}{r_{t-1}}$	0.122 (1.472)	0.122 (1.472)	0.122 (1.472)
$\ell n \frac{f_t}{f_{t-1}}$	0.029 (0.634)	0.029 (0.634)	0.029 (0.634)
u_t	-0.548 (-7.380)	-0.548 (-7.380)	-0.548 (-7.380)
v_t	0.070 (0.959)	0.070 (0.959)	0.070 (0.959)
λ		0.749	0.914
σ^2_α	0.055 (1.412)		
χ^2 statistic	3.841	df	14

The hypothesis of proportionality is not rejected at the 0.05 test level. The error variance of the second wave is greater in absolute terms than the error variances of the first and third wave as judged from the inverse of the proportionality coefficient λ. The variance of heterogeneity turns out to be not significant at the 0.05 test level.

REFERENCES

Andersen, E.B. (1973). *Conditional Inference and Models for Measuring*, Copenhagen: Mentalhygiejnisk Forsknings Institut.

Arminger, G. (1987). Misspecification, asymptotic stability and ordinal measurements in models for the analysis of panel data. *Sociological Methods and Research*, 15, 3, 336-348.

Arminger, G., and Ronning, G. (1991). Ein Strukturmodell für Preis-, Produktions- und Lagerhaltungsentscheidungen von Firmen. *IFO-STUDIEN, Zeitschrift für empirische Wirtschaftsforschung*, 37, 229-254.

Arminger, G. (1992). Analyzing panel data with non-metric dependent variables: Probit models, generalized estimating equations, missing data and absorbing states. *Discussion Paper No. 59*, Deutsches Institut für Wirtschaftsforschung, Berlin.

Hamerle, A., and Ronning, G. (1993). Analysis of discrete panel data, forthcoming in G. Arminger, C.C. Clogg and M.E. Sobel (Eds.). *Handbook of Statistical Modeling for the Behavioral Sciences*, New York: Plenum.

Heckman, J.J. (1981a). Statistical models for discrete panel data, in C.F. Manski and D. McFadden (Eds.). *Structural Analysis of Discrete Data with Econometric Applications*, 114-178.

Heckman, J.J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete stochastic process, in C.F. Manski and D. McFadden (Eds.). *Structural Analysis of Discrete Data with Econometric Applications*, 179-195.

Hsiao, C. (1986). *Analysis of Panel Data*, Cambridge, Massachusetts: Cambridge University Press.

Keane, M.P., and Runkle, D.E. (1992). On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. *Journal of Business & Economic Statistics*, 10, 1, 1-29.

Kiefer, J., and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887-906.

Küsters, U. (1987). *Hierarchische Mittelwert- und Kovarianzstrukturmodelle mit nichtmetrischen endogenen Variablen*, Heidelberg: Physica Verlag.

Maddala, G.S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources*, XXII, 3, 307-336.

McKelvey, R.D., and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.

Neyman, J., and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1-32.

Rosett, R.N., and Nelson, F.D. (1975). Estimation of the two-limit probit regression model. *Econometrica*, 43, 141-146.

Schepers, A., Arminger, G., and Küsters, U. (1991). The analysis of non-metric endogenous variables in latent variable models: the MECOSA Approach, in P. Gruber (Ed.). *Econometric Decision Models: New Methods of Modeling and Applications*, Springer Verlag, Heidelberg, 1991, 459-472.

Schepers, A., and Arminger, G. (1992). *MECOSA: A Program for the Analysis of General Mean- and Covariance Structures with Non-Metric Variables, User Guide*, SLI-AG, Züricher Str. 300, CH-8500 Frauenfeld, Switzerland.

Sobel, M., and Arminger, G. (1992). Modeling household fertility decisions: A nonlinear simultaneous probit model. *Journal of the American Statistical Association*, 87, 38-47.

Stewart, M.B. (1983). On least squares estimation when the dependent variable is grouped. *Review of Economic Studies L*, 737-753.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

LOGISTIC MODELLING OF LONGITUDINAL SURVEY DATA WITH MEASUREMENT ERROR

C.J. Skinner[1]

ABSTRACT

A logistic model relating a binary response at one occasion to the response at the previous occasion and other covariates is considered. Measurement error on the binary response can lead to biased estimates. Estimation procedures which adjust for measurement error are suggested for different measurement models. The procedures are illustrated using data from the U.S. Panel Study of Income Dynamics, where the response is whether an individual is in a job with a union contract.

KEY WORDS: Gross flow; Transition; Longitudinal; Measurement error.

1. INTRODUCTION

In the analysis of longitudinal survey data it is often of interest to estimate a *transition rate*, the proportion of units in the population in one state on one occasion which flow into another state on the successive occasion. For example, labour market analysts may be interested in the 3 x 3 matrix of transitions between the states: employed, unemployed, and not in the labour force. For analytical purposes, it is often of interest to study how rates vary across different subgroups of the population. For example, in the case of labour force states it may be of interest to study the dependence of transition rates on sex, age and region.

Transition rates are proportions and so may be estimated in the usual way from survey data. The off-diagonal cells of transition matrices are often relatively rare, however, and so, as the number of subgroups increases, the sample sizes upon which some estimates are based can become small implying large sampling errors. In this case some modelling of the relation between the rates and the covariates defining the subgroups is desirable. In Section 2, we describe a logistic model for representing the dependence of transition probabilities on covariates for the case of two states and two occasions. Similar models have been applied to longitudinal data in both biostatistical applications (*e.g.*, Korn and Whittemore 1979; Muenz and Rubinstein 1985) and econometric applications (*e.g.*, Hsiao 1986, Sect. 7.4; Maddala 1987).

A major problem, when estimating gross flows from survey data, is that of measurement error. Random errors in measured states can lead to severe upward bias in standard estimators of proportions moving between different states. A number of alternative estimators have been proposed which use reinterview data to reduce this bias (Meyer 1988). The aim of this paper is to extend this work to the estimation of the logistic models referred to above.

[1] C.J. Skinner, Department of Social Statistics, University of Southampton, Highfield, Southampton S09 5NH, U.K.

2. THE MODEL

Consider a finite population of size N, which is fixed over the two occasions $t = 1, 2$ and is partitioned into$I$ cells of sizes $N_1, ..., N_I$ ($\sum N_i = N$) by the levels of one or more factors defined at $t = 1$. Let y_t be a binary indicator variable for the two states and let N_{ijk} be the number of units in cell i for which $y_i = j$ and $y_2 = k$ $(i = 1, ..., I; j = 0, 1; k = 0, 1)$. Let $N_{ij.} = N_{ij0} + N_{ij1}$ and note that

$$\sum_{j=0}^{1} \sum_{k=0}^{1} N_{ijk} = \sum_{j=0}^{1} N_{ij.} = N_i, \; i = 1, ..., I .$$

Suppose that the finite population values are generated by a model in such a way that the N_{ijk} are multinomially distributed with parameters N and ϕ_{ijk} $(i = 1, ..., I; j = 0, 1; k = 0, 1)$, so that in particular $E(N_{ijk}) = N\phi_{ijk}$. Let

$$\pi_{ij} = \phi_{ij1} / \phi_{ij.} ,$$

where

$$\phi_{ij.} = \phi_{ij0} + \phi_{ij1}, \; i = 1, ..., I; \; j = 0, 1.$$

Then π_{ij} represents the model transition probability or gross flow rate in cell i between state j at $t = 1$ and state 1 at $t = 2$. We take our objective to be to study the dependence of π_{ij} on i and j and consider the following logistic model for π_{ij}:

$$\pi_{ij} = F(x_{ij}\beta), \; i = 1, ..., I; \; j = 0, 1, \tag{1}$$

where

$$F(t) = e^t / (1 + e^t),$$

the x_{ij} are $1 \times s$ vectors of known constants and β is a $s \times 1$ vector of unknown parameters. Note that (1) may be expressed alternatively as

$$\log[\pi_{ij} / (1 - \pi_{ij})] = x_{ij}\beta . \tag{2}$$

Some special cases of this model are given below to illustrate its interpretation, but first we record some notation. For a series of $1 \times k$ vectors, a_{ij} $(i = 1, ..., I; j = 0, 1)$, let $[a_{ij}]$ denote the $2I \times k$ matrix with rows $a_{10}, a_{11}, a_{20}, a_{21}, ..., a_{I0}, a_{I1}$. Let $X = [x_{ij}]$, $\ell = [\ell_{ij}]$, $\pi = [\pi_{ij}]$, $\phi = [\phi_{ij.}]$, $f(\beta) = [f_{ij}(\beta)]$, where $\ell_{ij} = \log[\pi_{ij} / (1 - \pi_{ij})]$, $f_{ij}(\beta) = F(x_{ij}\beta)$.

Then (1) may be re-expressed as

$$\pi = f(\beta) \tag{3}$$

and (2) may be re-expressed as

$$\ell = X\beta . \tag{4}$$

Examples of Models

(i) **Constant transition rates**

Let $s = 2$, $x_{ij} = (1 j)$ and $\beta = (\beta_1 \beta_2)'$. Then $\pi_{i0} = F(\beta_1)$ and $\pi_{i1} = F(\beta_1 + \beta_2)$ for all i.

(ii) **Additive model**

Let $s = r + 2$, $x_{ij} = (1 z_i j)$ and $\beta = (\beta_1 \beta_2' \beta_3)'$, where z_i is a $1 \times r$ vector of known constants derived from the factor levels defining the I cells and β_2 is an $r \times 1$ vector of unknown parameters. For example, the cells may arise from crossing I/2 age groups by 2 sexes and z_i may be $(a_i a_i^2 s_i)$ where a_i is the midpoint of the age group and s_i is a dummy variable representing sex for cell i. The ratio of the odds that $y_2 = 1$ given $y_1 = 1$ versus $y_1 = 0$ is $\exp(\beta_3)$ which is constant across cells.

(iii) **Separate models for different previous states**

Let $s = 2r + 2$, $x_{ij} = (1\ z_i\ j\ jz_i)$ and $\beta = (\beta_1\, \beta_2'\, \beta_3\, \beta_4')'$, where z_i is as in (ii). Unlike in (ii), this model allows for interaction between y_1 and cell i. Transition rates are now $\pi_{i0} = F(\beta_1 + z_i\beta_2)$ and $\pi_{i1} = F[(\beta_1 + \beta_3) + z_i(\beta_2 + \beta_4)]$.

(iv) **Saturated model**

Let $s = 2I$ and suppose X is nonsingular. Then there is a 1-1 mapping between π and β since (4) may be inverted to give $\beta = X^{-1}\ell$.

In general, we take β to be the vector of parameters of interest. As set out above, β is only well defined if model (1) holds. In practice, however, it may still be of interest to fit a model, such as the main effects model in (ii), even if that model only holds approximately. Under just the multinomial assumption and some specification of x_{ij}, but not necessarily the validity of (1), we define β to be the solution to

$$\sum_i \sum_j x_{ij}\, \phi_{ij.}\, (f_{ij}(\beta) - \pi_{ij}) = 0 . \tag{5}$$

3. ESTIMATION

Let $\hat{N}_{ijk}$ be an estimator of N_{ijk} which may involve weighting or other survey adjustments. Consider an asymptotic framework in which N and the sample size n increase but where I and the ϕ_{ijk} are fixed. Let

$$\hat{N}_{ij.} = \sum_k \hat{N}_{ijk},\ w_{ij.} = \hat{N}_{ij} / \hat{N},\ p_{ij} = \hat{N}_{ij1} / N_{ij},\ \hat{N} = \sum_i \sum_j \hat{N}_{ij.},$$
$$w = [w_{ij.}],\ p = [p_{ij}].$$

We then assume $(p'\, w')'$ is consistent for $(\pi'\, \phi')'$ and the asymptotic distribution as $n \to \infty$ of $\sqrt{n}\, [(p'\, w')' - (\pi'\, \phi')']$ is normal with mean vector zero and covariance matrix

$$V_{(p,w)} = \begin{bmatrix} V_p & C_{pw} \\ C_{pw}' & V_w \end{bmatrix} . \tag{6}$$

Given w and p, β may be estimated by the solution $\hat{\beta}$ of equations (5) with $\phi_{ij.}$ and π_{ij} replaced by $w_{ij.}$ and p_{ij} respectively.

We now define some further notation. For a series of scalars a_{ij} $(i = 1,...,I;\ j = 0,1)$ let diag$[a_{ij}]$ denote the $2I \times 2I$ diagonal matrix with diagonal elements $a_{10}, a_{11}, ..., a_{I0}, a_{I1}$. Let

$$D(w) = \text{diag}[w_{ij.}],\ D(\phi) = \text{diag}\,[\phi_{ij.}],$$
$$D(\epsilon) = \text{diag}[\epsilon_{ij}],\ \Delta = \text{diag}[\phi_{ij.} f_{ij}(\beta)\{1 - f_{ij}(\beta)\}],$$

where $\epsilon_{ij} = f_{ij}(\beta) - \pi_{ij}$ is the model approximation error for cell i and $y_1 = j$. Then $\hat{\beta}$ solves the estimating equations

$$X' D(w) f(\hat{\beta}) = X' D(w) p , \tag{7}$$

(*c.f.* Roberts *et al*. 1987, equation 2.3). The asymptotic covariance matrix of $\hat{\beta}$ is given by:

$$V(\hat{\beta}) = n^{-1} (X'\Delta X)^{-1}\, X' \sum\, X (X'\Delta X)^{-1} , \tag{8}$$

where

$$\sum = [D(\phi)\ D(\epsilon)]\ V_{(p,w)}\ [D(\phi)\ D(\epsilon)]' . \tag{9}$$

If the logistic model holds then $D(\epsilon) = 0$, $\sum$ reduces to $D(\phi)\, V_p D(\phi)$ and $V(\hat{\beta})$ reduces to an expression analogous to equation (2.4) of Roberts *et al.* (1984).

Given estimators of V_p, V_w and C_{pw}, $V(\hat{\beta})$ may be estimated by substituting w, p, and $\hat{\beta}$ for ϕ, π, and β respectively in Δ and $\sum$.

Example, Simple Random Sampling

Suppose n units are selected from the population by simple random sampling, and $n_{ij.}$ and n_{ij1}, the sample quantities analogous to $N_{ij.}$ and N_{ij1}, are observed $(i = 1, ..., I;\ j = 0, 1)$. Then the maximum likelihood equations are, by analogy to (5):

$$\sum_i \sum_j x_{ij} [n_{ij.} f_{ij}(\beta) - n_{ij1}] = 0 ,$$

which corresponds to taking $w_{ij.} = n_{ij.}/n, p_{ij} = n_{ij1}/n_{ij.}$ in (7). Then p and w are asymptotically uncorrelated $(C_{wp} = 0)$,

$$V_p = \mathrm{diag}[\pi_{ij}(1 - \pi_{ij})\phi_{ij.}^{-1}] ,\quad V_w = D(\phi) - \phi\phi'$$

and from (9),

$$\sum = \mathrm{diag}[\pi_{ij}(1 - \pi_{ij})\phi_{ij.}] + D(\epsilon)[D(\phi) - \phi\phi']D(\epsilon) .$$

If the logistic model holds, then $\sum = \Delta$ and $V(\hat{\beta}) = n^{-1}(X'\Delta X)^{-1}$, the standard formula.

4. THE EFFECT OF MEASUREMENT ERROR

Now suppose we do not observe $\hat{N}_{ijk}$, but only $\hat{N}_{ijk}^{*}$, an estimator of N_{ijk}^{*}, the number of units in cell i for which $y_1^{*} = j$ and $y_2^{*} = k$, where y_1^{*} and y_2^{*} are versions of y_1 and y_2 respectively, measured with error. We suppose that the N_{ijk}^{*} are multinomially distributed with parameters N and ϕ_{ijk}^{*} and define $\hat{N}_{ij}^{*}, \phi_{ij}^{*}, w_{ij}^{*}, p_{ij}^{*}, \pi_{ij}^{*}, \phi^{*}, w^{*}, p^{*}, \pi^{*}, D(w^{*})$, and $D(\phi^{*})$, analogously to their non-asterisked versions.

Let $\hat{\beta}^{*}$ be the solution of estimating equations (7) with w and p replaced by w^{*} and p^{*} respectively. Then, assuming w^{*} and p^{*} are consistent for ϕ^{*} and π^{*} respectively, $\hat{\beta}^{*}$ will be consistent for the solution β^{*} of the equations:

$$X'D(\phi^{*})f(\beta^{*}) = X'D(\phi^{*})\pi^{*} . \tag{10}$$

In general, β^{*} will not equal β unless $\phi^{*} = \phi$ and $\pi^{*} = \pi$. Hence, measurement error induces bias even in large samples.

5. ADJUSTMENT FOR MEASUREMENT ERROR

The nature of the measurement error adjustment will depend on the specification of the measurement error model which will in turn depend, in practice, on the nature and extent of the validation data available. We suppose first that there is no misclassification of the cells so that $\phi_{i..}^{*} = \phi_{i..}$, where $\phi_{i..} = \phi_{i0.} + \phi_{i1.}, \phi_{i..}^{*} = \phi_{i0.}^{*} + \phi_{i1.}^{*}$.

We suppose next that only cross-sectional validation data is available in which case it is natural, following *e.g.* Abowd and Zellner (1985), to assume *independent measurement errors within cells*:

$$\phi^*_{ijk} = \sum_{\ell=0}^{1} \sum_{m=0}^{1} \theta^1_{ij\ell}\, \theta^2_{ikm}\, \phi_{i\ell m}, \tag{11}$$

where $\theta^t_{ijk} = pr(y_t^* = j \mid y_t = k,\ \text{cell } i)$ is the probability of misclassifying state k as state j in cell i at time t. Without longitudinal validation data it is difficult to know how to specify a model for dependent errors although some sensitivity analysis to departures from independence is possible (Rao and Singh 1991).

Letting $\phi^t(i)$, $\phi(i)$ and $\phi^*(i)$ denote the 2×2 matrices with jk^{th} elements ϕ^t_{ijk}, ϕ_{ijk} and ϕ^*_{ijk} respectively, (11) may be reexpressed as

$$\phi^*(i) = \theta^1(i)\ \phi(i)\ \theta^2(i)'\ .$$

If estimators $\hat{\theta}^t(i)$ of the $\theta^t(i)$ are available from validation studies then an adjusted estimator of $\phi(i)$ is

$$\hat{\phi}(i) = \hat{\theta}^1(i)^{-1}\ \hat{\phi}^*(i)\ [\hat{\theta}^2(i)']^{-1}, \tag{12}$$

where the jk^{th} element of $\hat{\phi}^*(i)$ is $\hat{N}^*_{ijk} / \hat{N}^*$. An adjusted estimator of β is then obtained by solving (5) with $\phi_{ij.}$ and π_{ij} replaced by $\hat{\phi}_{ij}$ and $\hat{\pi}_{ij} = \hat{\phi}_{ij1} / \hat{\phi}_{ij.}$ respectively. Assuming consistency of $\hat{\phi}^*_{ijk}$ and $\hat{\theta}^t_{ijk}$ for ϕ^*_{ijk} and θ^t_{ijk} respectively, the adjusted estimator will be consistent and its asymptotic covariance matrix is as defined by (8) and (9) where $V_{(p,w)}$ is replaced the asymptotic covariance matrix of the vector of $\hat{\pi}_{ij}$ and $\hat{\phi}_{ij}$. This matrix can be estimated by the δ-method provided estimates of the covariance matrix of the $\hat{\theta}^t_{ijk}$ are available.

A problem with this approach is that the values of $\hat{\phi}_{ijk}$ implied by (12) may fall outside the interval [0,1], a situation which may often arise since the $\hat{\phi}^*(i)$ are likely to display appreciable sampling variability, given that this is the reason that a logistic model is being used in the first place. This suggests either imposing a constrained inference procedure or considering a narrower measurement error model specification. One such more restricted assumption following Chua and Fuller (1987) is to assume *unbiased measurement errors*:

$$\phi^*_{ij.} = \phi_{ij.}\,,\ \phi^*_{i.k} = \phi_{i.k} \quad i=1,\ ..,\ I,\ j=0,1,\ k=0,1\,.$$

Let

$$\alpha^t_i = pr(y_t^* = 1 \mid y_t = 0,\ \ \text{cell } i) / pr(y_t^* = 1 \mid \text{cell } i) \tag{13}$$

measure the 'amount' of measurement error at time t in cell i. A consequence of the unbiasedness condition is that the right-hand side of (13) is unchanged if 1 is replaced by 0 and vice versa. Letting $\pi^*_{ij} = \phi^*_{ij1} / \phi^*_{ij.}$ it follows from the assumption that the measurement errors are both independent and unbiased that

$$\pi^*_{ij} = (1-\alpha^1_i)\ (1-\alpha^2_i)\,\pi_{ij} + [1-(1-\alpha^1_i)\ (1-\alpha^2_i)]\ \phi_{i.1}/\phi_{i..}\,.$$

Given an estimator $\tilde{\gamma}_i$ of $\gamma_i = [(1-\alpha^1_i)(1-\alpha^2_i)]^{-1}$ we may estimate π_{ij} by

$$\tilde{\pi}_{ij} = \tilde{\gamma}_i p^*_{ij} - (\tilde{\gamma}_i - 1) p_i^*\,, \tag{14}$$

where $p^*_{ij} = \hat{\phi}_{ij1} / \hat{\phi}^*_{ij.}$ and $p_i^* = (w^*_{i0}\, p^*_{i0} + w^*_{i1}\, p^*_{i1}) / (w^*_{i0} + w^*_{i1})$.

An adjusted estimator of β is then obtained by solving (5) with ϕ_{ijl} and π_{ij} replaced by $\phi^*_{ij.}$ and $\tilde{\pi}_{ij}$ respectively, that is

$$X'D(w^*)f(\tilde{\beta}) = X'D(w^*)\tilde{\pi},$$

where $\tilde{\pi} = [\tilde{\pi}_{ij}]$. Letting $\gamma = (\gamma_1, ..., \gamma_I)'$, $\tilde{\gamma} = (\tilde{\gamma}_1, ..., \tilde{\gamma}_I)'$, we assume $n^{½}(\tilde{\gamma}-\gamma) \rightarrow N[0, V_\gamma]$ in law as $n \rightarrow \infty$ and $\tilde{\gamma}$ is asymptotically independent of $(p^{*\prime}, w^{*\prime})$. Since w^* and $\tilde{\pi}$ are consistent for $\phi^* = \phi$ and π respectively, $\tilde{\beta}$ is consistent for β. The asymptotic covariance matrix of $\tilde{\beta}$, is given by expression (8) with $V_{(p,w)}$ in (9) replaced by the (normalized) asymptotic covariance matrix of $(\tilde{\pi}', w^{*\prime})'$. Given consistent estimators of $V(p^*)$, $V(w^*)$, and $V(\tilde{\gamma})$, a consistent estimator of $V(\tilde{\beta})$ may be obtained as before.

6. AN EXAMPLE

The example is based on data from the U.S. Panel Study of Income Dynamics (PSID). Table 1 cross-classifies the variable:

y_t^* = 1 if individual is in job covered by union contract
= 0 otherwise,

for the two years t=1 (1983) and t=2 (1987), based on men in the self-weighting 'Survey Research Centre sample' (Hill 1992, p.9) who are currently working in both years but are not self-employed nor working for government.

Table 1: Sample Counts for Observed Variables.

		y_2^*	
		0	1
y_1^*	0	684	33
	1	43	191

Two factors which might be expected to affect transitions between the states are considered. The first factor is age, which is divided into four categories, 18-29, 30-34, 35-44, 45+, which are of roughly equal size for the sample considered. The second factor partitions employment sectors into two categories, roughly according to tendency to be unionized. The first less-unionized category includes professional, managerial, sales and farming employment. The second more-unionized category includes manual and clerical employment. These 2 factors together define I=8 cells.

For simplicity, we ignore here the complexity of the sampling design and assume simple random sampling. Fitting alternative logistic models with y_2^* as the response and examining the likelihood-ratio chi-squared statistics suggests a model with

$$x_{ij} = (1 \;\; j \;\; \text{age}(2) \;\; \text{age}(3) \;\; \text{age}(4) \;\; \text{work} \;\; j.\text{age}(2) \;\; j.\text{age}(3) \;\; j.\text{age}(4)),$$

where j is the value of y_1^*, age (2) - age (4) are binary indicators representing the age factor and 'work' is a binary indicator of the second factor. Thus the model includes an interaction between age and y_1^*, which reflects the fact that as age increases there is declining mobility either from $y_1^* = 0$ to $y_2^* = 1$ or from $y_1^* = 1$ to $y_2^* = 0$. On the other hand, there seems little evidence of an interaction between y_1^* and the second factor or between the two factors. Parameter estimates and standard errors are given in the first column of Table 2.

Table 2: Parameter Estimates for Logistic Model.

	Measurement error ignored			Adjusted for measurement error	
Covariate	Estimated Coefficient	s.e. model-based	robust	Estimated Coefficient	s.e.
Constant	-2.81	0.34	0.33	-2.75	0.33
y_1	3.13	0.44	0.44	3.61	0.48
age(2)	-0.69	0.47	0.47	-1.11	0.49
age(3)	-1.02	0.53	0.53	-1.74	0.54
age(4)	-0.93	0.53	0.53	-2.26	0.55
y_1.age(2)	0.80	0.65	0.65	1.19	0.71
y_1.age(3)	1.92	0.75	0.74	2.74	0.80
y_1.age(4)	2.54	0.76	0.76	4.76	0.84
work	0.63	0.30	0.29	0.32	0.29

One source of information on measurement error is the PSID Validation Study (Hill 1992, p.29). This study involved comparing responses to the PSID instrument with company records for a sample of workers from one large firm. A cross-classification of validated and survey responses on the response variable in 1987 is given in Table 3.

Table 3: Validated by Survey Responses from Validation Study.

		Survey y_2^*	
		0	1
Validated y_2	0	140	8
	1	2	302

Assuming that the misclassification matrices in the validation study and the general population are the same and that errors are independent and identically distributed over time, observed counts in Table 1 are adjusted according to the approach in (12) to the first table in Table 4.

Table 4: Adjusted Counts under Alternative Measurement Models.

Common Misclassification Matrices

		y_2	
		0	1
y_1	0	764	- 8
	1	3	192

Unbiased Errors common α

	y_2	
	0	1
	690	27
	37	197

Under this measurement model it appears that essentially all the observed transitions can be explained by measurement error and hence there is no purpose in continuing to fit a logistic model. As an alternative measurement model suppose now that measurement errors are now not only independent and identically distributed over time but they are also unbiased in both the general population and the Validation Study. The estimated value of α in (13) for the validation study is $\hat{\alpha} = 0.02$ and assuming that this (rather than the entire misclassification matrices) is the same in the Validation Study and the general population and is the same over time the adjusted count matrix, following the approach in (14), is the second table in Table 4. This adjustment is quite different to the first and implies only a moderate adjustment of Table 1. The reason for the difference is that the marginal distribution of y_2^* is very different in the Validation Study and the general population.

Hence assuming unbiased measurement errors with common α in both populations implies very different misclassification matrices.

Extending the adjustment under the unbiased error model to the logistic model following the approach in Section 5 and assuming a common α over all cells gives the adjusted estimates in Table 2. Note that even though the adjustment appears small with $\hat{\alpha}$ only equal to 0.02, the effect on the coefficients of age and of the interaction between age and y_1 are very marked. The adjusted standard errors allow for the error in estimating α and it is reassuring that these are not much larger than the original standard errors.

ACKNOWLEDGEMENTS

Research was supported by grant number R000 23 2522 from the Economic and Social Research Council. The author is grateful to Keith Humphreys for carrying out the computational work in Section 6. The data were made available through the ESRC Data Archive by the Inter-University Consortium for Political and Social Research, Ann Arbor, Michigan. The data were originally collected by J.N. Morgan. Neither the original investigator, nor the Consortium, nor the Archive bear any responsibility for the analysis presented here.

REFERENCES

Abowd, H.M., and Zellner, A. (1985). Estimating gross flows. *Journal of Business and Economic Statistics*, 3, 254-283.

Chua, T.C., and Fuller, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, Sage.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Korn, E.L., and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics* 35, 795-802.

Maddala, G.S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources*, 22, 307-338.

Meyer, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics*, 6, 385-390.

Muenz, L.R., and Rubinstein, L.V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 41, 91-101.

Roberts, G., Rao, J.N.K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

Singh, A.C., and Rao, J.N.K. (1991). Classification error adjustments for gross flow estimates. Technical Report No. 183, Laboratory for Research in Statistics and Probability, Carleton University.

SESSION 10

Quality Issues

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

DATA QUALITY AND THE OPCS LONGITUDINAL STUDY

I. Macdonald Davies[1]

ABSTRACT

The OPCS Longitudinal Study involves no direct data collection. It is a record linkage study of many of the events routinely processed by OPCS, including the 1971, 1981 and 1991 Censuses. At any one time, it covers about 500,000 people in England and Wales - 1% of the population.

Linkage depends on information provided at most events - name, sex, date of birth. The paper considers the quality issues that this type of linkage raises, the challenges of dealing with a dynamic sample and the insight that the LS can provide into the quality of constituent data sources.

KEY WORDS: Longitudinal; Record linkage; Data quality.

1. INTRODUCTION

1.1 The Longitudinal Study

The Longitudinal Study (LS) is a record linkage study run by the Office of Population Censuses and Surveys (OPCS). At any one time, it includes about 500,000 people in England and Wales - 1 per cent of the population. Currently, it includes a sample of the census records from 1971 and from 1981 and of vital events (births to women in the sample, deaths, cancers, widowerhoods *etc.*) (OPCS 1989). Only information that is routinely collected for everyone by OPCS is included. There is no specific data collection nor are any data from other Government departments incorporated.

1.2 Membership of the sample

Membership of the LS is determined by day and month of birth. Everyone born on one of four fixed days in every year is a member - a sampling fraction of 4/365 or 1.1%. The original LS sample was taken from the 1971 Census. Throughout the intercensal period, the sample was updated by adding births occurring on the relevant dates and immigrants with these birth dates who registered with the National Health Service. Deaths and emigrations were removed. Another sample was taken of all those records in the 1981 Census with the appropriate birth dates and the records linked into the LS dataset. The sample continued to be updated during the 1980's. Once again, a sample was taken from the 1991 Census of all those with the appropriate birth dates. Currently, these records are being linked to the LS dataset. For each LS member, this record linkage can be regarded as giving a profile of their life as recorded by events known to OPCS.

1.3 The National Health Service Central Register

A key element in the successful running to the LS is the National Health Service Central Register (NHSCR). This Register is updated with births as they occur and with immigrants who register with a General Practitioner. Members of the population can be "flagged" in the Register and information on death, cancer registration or exit from risk (primarily emigration) can then be made available for research.

[1] I. Macdonald Davies, Office of Population Censuses and Surveys, Health Statistics, St. Catherine's House, 10 Kingsway, London WC2B 6JP.

1.4 Organization of linkage and holding of data

Each member of the LS has an LS number which is generated as he/she **joins** the LS. It is used to link the component records with the LS dataset. The number is put on the main Register at NHSCR which holds only the LS number, name, date of birth and sex. The statistical records forming the LS dataset are held separately and these records do not include name but do include LS number. For those LS members flagged on the Register, NHSCR provides the mechanism for adding the LS number to the appropriate sample generated from the routine processing of census and events records. This LS number is then used to link the statistical part of the census/event record into the LF dataset. This system is necessary because routine processing of census and events data does not involve entering people's names on the computer. NHSCR also provide details of entry to the LS via immigration and of exit from the LS (primarily emigrations). Finally, NHSCR records deaths and cancer registrations occurring among LS members flagged on the Register. These people are compared with those people who quote an LS data of birth at death/cancer registration in the main OPCS processing of death and cancer registration. In an ideal world, these two sets of records would be identical! However, some people inconsistently quote their date of birth. This cross check helps to minimize the loss of events occurring to the LS members.

2. DATA QUALITY

2.1 Introduction

The LS has the strength of confirming information using multiple sources, though this can introduce doubt where sources differ. The quality is dependent on the quality of the feeder sources but can be better than the original sources because additional checks can be built in. Other factors affecting quality are the appropriateness of the sample and the effectiveness of the internal linkage process within the LS. In this paper some indicators of quality are considered: tracing rates; sampling fractions; linkage rates and representativeness of those linked. These are described with some examples.

2.2 Constituent data sources

In England and Wales, there is a legal requirement to complete a Census return, and extensive efforts are made to ensure good coverage. Coverage is almost complete (OPCS 1983, OPCS 1988, OPCS 1992). Information on births and deaths is provided by the administrative system for registering births and deaths. Registration is a legal requirement. There are also practical reasons why registration is important (*e.g.*, a death certificate is needed before burial or cremation can take place). This helps to ensure that registration is effectively complete. Information on immigration and emigration is provided by NHSCR. This relies on people registering with a doctor (for immigration) and on the NHSCR being notified that they have left the country (emigration). The notification of emigration is particularly deficient, possibly around fifty per cent (OPCS 1988). The final routinely collected data within the LS is cancer registration. This is a voluntary system generated from hospital records and compiled nationally by OPCS. The coverage has improved since it started in 1971 but there are been and continue to be differences in the coverage between regions (Swerdlow 1986).

2.3 Tracing at NHSCR

Flagging the LS member's entry on the Register at NHSCR (tracing) is essential for high levels of linkage of subsequent information. Almost 98 per cent of the original 1971 Census sample are now traced at NHSCR. The majority of these were traced initially, but 5,400 were traced when additional information became available in the 1981 Census. All people entering the LS by birth or immigration are automatically flagged on the Register as the new births and immigrants form part of the main Register. For the 1981 Census sample, over 98 per cent of the sample were traced. Tracing rates differ for subgroups of the population. Fox and Goldblatt (1982) noted that women had slightly lower rates (explained by changing of name at marriage/divorce). The highest no trace rates were for people born in Pakistan and Hong Kong. For people born in these countries, name does not provide as useful an identifier as it does for people born in the United Kingdom. Data of birth may also be inconsistently reported.

2.4 Expected sampling fractions in the 1971 Census

The expected sampling fraction for the LS is 1.1%. Fox and Goldblatt (1982) reported on the actual 1971 Census sampling fractions. Taking only those traced at NHSCR, they found an actual sampling fraction of 1.06 for males and 1.05 for females. The variation in sampling fractions by age and marital condition was not greater than expected from a random sample of this size. Sampling fractions were low for people in non-private households. This group probably included a high proportion of hard to trace cases and a high proportion of people for whom no date of birth was recorded. Table 1 shows sampling fractions for people by country of birth and sex. In many cases observed sampling fractions are markedly different from those expected on the basis of overall sampling fractions and no trace rates. The high no trace rates for people born in the Asian New Commonwealth suggested that these groups would be under-represented in the final sample. However, they were over-represented and markedly so. This is because a relatively high proportion of people born in these countries recorded an LS data of birth on the 1971 Census form.

Table 1: 1971 Census LS sample observed and expected sampling fractions by sex and selected country of birth.

	Male		Female	
	Observed	Expected	Observed	Expected
England and Wales	1.06	1.07	1.05	1.06
Irish Republic	1.05	0.96	1.08	0.97
Old Commonwealth	0.98	0.98	0.98	0.98
Asian New Commonwealth*	1.37	0.88	1.21	0.88
Europe	1.16	0.99	1.15	0.97

* Includes Oceanian New Commonwealth.

2.5 Expected sampling fractions for events

Babb and Hattersley (1992) considered women LS members who were born in 1950 or later and whose babies were born after the 1971 Census. These births were linked into the LS via the recording of their mother's data of birth at birth registration. The quality of the data between 1971 and 1981 was improved by using the 1981 Census to find births "missed" originally, generally because the mother's data of birth was incorrectly reported. They found that the sampling fractions for the women LS members themselves were good (ranging from 1.0 to 1.2). The sampling fractions for the births occurring to the women LS members showed greater variation and tended to be lower for births to older members and particularly high for births to 15-17 year olds.

2.6 Linkage rates: Census to Census

Linkage rates measure the success of adding new information to the LS. Consider the linkage between the 1971 and 1981 Census LS samples. A sample was taken from the 1971 Census containing all people born on LS dates. An equivalent sample was taken from the 1981 Census. Logically, the 1971 Census updated with entries and exits between 1971 and 1981 should be identical to the sample from the 1981 Census. However it is not! A census record may be unavailable - a person may not be recorded in a census or the census computer record may not have an LS date of birth. A person may have entered or exited from the LS but this has not been recorded (primarily due to inconsistent reporting of date of birth). Census based linkage rates can be measured in two date ways - forward linkage (what happened to those people in the 1971 Census based LS sample?) and backward (had all the people in the 1981 Census based sample entered the LS previously?). Considering only those traced at NHSCR, the forward linkage rate was 91 per cent. Table 2 shows that of the 513,000 people in the 1971 Census, 59,000 had died and 6,000 had emigrated. Thus 448,000 should have been recorded in the 1981 census; 408,000 were actually found. The remaining 40,000 failed to be linked for a variety of reasons, the most important being inconsistent recording of date of birth (37%) and household or person not enumerated at the person's usual address (38%) (OPCS 1988). The forward linkage rates varied for subgroups of the 1971 Census. Linkage rates for the elderly, aged 75 and over in 1971 were below average at 86 per cent, probably due to failure to link deaths. For the very old, aged 90 and over in 1971, 96% were known to have died by 1981. Only half of the remainder could be found giving a linkage rate of 49 per cent. The missing 2 per cent had probably died and their deaths had not been linked to the sample, probably due to inconsistent reporting of date of birth. The backward linkage rate 1971-1981 Census was 93% - of the 528,000 people in the 1981 Census, 414,000 were recorded in the 1971 Census. During the decade 64,000 had entered via birth and 14,000 via immigration. The remaining 36,000 effectively entered the LS at the 1981 Census as they hadn't quoted an LS date of birth at birth registration (1,000), immigration (4,000) or 1971 Census (31,000).

Table 2: 1971 and 1981 Census LS samples: forward and backward linkage.

Forward Linkage		Backward Linkage	
	Number		**Number**
1971 Census sample[1]	513,000	1981 Census sample[2]	528,000
Died before 1981 Census	59,000	Born after 1971 Census	64,000
Emigrated before 1981 Census	6,000	Immigrant after 1971 Census[3]	14,000
		Birth entering via 1981 Census	1,000
		Immigrant entering via 1981 Census	4,000
Eligible to be in 1981 Census	448,000	Should have been in 1971 Census	445,000
Recorded in 1981 Census	408,000	Recorded in 1971 Census	414,000
Forward linkage rate	91%	Backward linkage rate	93%

[1] Those traced at NHSCR prior to 1981 Census.

[2] Those traced at NHSCR.

[3] Immigrant to England and Wales, includes those people resident in Scotland in 1971.

2.7 Linkage of events to existing LS sample

Linkage of events to the LS is very high. The cross checking mechanism for deaths and cancer registration should ensure almost complete linkage of deaths (Fox and Goldblatt 1982) and cancer registrations (Leon 1988). Events which are linked solely via quoting of an LS date at registration (births to LS women and widowerhoods) have a lower success rate as linkage relies entirely on the correct date of birth being recorded. Babb and Hattersley (1992) showed a linkage rate of 86 per cent for births to LS women during 1981-1988. Using census data to find "missed" births results in a higher linkage rate - 94 per cent of births to LS women between 1971 and 1981 were linked to the LS. More generally, the LS is fortunate in being revalidated every ten years by the drawing of the appropriate sample from the census. This provides some measure of the missed entries (births and immigration) and exits (death and emigration). It can also be used to check linkage of events during (*e.g.*, births to LS women, widowerhoods).

2.8 The future

The recent computerisation of the Register at NHSCR has already resulted in a significant number of missed events being added to the LS dataset. Future automation of much of the event handling will reduce the scope for error. The availability of 1991 Census data will allow more quality checking. OPCS are collaborating with City University, London on a Technical Report which will cover all aspects of data quality in the LS up to and including 1991 Census. The LS will move to a new computer system which will make more detailed inter-record checking easier. Thus the overall quality of LS data will continue to improve.

3. THE LS IS A DYNAMIC SAMPLE

The LS sample is constantly changing. People enter the LS through birth/immigration or recording an LS date for the first time on a census form. They exit from the LS permanently (through death and permanent emigration) or temporarily (through emigration and subsequent re-entry). More challenging is the issue of inconsistent recording of data of birth. If an LS date of birth is recorded at birth, on registration with the NHS (immigrant), or at a census, then the person becomes an LS member irrespective of the date of birth quoted at other times (*e.g.*, other in censuses, at cancer or death registration). The LS database is structured so that researchers can take the appropriate subsample for their particular area of interest. For example, it is possible to select only those LS members traced at NHSCR (strongly recommended), all those LS members present at one census, only those members present at two/three censuses, to move the sample over time so that LS members are selected as they come into risk. Each type of subsample has its own advantages and disadvantages which are discussed by Hattersley (1992).

4. USE OF THE LS TO ASSESS QUALITY OF CONSTITUENT DATA SOURCES

4.1 Introduction

The LS can be used to provide better quality information for outcome based research. The most widely recognized example is occupational mortality (Fox and Goldblatt). Traditionally, analysis relied on using occupational and socio-economic information collected at death registration and relating it to population figures derived from the census. The LS provides census based information for people who subsequently died. Thus it removes the bias arising from using data collected from different sources. The availability of census information for all the LS members, also means that outcomes (death, widowerhood, fertility *etc.*) can be examined by the range of socio-demographic information collected at census. While the primary output is research on the specific topic, insights can also be provided into the quality of the constituent data.

4.2 Consistency of recording between 1971 and 1981 Censuses

There are some variables which do not change for an individual and should be the same on all records for an LS member. There are also transitions which are impossible. In reporting on the comparison between the 1971 and 1981 Censuses, OPCS (1988) noted inconsistencies - 0.3 per cent for the recording of sex, 0.4 per cent for the recording of country of birth (over 5 per cent for some countries outside the United Kingdom) and 3.4 per cent for date of birth (Figure 8). There were 0.4 per cent of people for whom an impossible transition of marital status was recorded (*e.g.*, married/widowed/divorced in 1971 and single in 1981).

4.3 Double enumeration in 1981 Census

At census, people can be included on two census returns - the one relating to their usual residence and the other where they were enumerated (for those away from their usual residence). As a person based dataset, the LS brings together both the census records for an LS member. In the 1981 Census, 525,000 LS members were enumerated at their usual address. A further 4,300 were enumerated away from their usual address and were included on the form relating to their usual residence. The remaining 6,500 were enumerated away from their usual address and only on that form. This provides a measure of the level of double-counting which can assist in the interpretation of data of different census population bases. It is also possible to consider the consistency of the information provided on the two occasions.

5. SUMMARY

The OPCS Longitudinal Study uses record linkage of routinely collected data to add value to the data. It was established as a national resource for the study of social and demographic change and social and demographic variation in event rates. The sample is a representative one and the quality of the record linkage is very high. It is a challenging dataset to use as it is constantly changing and inconsistent recording of date of birth can result in missed data. In addition to supporting a wide range of research, the LS can be used to examine the quality of the constituent data sources.

REFERENCES

Babb, P., and Hattersley, L. (1992). An examination of the quality of OPCS Longitudinal Study data for use in fertility analyses. LS User Guide Number 10, London: Social Statistics Research Unit, City University.

Britton, M., and Birch, F. (1985). 1981 Census post-enumeration survey. London: HMSO.

Hattersley, L. (1992). Selecting samples for analysis for the LS. Longitudinal study newsletter no. 7, London: OPCS.

Fox, J., and Goldblatt, P.O. (1982). 1971-1975 Longitudinal study: Socio-demographic mortality differentials. LS series no. 1, London: HMSO.

Leon, D.A. (1988). 1971-75 Longitudinal study: Social distribution of cancer, LS series no. 3. London: HMSO.

OPCS (1983). Census 1971 General Report, Part 3 Statistical assessment. London: HMSO.

OPCS (1988). Census 1971-1981 The Longitudinal study: Linked census data, England and Wales. London: HMSO.

OPCS (1989). Longitudinal Study Newsletter no. 1. London: OPCS.

OPCS (1992). Provisional mid-1991 population estimates for England and Wales and constituent local and health authorities based on 1991 Census results. OPCS Monitor, PP1 92/1, London: OPCS.

Swerdlow, A.J. (1986). Cancer registration in England and Wales: Some aspects relevant to interpretation of the data. *Journal of the Royal Statistical Society Series A*, 149, 146-160. London.

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

EXPLORING NONSAMPLING ERROR IN A LONGITUDINAL SURVEY OF INDIVIDUAL TAXPAYERS

S. Hostetter[1]

ABSTRACT

To respond to the expanding and varied data needs of tax policy researchers, the U.S. Internal Revenue Service (IRS) has built longitudinal studies into its statistical samples of individual income tax filers. This paper examines the largest of those panel studies -- a 1987-based panel of 90,000 individual taxpayer families. Both anecdotal and empirical evidence of nonsampling errors is presented, and the methods IRS has used to correct or compensate to improve data for estimation and policy modelling purposes are described.

KEY WORDS: Tax statistics; Nonsampling error; Tax policy.

1. INTRODUCTION

This paper addresses the process we, in IRS, used for correcting data. This discussion of process compliments some of the more quantitative work also presented at this conference (Czajka and Schirm 1992). The U.S. Internal Revenue Service (IRS) has developed an important new panel of 90,000 tax families. A recent, in-depth, inferential review of taxpayer reporting error and IRS linking procedures for tax families in this panel provides valuable information on nonsampling error.

In 1987 the U.S. Treasury's Office of Tax Analysis asked Statistics of Income at IRS to begin a major redesign of our Individual Tax Return Sample, to improve it for more accurate modelling of the effects of tax policy recommendations (Hostetter and O'Conor forthcoming). We were asked to do three things:

- **To design and implement a Tax Family Unit**, so that Treasury and Congress' Joint Committee on Taxation could model the effect of tax law changes on family economic units (Nelson 1986);

- **To redesign the stratification of the cross-sectional sample selection**, to strengthen the sample of income components of importance to tax policy, and to obtain better coverage of certain demographic groups (Hostetter *et al.* 1990; Schirm and Czajka 1991); and

- **To design and implement a panel of individual tax returns**, to measure the effect of tax policy on individual taxpayer behavior over time, as opposed to measuring change in aggregates.

Previous papers have described our efforts with regard to the first two objectives, and we welcome the opportunity the Symposium on Longitudinal Surveys provides for sharing as well as gaining information and perspective concerning the design and implementation of the IRS panel of tax returns. This paper will briefly cover some important historic panels, and it will mention some general limitations to panels. However, the focus of the paper is the description of the IRS methodology used, and the preliminary results from the review of about 331,000 tax return records, linked as both panel and tax family units. The paper will conclude with a discussion of our plans for maintaining and using the panel and IRS's plans for improving the quality of administrative records.

[1] S. Hostetter, Internal Revenue Service and Joint Committee on Taxation, Room 1015 Longworth HOB, Washington, DC, U.S.A. 20515.

2. HISTORICAL PANELS AND PANEL LIMITATIONS

2.1 Early Panels

Much early work in panel development led the way for the IRS panel. The first major panel -- The Continuous Work History Sample (CWHS) -- was begun in the late 1930's by the Social Security Administration (Buckler and Smith 1980), and the earliest research using these data was presented in the 1970's (Ruggles and Ruggles 1974). These CWHS data are particularly pertinent to the development of the IRS panel because IRS has included a panel of 20,000 CWHS Social Security Numbers (SSN) embedded in its annual individual sample since 1979, and these same 20,000 individuals are also embedded in the current panel. Many other early panels also served as a basis for the design and development methods used in the IRS panel. Some of the other important ones were:

- **The Panel Study of Income Dynamics (Duncan *et al.* 1984),**
- **The Canadian 10 Percent Longitudinal File (Hoskins and Yazdani 1985)**, and
- **The Survey of Income and Program Participation (Kasprzyk and Frankel 1985).**

Other work that laid the groundwork for the current effort included integration of administrative sources with survey data (Scheuren 1985; Scheuren 1975), and the development of a public use file (David 1989).

2.2 Limitations of Panel Data

Clearly, the panels mentioned above offer a considerable source of knowledge, particularly concerning life cycles of income and wealth or individual behavior caused by tax policy or economic changes. Nevertheless, we realized there is a cost and some limitations to these gains, particularly when you begin with administrative records. For example:

- A panel is good for reviewing the before and after picture of **tax law changes**, but, for some data uses, major changes cause a break in continuity of data which may be a problem.
- Large quantities of **resources** are necessary to initiate and maintain a panel, mainly to keep track of all the people. Keeping knowledgeable and trained staff devoted to a project for many years is difficult.
- The **weighting issues** for a series of panels are very complex. Generally both design-based and model-based weights will be needed. Sometimes weighting difficulties can be mitigated by placing greater emphasis and effort on the initial stratification at the time the panel is selected (Czajka and Schirm 1992).
- Creating and continuing a large panel is **expensive** if it's to run a long time or if it's to provide information on diverse characteristics.

3. DESCRIPTION OF THE IRS PANEL

The **Statistics of Income, or SOI, panel** consists of about 90,000 tax family units. It was initiated for Tax Year 1987 by selecting *90,000 nondependent (parent) returns* and returns for any *dependents claimed on the parent returns*. The already selected, 1987 SOI cross-sectional sample was designated as the panel sample, with only minimal adjustments (Czajka and Walker 1989). The SOI cross-sectional sample includes "prior year" returns (those filed in the same calendar year as those for 1987, but covering an earlier tax period). Such returns tend to have different characteristics from timely returns, and they are designed to represent the returns that would be filed late for the current Tax Year. So, the initial composition of the panel was determined by the sample design of the basic annual individual sample, which, in turn, was designed to meet many needs for several customers, none incorporating longitudinal concepts. This decision was expedient at the time, and without the knowledge that we now have, it is unlikely that a differently selected panel in 1987 would have served all our needs better than this one.

In the U.S. the Social Security Administration assigns a uniquely identifying number to almost all individuals, comparable to the SIN in Canada. The U.S. number is aptly called a social security number, and it's abbreviated as SSN. We created a file with all SSN's reported on the 90,000 returns -- over 200,000 of them. The SSN is used to define individuals who are members of the panel. Because it is imperfect, using the SSN causes you to miss some individuals that are members and also to collect others that you do not want because they are not members. For processing purposes, this file defines the actual "card-carrying" members of the panel. With expansion due to changes in family structure, SOI has about 135,000 returns in the panel for 1990, but they still represent, and will be weighted as 90,000 panel units.

4. PLANNING PANEL REVIEW

4.1 Objectives of Panel Review

The four major goals of the Panel Review Project were to:

- **Define the panel membership,**
- **Develop a "clean" panel database covering the first three years,**
- **Establish accurately linked tax families**, and
- **Gain information to build models for future panel review**.

The process for the project was to review panel links, panel units, family links and individuals for all three years, mechanically when possible, and manually when units failed the initial, rigid matching criteria. By fall of 1991 (four years after starting the panel) SOI staff had created a database containing all panel returns selected for 1987, 1988 and 1989.

Five years after initiating the panel (fall of 1992) SOI has recently completed editing 331,000 returns, including manual review of over 150,000 returns covering the first three years. This time span is key to the correction and improvement timetable because you cannot review the relationships and activities of panel members intelligently until you have at least three years of data. Unlike a single year's data, three years provides enough information to delineate between change and error. Therefore, SOI staff could not even link the returns into families *reliably* until this cleanup work was completed.

4.2 Definition of the Tax Family Unit

The tax family is the nondependent taxpayer(s) on an individual tax return and all dependents claimed by the nondependent taxpayer(s) for a *specific year*. The nondependent taxpayer can be either the first person listed -- the primary filer on a return with a single filing status or the primary and secondary (spouse) filers on a joint return. In over 90 percent of joint returns the male is reported as primary. To establish the tax family, SOI linked any tax returns filed by dependents to the tax return on which they were claimed -- the parent return. Most of these dependents are children who met the filing requirements even though their parents claimed them as exemptions.

The tax family is implemented administratively based on tax return information rather than through survey collection and contact with respondents, which are usually used to establish "households," more commonly used for grouping income and demographic characteristics. The purpose of tax families for tax policy analysis is to capture an economic unit sharing the same pool of income. We realize that they, like households and other family definitions that are used broadly, may be imperfect. However, this is the unit available to us in IRS.

SOI is constructing tax families units for both its cross-sectional and panel samples, however, this paper will focus on the panel. Tax families describe an organizational relationship for a single year, unlike the panel link, where individuals stay in the same panel unit for life (or rather the life of the panel). Tax families may stay constant, but they also may change. For example, a dependent who grows up, leaves home, and is no longer claimed as a dependent, represents a separate family unit if he files a return. Part of the review effort was to determine whether change represented actual change in the family dynamics, whether probable such as our example, or unusual; or whether change represented a different individual and needed correction. About 70 percent of returns represent married taxpayers, almost all of which are joint returns and only a few of which are "married filing separate" returns. The other 30 percent of returns are single taxpayers, and about 20 percent of those are "head of household" returns. Taxpayers filing "head of household" returns are single and must have a dependent who meets certain conditions. Single taxpayers could also have dependents.

4.3 Dependent SSN Problems

The reason the panel was initiated for 1987 is that it was the first year that taxpayers were required to supply the SSN for dependents, and so it was the first opportunity to link individuals into families from tax return information. Until that year SOI could not have linked dependents to the parent return. The law was phased in over three years, finally requiring such reporting for children one year old or over. Not surprisingly, taxpayers were slow to comply, and they found a great assortment of methods for not complying. In our review we considered a dependent a panel member if he was claimed as an exemption in 1987, whether or not a viable SSN was reported. However, we could not include such dependents on the file until their correct SSN appeared in a later year.

Some taxpayers used their own SSN's for dependents, some omitted the SSN, and some borrowed SSN's from other families or relatives. In one case, a taxpayer used the SSN of an ex-spouse for one of her dependents. In a later year the ex-spouse and his family were pulled into the original panel family because of this common SSN, and -- surprise, he was claiming the same two children on his return. Although this specific example did not occur frequently, a variety of similar cases did occur frequently.

4.4 Initial Linking of Panel Returns

The SSN was the consistent basis for creating and reviewing tax family links, structure, and change. The panel database that SOI staff established in 1991 was our first linked panel file for review. The first effort, as in most new work, was difficult because we did not know what level of error to assume for planning purposes; we did not know what methods of review would prove most fruitful; and we did not know how the errors would group themselves for intelligent review. In an IRS production environment these deficiencies are particularly acute.

For this first effort we chose not to make assumptions regarding the validity of family or panel links -- rather, we used the most simple and *inclusive* linking procedure. Since, in the U.S., dependents can file separate returns but be listed on their parents' return for exemption purposes, the initial linking process simply linked *to the parent return* any returns filed by dependents claimed on the parent return. Without review we were unwilling to delete, or fail to link, any return with its primary or secondary SSN claimed on a panel parent return.

4.5 Taxpayer Reporting Behavior

Our ability to construct accurate tax family units through selecting and linking was clearly dependent upon human behavior that affects taxpayer filing and reporting -- and not always in positive ways. Tax reporting behavior is a major portion of our nonsampling error. How do we assign each individual to the correct panel, panel unit, and tax family? It's ***very difficult***, and we have developed three specific files to help track and manage change in these characteristics.

Creating and maintaining tax families over a number of years is also complicated by the fact that taxpayers commit what we have dubbed "family matching sins," only seven of which are covered here.

- **Marrying** -- change in status requires identification and review. Worst case is marrying another panel member, which additionally requires weighting adjustments,
- **Divorcing** -- results in two families in the same panel unit,
- **Remarrying** -- introduces a nonmember (a "VISITOR") to the panel and maybe dependent visitors that arrived with him or her (A visitor is anyone appearing on a panel return who is not a member of the panel),

- **Claiming Dependents** -- (kids or parents) who are not panel members, thereby creating more visitors,
- **Divorcing a Visitor** -- the visitor must be deleted from active selection and subsequently, any dependent visitors that are not claimed by our panel member must also be deleted,
- **Sharing your SSN**, and
- **Reporting the Wrong SSN.**

Only the last two represent errors. The rest simply create the need for review with verification or correction. This process will be ongoing and will benefit from the review process described in this paper, which gave us considerable insight into the problems caused by behavioral issues such as these.

5. THE PANEL REVIEW EFFORT

5.1 Defining and "Perfecting" the Individual Taxpayer

Having linked returns to establish "preliminary" tax families, we then had to verify that the returns were reasonably linked and, in fact, created a tax family. For this cleanup process the most important field to review or change during the cleanup process was, of course, the SSN. Information about **individuals** was reviewed and corrected based on **tax return** information. Although potential errors relating to individuals or tax returns were identified for review, once we examined a panel case we looked at the entire unit with *ALL returns for ALL three years*. Correcting the SSN's was the main goal, but review and correction procedures also covered the panel and family identification. To clarify corrections, codes were added to designate the specific individual on the return to whom the correction applied. Additionally, we assigned status codes describing the kind of action -- such as a delete or change -- and reason codes describing the cause or condition for the error. These additional codes will be used in future development of computerized review models.

The six data fields we corrected were:

- **Panel identification number,**
- **Family identification number,**
- **Primary taxpayer SSN,**
- **Secondary taxpayer SSN,**
- **Dependent SSN's (up to 10 dependents),**
- **Filing status code at time of entering panel.**

A few income and tax items were included for informational purposes on the review documents, but none were changed. Also, we made no effort to correct information to make it comply with the tax code. Rather, the effort was aimed at capturing taxpayer information without reporting or processing error. In other words, our goal was to eliminate and measure the nonsampling error relating to identification of individuals and tax families.

5.2 Information Documents Used for Verification and Correction

We had a broad range of pertinent information in front of us during our review. From the tax return, via the IRS master accounting file, we extracted and used:

- **Full name of taxpayer(s),**
- **Address of taxpayer(s),**
- **Primary SSN,**
- **Secondary SSN,**
- **All dependent SSN's,**

- **Filing (marital status),**
- **Dependent status indicator,**
- **Number and type of all exemptions,**
- **Names of all dependents,**
- **Selected income and tax items.**

From the Statistics of Income sample selection and initial panel coding process we extracted the following information:

- **Sample strata defining codes,**
- **Questionable family match code,**
- **Name control from tax return,**
- **Tax Year covered by the return,**
- **Family ID,**
- **Panel ID.**

Finally, from the Social Security Administration (SSA), we obtained and used three very important fields:

- **Name control (first four letters of each of the last names filed with Social Security for the specific social security number) for *all identified SSN's*,**
- **Date of birth for *all identified SSN's***, and
- **Date of death for *all identified SSN's*.**

These last data had a strategic impact on the accuracy of our review. For example, if the name control for the primary taxpayer (extracted by IRS from the name written on the return) did not match the SSA name control for the SSN, there was a strong indication that the entire return was incorrectly represented by the controlling SSN -- that of the primary taxpayer. Frequently, we would see either another return for the same year or the next year, with the same SSN and with a different individual represented (Steffick 1992).

5.3 Organizing Panel Units for Review

After completing our initial links and experimental work the SOI staff refined the definitions of the initial review groups and developed many additional ones. The review groups fell into two categories -- "Clean" and "Questionable." The "Clean" groups were identified mechanically, and the accuracy of the definitions was tested with manual review of a sample of units. "Clean" groups were carefully restricted, and had the following characteristics:

- **No nonmatches,**
- **No change,**
- **Zero, one, or two dependents**, and
- **The base year return did not represent a prior year.**

Nonmatches were defined as name controls for the SSN not matching names reported on the return, names of secondary taxpayers or dependents not matching the name of the primary taxpayer, or ZIP Codes of dependents that did not match their parents. "No change" meant **no change** from year to year -- in filing status, SSN, dependents, or ZIP Code. Initially, prior year returns -- those covering Tax Years 1986 or earlier -- were excluded from this "clean" group, regardless of their other characteristics. Subsequently, after some manual review of returns selected in the base year with a prior year return, we eliminated that restriction, so that if all other characteristics remained constant prior year returns could be considered "clean." This decision eliminated 1,517 returns from the manual review process.

After completing and testing the computer review, 153,153 of the total 330,956 returns remained for manual review. Using the criteria described, we were able to designate 54 percent of the returns as "clean" based on computer checks.

5.4 Developing Questionable Groups for Manual Review

Even though we had little insight as to the pattern of errors we could expect, for purposes of managing the review and for gathering summary information about various kinds of errors, we decided to group returns for review by certain error conditions. Since the groups identified certain types of errors, but did not exclude others, a given return was not necessarily described uniquely by a review group. This meant that the order in which the groups were processed made a difference. Specifically, the few groups that we wanted fully inclusive -- so that we could monitor the frequency or extent of a specific type of error -- needed to be selected first. In all cases, all potential error conditions in a panel unit were reviewed when the unit was identified in a group. Listed below are some abbreviated descriptions used in developing control groups for manual review. Definitions and priority order were still subject to change (and, in fact, some were changed) in the first phase of production.

- **The base year return had a dependent status code,**
- **Prior year returns that were not designated "clean",**
- **Returns with primary and secondary taxpayers representing different panel units,**
- **Returns with visitors (not panel members),**
- **Returns linked as dependents, but with a nondependent status code,**
- **Name control on the return not matching the social security name control,**
- **A married filing separate return in the panel unit,**
- **The secondary SSN is inconsistent among years.**

We found that having returns grouped by type of error was helpful, not only for managing progress of the project, but also for training staff, because it reinforced learning in a short time span. The consistency of reviewing like conditions together was also useful for **understanding and documenting systematic error conditions.** During evaluation and planning meetings we were able to distinguish between systematic and exceptional error. This, in turn, enabled us to set procedures for the systematic conditions, and to correctly identify and handle exceptions as such.

5.5 Additional Research for Exceptional Cases

Even with the considerable detail available to our reviewers, some error conditions could not be corrected based on information provided on the returns for all three years. For example, our data would not provide the information to correct the record for a taxpayer selected in the panel based on the 1987 return, but with the wrong primary SSN provided. It was the specific person with the selected tax characteristics that we wanted to include, but if we didn't have his SSN we could not find him again. If our three years of data did not provide information to correct the SSN, field staff reviewed IRS Master File Account Data, using a name search to attempt to identify the correct SSN. Although this additional effort was time consuming and expensive, we felt it was worth the trouble for this first cleanup effort. This rationale was partly based on our view that this initial cleanup was **both production and research** work. The production effort was clearly to clean the file as we have described. The research effort was to:

- **Use different methods of review,**
- **Document our methods,**

- **Document our results,**
- **Evaluate these methods.**

The evaluation will be an ongoing process to study methods for efficiency, technical accuracy, and cost effectiveness.

6. INITIAL RESULTS OF PANEL REVIEW

Figure 1 shows a breakdown of error rates for each category of SSN. Surprisingly, dependent SSN's had the same error rate as secondary SSN's. And, a greater percentage of them were corrected than for secondary SSN's. This may be because dependent SSN's showed steady improvement as taxpayers began to comply to the new requirement. Whereas, when an error for a secondary taxpayer occurs, it is often repeated every year by copying the previous year's data. An SSN was corrected only when there was conclusive evidence -- using name control, age, and the name and address written on the return -- that we knew the correct SSN. Secondary's had the highest percentage of SSN's in the "uncorrected" category, where we were unable to correct what appeared to be an incorrect SSN. Total errors for both secondary's and dependents were 3.5 percent. As expected, error and "uncorrected" SSN's for primary taxpayers were minimal. Our knowledge of IRS administrative procedures for reviewing, rejecting, and correcting SSN's for primary taxpayers correctly guided us to the assumption that there would be little error for this group at the point that SOI selects the return from the Master File. We know that IRS verifies that the primary SSN is a legitimate SSN and that the name control matches for the SSN. The IRS processing not only eliminates reporting errors such as transposition of digits, but it also catches and precludes most keypunching errors.

Figure 1: Error Rate by Type of SSN.

Type of SSN	Percent in Error	Percent Corrected	Percent Uncorrected
Primary	0.18	0.03	0.15
Secondary	3.50	1.10	2.40
Dependent	3.50	1.70	1.80

Figure 2: SSN Errors by Number of Incorrect Digits.

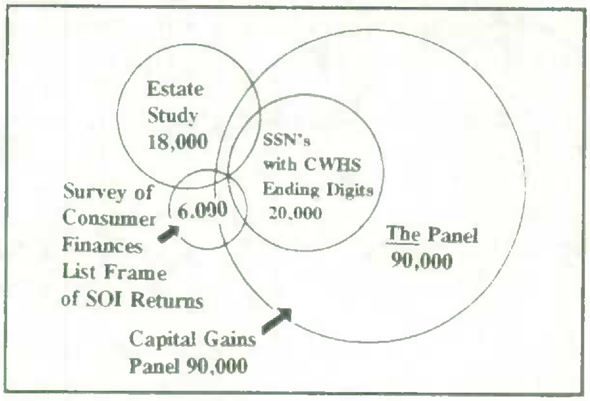
Type of SSN	Percent by Error Type		
	1-2 Digits	3-4 Digits	5+ Digits
Primary	33.7	1.3	56.0
Secondary	76.7	2.5	20.8
Dependent	81.4	4.3	14.3

Figure 2 looks at the nature of error in SSN's. For primary SSN's the largest percentage of errors had five or more incorrect digits, considered a wrong SSN. Note, however, that secondary and dependents SSN's had the highest percentage of known error with 1 or 2 digits wrong. These were most likely transposition errors by the taxpayer or the keypunch operator. The pattern is consistent with what we know of IRS processing standards. These data show that we do have a nontrivial amount of error for secondary and dependent SSN's. However, these estimates are preliminary, and the frequency of error -- particularly in regard to secondary and dependent SSN's -- in the panel sample cannot be used as a basis for inferences to the population of tax returns. As discussed in the final section of this paper, IRS plans to improve the quality of both of these categories of SSN's.

7. THE SOI CONSOLIDATED PANEL

This concludes the description of the initial panel review process. The basic panel, just described, however, is only the starting point. Our full plan is to develop a **consolidated panel** which overlaps with the basic panel, in order to benefit multiple uses for policy analysis. In other words, by overlapping several sets of longitudinal data on individual taxpayers, we can provide richer data for each of the specific studies and at a lower cost. (See Figure 3.) The discussion of nonsampling error in this paper was limited to a review of *the* main or basic individual panel, but it should be viewed in the context of the Consolidated Panel since similar work is planned for the other panels in the future (Hostetter 1992). The four other major panels that will be included in the SOI panel design, management and processing are the following:

Figure 3: Components of Consolidated Panel


Estate Study 18,000
SSN's with CWHS Ending Digits 20,000
Survey of Consumer Finances List Frame of SOI Returns
6,000
The Panel 90,000
Capital Gains Panel 90,000

- As mentioned earlier, **A CWHS Sample** of 20,000 SSN's is included in the panel. As part of the panel, tax families were also initiated for this group. This group provides a direct overlap with the cross-sectional sample, for use in comparative time series studies.

- **The Survey of Consumer Finances** is a triennial household survey conducted by the Federal Reserve Board in cooperation with SOI. It is used by the Federal Reserve, Congress, Treasury, and researchers to study a broad range of financial characteristics of households (Kennickell and Woodburn 1992).

- Based on knowledge gained in a pilot study of 18,000 decedents and beneficiaries listed on 1989 estate returns, SOI will initiate a new 1993 **Estate Collation Study** with partial overlap of beneficiaries and their tax families (Johnson and Woodburn 1992).

- Beginning with 1993 SOI will incorporate **Capital Gains** data for the entire panel of 90,000 units. The panel units and, within those, the family units will remain; the difference is in the additional data -- the asset codes and transactions information for all capital gains or losses reported on the panel returns (Holik, Hostetter and Labate 1989).

8. FUTURE PLANS

This brings us to our plans for the future -- both for SOI and for IRS Processing. For 1992 and 1993, SOI will improve panel cleanup methods, using models based on five years and based on information about taxpayer reporting error gathered from the cleanup process described in this paper. Accuracy and timeliness will be improved by incorporating the panel cleanup into the SOI on-line production processing, first at the tail end and, in 1993, at the front end of the SOI statistical processing.

Also, for 1993, SOI will perfect its match study of information returns to tax returns, which uses the panel described in this paper, by correcting, to the extent possible, the SSN's on our panel file. IRS has already started verifying secondary SSN's as part of its regular Master File processing, and will verify the SSN's of the first two dependents for 1993. This will mean that the error rate for matching SSN's, both for improving the panel quality and improving IRS's ability to make important contributions to U.S. population estimates, will reflect the low error rate currently associated with primary SSN's -- 0.2 percent.

REFERENCES

Buckler, W., and Smith, C. (1980). The continuous work history sample (CWHS): Description and contents. *Economic and Demographic Statistics*, Social Security Administration.

Czajka, J.L., and Schirm, A.L. (1992). Selection and maintenance of a highly stratified panel sample. *Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys*, Ottawa, Ontario, Canada.

Czajka, J.L., and Walker, B. (1989). Combining panel and cross-sectional selection in an annual sample of tax returns. *American Statistical Association 1989 Proceedings of the Section on Survey Methods*.

David, M.H. (1989). Managing panel data for scientific analysis: The role of relational database management systems. *The American Statistical Association International Symposium on Panel Surveys*, John Wiley & Sons.

Duncan, G., *et al.* (1984). The role of panel studies in a world of scarce resources. *The Collection and Analysis of Economic and Consumer Behavior Data* (S. Sudman and M.A. Spaeth, Eds.), Bureau of Economic and Business Research, Champaign, IL.

Holik, D., Hostetter, S., and Labate, J. (1989). Sales of capital assets. *American Statistical Association 1989 Proceedings of the Section on Survey Research Methods.*

Hoskins, E., and Yazdani, M. (1985). Some longitudinal methodologies and issues relevant to the modelling and analysis of tax policies and programs. *Multinational Tax Modelling Symposium Proceedings*, Revenue Canada Taxation.

Hostetter, S. (1992). Managing multiple uses of panels. *American Statistical Association 1992 Proceedings of the Section on Social Statistics.*

Hostetter, S., and O'Conor, K. (forthcoming). Satisfying the need of income policy modelers while preserving the reliability of descriptive statistics. *Statistics of Income Methods and Results – From Data to Information: 1991-1992*, Internal Revenue Service.

Hostetter, S., *et al.* (1990). Choosing the appropriate income classifier for economic tax modeling. *American Statistical Association 1990 Proceedings of the Section on Survey Research Methods.*

Johnson, B., and Woodburn, L. (1992). The underlying methodology of the estate multiplier technique: Recent improvements and estimates for 1989. Paper presented at the 1992 Joint Statistical Meetings, Boston, MA.

Kasprzyk, D., and Frankel, D. (Eds.) (1985). *Survey of Income and Program Participation and Related Longitudinal Surveys: 1984*, Bureau of the Census.

Kennickell, A.B., and Woodburn, L. (1992). Methodological issues in the estimation of household net worth: Results from the 1989 survey of consumer finances. *American Statistical Association 1992 Proceedings of the Survey Research Section.*

Nelson, S.C. (1986). Family economic income and other income concepts used in analyzing tax reform. *Compendium of Tax Research, 1986*, Department of Treasury, Office of Tax Analysis.

Ruggles, N.D., and Ruggles, R. (1974). The anatomy of earnings behavior. *The Distribution of Economic Wellbeing*, (F. Thomas Juster, Ed.), Cambridge, MA, Ballinger.

Scheuren, F. (1985). Methodological issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*, Internal Revenue Service.

Scheuren, F. (1975). ORS management of the HEW income security survey -- some administrative issues. Working Paper, Office of Research and Statistics, Social Security Administration.

Schirm, A.L., and Czajka, J.L. (1991). Alternative designs for a cross-sectional sample of individual tax returns: The old and the new. *American Statistical Association 1991 Proceedings of the Section on Survey Research Methods.*

Steffick, D. (1992). Analyzing longitudinal data linkages in a panel of individual tax returns. *American Statistical Association 1992 Proceedings of the Section on Social Statistics.*

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

USING ADMINISTRATIVE RECORD INFORMATION TO EVALUATE THE QUALITY OF THE INCOME DATA COLLECTED IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

J.F. Coder[1]

ABSTRACT

The Survey of Income and Program Participation (SIPP) collects a comprehensive set of information about the economic situation of American households. Interviews are conducted at 4-month intervals and income and work experience data are recorded for each month of a 4-month reference period. This paper explores the accuracy of the wage and salary income data collected in the 1990 SIPP using an "exact" match between the survey data and data obtained from 1990 Federal individual income tax returns. Various measures of the magnitude and characteristics of survey response and imputation errors are presented.

KEY WORDS: Response error; Income; Data quality; Data linkage.

1. INTRODUCTION

In this paper I examine the magnitude and characteristics of measurement error[2] for the wage and salary income data collected in the Survey of Income and Program Participation (SIPP). Measurement error has been defined here to be the difference between survey responses on wage and salary income amounts and "comparable" amounts reported on Federal individual income tax returns for calendar year 1990. Investigations based on comparisons of survey responses with administrative data, such as the tax return information used in this study, are relatively rare since access to administrative record information is very restricted and the cost of linking survey and administrative data is high. This assessment of the nonsampling error for wage and salary income was made possible through a joint agreement between the Bureau of the Census and the Internal Revenue Service (IRS) that permits the Bureau of the Census to link survey respondents with their tax returns for purposes of evaluating data quality[3].

Previous evaluations of the quality of income data collected in household surveys typically indicate downward biases in the survey estimates of income amounts when these estimates are compared to independent sources such as the National Income and Product Accounts (NIPA). Comparisons available for the March CPS income supplement[4] indicate that downward biases of from 1 to 3 percent have been experienced over the past 10 years

[1] J.F. Coder, Housing and Household Economic Statistics Division, Bureau of the Census, Washington, DC, U.S.A. 20233.

[2] In this paper the terms "measurement error", "nonsampling error", "survey error", and "response error" are used interchangeably to indicate the difference between the survey and tax return data. It is acknowledged that in some contexts these terms may have somewhat different definitions.

[3] Under an agreement with the Internal Revenue Service (IRS), the Bureau of the Census receives an extract of the content of all Federal individual tax returns annually. This extract is used for developing estimates of population for States and counties during the postcensal period and for evaluating the quality of data collected in surveys. In accordance with Title 13, the release of information that would allow individual survey respondents to be identified by the IRS or anyone else is prohibited. Linkage of survey respondents to their tax return information occurs at the Bureau of the Census and the resulting linked files are maintained in secured areas within the Bureau of the Census.

[4] The March Current Population Survey (CPS) income supplement collects income and work experience data for a sample of about 60,000 households. This survey, which is conducted annually, asks detailed questions about sources and amounts of income received during the previous calendar year. It has been the main source of income and poverty data for the United States since 1947.

in the survey estimates of aggregate wage and salary income (U.S.Census 1991). Similar comparisons for SIPP for calendar year 1984 indicated a possible 6 to 7 percent downward bias (Vaughan 1989).

Evidence exists showing that these downward biases in the aggregate are symptomatic of an underlying, complex pattern of error. Recent research (for example Brownstone 1992; Coder 1990; Scholz 1990; and Lillard *et al.* 1986) has shown that the measurement error related to wage and salary income is not randomly distributed. Using the PSID Brownstone finds that measurement error is correlated with "true" earnings. Coder also finds this correlation and significant downward biases in measures of inequality for the March CPS. In his study focusing on the earned income tax credit, Scholz, using the SIPP 1984 panel, observed underestimates of families with wage and salary incomes of $50,000 or more. Finally, in an examination of procedures for imputing missing responses to wage and salary amount in the March CPS, Lillard *et al.* alleged serious downward biases in the assignment of wage amounts to nonrepondents.

2. COLLECTION OF THE SIPP WAGE AND SALARY DATA

This assessment of measurement error in wage and salary income is based on the sum of wage and salary amounts recorded for each month of calendar year 1990. Since the SIPP design used a 4-month reference period with interviews commencing in February 1990 for the first rotation, the aggregation to calendar year amounts required that data from either 3 or 4 interviews be used depending on the rotation.

Persons in the wage and salary workers universe included (1) employees of private companies or businesses (2) employees of private nonprofit organizations, (3) employees of Federal, State, or local governments, (4) members of the Armed Forces (military) and (5) self-employed owners of incorporated businesses.

The SIPP questionnaire contains two identical sections related to employment for wage and salary workers (excluding self-employed incorporated) so that details about two different wage and salary jobs can be recorded for the 4-month reference period. These sections begin with questions that provide information needed in the assignment of standard occupation and industry codes for the job and business. Next are questions to determine the period during which the job was held, hours working, reasons for leaving the job (if not held the entire 4-month period), mode of payment, *e.g.*, hourly pay rate, annual salary and labor union membership.

The wage and salary employment sections conclude with questions covering amount of wage and salary income received each month. The income concept used is the "gross" monthly wages and salaries "before deductions". Counted in wages and salary income are regular time pay and earnings in the form of tips, bonuses, commissions, and overtime pay. This definition excludes fringe benefits or pay in-kind, such as meals, lodging, use of automobiles, *etc.* The before-deduction definition requires that the amount reported must **not** exclude amounts such as income and payroll taxes, employee contributions to health insurance and pension plans, saving plans, union dues, *etc.* In terms of accounting principles, amounts are recorded in the month received, not in the month accrued (the month the work was performed).

Income received by self-employed owners of incorporated businesses is also counted as wage and salary income even though it is not collected in that part of the questionnaire described above. The concept used to collect amounts of income for the self-employed incorporated is much less well-defined than that for the more typical wage and salary worker. For the self-employed the definition of monthly income covers any regular "salary" or other money received from the business.

Amounts of wage and salary income are imputed or assigned when the respondent failed to provide an answer. The imputation process is based on a standard "hot deck" technique. Using this technique, nonrespondents are "matched" to respondent having "similar" characteristics. The wage and salary amount from the respondent is then assigned to the nonrepondent. Characteristics used in the matching process include variables such as age, race, sex, education, occupation, hours worked, weeks worked, place of residence, *etc.*

3. LINKAGE OF SIPP AND TAX RETURN WAGE INFORMATION

This assessment of the quality of the wage and salary data collected in the SIPP was made possible through a linkage between data collected in the survey and data contained on Federal individual income tax returns. Information from tax returns was provided to the Bureau of the Census by the IRS for the explicit purposes of improving small-area population estimates and conducting data quality evaluations.

Data from the SIPP and tax returns were linked using the social security numbers (SSNs) collected in the survey and reported on each tax return. The SSNs collected in the SIPP were validated by the Social Security Administration (SSA) to ensure their accuracy prior to the linkage with tax returns.[5]

Following validation, a linkage was attempted between the 51,500 survey SSNs and the roughly 113 million tax returns filed for 1990. Any tax return having an SSN matching that of a SIPP sample person was extracted. This process yielded an extract file containing a total of about 31,000 tax returns (duplicate returns were extracted if the SSNs of both the husband and wife matched the same married, filing jointly tax return). Data from this file of matching tax returns was then merged with the survey data to create a linked or "exact match" file that provided the basis for the comparisons presented in this paper.

4. THE STUDY UNIVERSE AND LIMITATIONS

4.1 The Study Universe

The universe for this study is not representative of the overall population but of a very specific subset of the SIPP sample. This subset consists of married couples[6] having validated SSNs for both the husband and wife, matching to a married-joint tax return, and having a nonzero wage and salary income amount on either SIPP or the tax return. There were a total of 5,703 married couples meeting these restrictions on the file created for this study. This represents about 62 percent of the total 9,267 husband-wife units present in the 1990 SIPP sample as of March 1991. Of the 9,267 total, 864, or about 9 percent, were excluded because one or both SSNs could not be validated. Of the remaining 8,403 cases, matches to tax returns were found for 6,548. Those without matches consisted of 902 cases for which no match was found even though validated SSNs were available for both spouses. Most of these were couples that did not file a tax return and therefore had no chance of matching. Situations in which the SSN of one spouse matched to a tax return that was either not a joint return or did not contain a "secondary" SSN (the SSN for the other spouse) accounted for another 857 of the cases dropped from this analysis[7]. Finally, 354 additional cases were excluded, mainly because only one of the spouses' SSNs matched to a joint return containing both a primary and secondary SSN.

Excluded entirely from consideration was the universe of unmarried sample individuals. This group was not included in this analysis because the reporting error problems may differ significantly from those of married couples whose wage and salary income represents the sum of the amounts for both the husband and wife. An examination of the unmarried universe will be undertaken separately.

[5] SSN validation was a two-part operation. For those persons reporting an SSN in the survey their SSNs were validated by matching to SSA administrative files. In this process the name, date of birth, race and sex were used to make the validation. For persons not reporting an SSN in the survey a search of the SSA record systems was made based on name, date of birth, race and sex. No searches were made for persons who specifically refused to provide an SSN to the SIPP interviewer.

[6] Married couples were more precisely defined as married couples as of March 1991. Since the validation of SSNs was restricted to only those persons interviewed in Wave 1 of the panel, the universe of married couples was also restricted to those couples where both the husband and wife were interviewed in the initial interview.

[7] For the most part, married, filing joint tax returns contain the social security numbers of both spouses and the SSN of the primary filer is validated (SSNs are categorized as either the primary and the secondary SSN for these returns). In some cases, however, the secondary SSN is missing. These cases were eliminated even though a match was found based on the primary SSN.

4.2 Different Wage and Salary Concepts

Tax return and SIPP wage and salary income concepts differ on several counts. Wages and salary amounts reported on tax returns may not reflect the actual amount earned. Several significant concerns arise here. First, the SIPP concept, at least on paper, is defined as the gross amount earned. This would include amounts earned, but deferred from current taxation. The amount reported on tax returns, however, excludes these deferred earnings amounts[8]. Second, wage and salary income for tax purposes can include some types of pay in-kind, none of which is included in the SIPP definition. While the inclusion of the in-kind amounts on tax returns leads to overstatement of the "true" error, the number of such cases is small. On the other hand, the large proportion of persons now having deferred earning plans must lead to an understatement of the level of error measured in this study unless deferred earnings amounts are largely excluded from the amounts reported in SIPP as well.

4.3 Sampling Frame, Population Undercount and Survey Undercoverage

This analysis is restricted to errors attributable to survey reporting problems. Errors related to (1) sampling frame problems, (2) Census of Population undercounts and (3) differential undercoverage of population subgroups by the survey have not been examined. Since we know that there are problems in these areas and in the weighting methodologies employed to adjust for these problems, the picture of total error is not complete without an evaluation their contribution.

5. ASSESSMENT OF NONSAMPLING ERRORS

In considering nonsampling errors in SIPP I have chosen to explore a broad, but basic range of measures in order provide a picture of the error problem. Because missing data problems contribute a major portion of the total level of error, most information presented in this paper is shown separately by reporting status. Two groups were defined based on reporting status. The first, termed "complete", includes only those cases with complete data for all 12 months of calendar-year 1990 (*i.e.* no months of noninterview and no item nonresponse for wage and salary amounts). These cases account for about 79 percent of the total. The second, accounting for the remaining 21 percent of all cases, were termed "incomplete". This group consists of cases with one or more months of noninterview during the year or with one or more item nonreponses for the monthly amounts of wage and salary income.

A restriction was placed on the size of wage and salary amounts in order to reduce the effect of "outliers". In the early stages of the study it was found that a small number of cases having very large wage and salary amounts reported on tax returns (but much smaller amounts reported in the survey) seemed to unduly distort the error picture. For this reason I limited all amounts of wage and salary income to be no greater than the largest amount reported in the survey. This amount was approximately $700,000[9]. All measures shown in this paper, therefore, reflect this adjustment. It affected only a small number of tax return amounts.

5.1 Differences in Basic Summary Measures

5.1.1 Total Universe

The data in table A summarize the downward biases in the SIPP wage and salary distribution relative to the distribution derived from tax returns. A 7-percent underestimate of the mean and a 5-percent underestimate of the median are noted. There also appears to be a slight downward bias of about 1 percent in the number of wage and salary recipients. The combined effect of the SIPP underestimates of wage recipients and amount leads to an underestimate in the aggregate wage amount of about 8 percent.

[8] About 20 percent of the cases reported participating in a plan which deferred earnings. The mean amount deferred was about $2,800.

[9] Rather than eliminate these "outliers" I chose to include them but limit the amount. Results based on a strategy that eliminates these cases were very similar to those shown in this paper.

Downward bias in the variance of the SIPP-based wage and salary distribution appears to be more serious. The variance of the wage and salary distribution based on SIPP is only about 47 percent of the variance of the comparable distribution based on tax returns[10]. This large difference in variances results mainly from differences at the high end of the distribution, even after the $700,000 limit has been applied. In examining the sensitivity of variances to very high wage amounts it was found that rapid deterioration occurs above the $400,000 level. For cases having tax amounts below $200,000 the SIPP variance is about 90 percent of the tax return variance. Lifting the restriction on the universe to $400,000 produces only a slight decline in the ratio to about 86 percent. At $700,000, however, the ratio declines to the 47-percent figure reported in table A.

The simple correlation coefficient relating the SIPP and tax return amounts was .754. This correlation coefficient proved to also be very sensitive to the large wage and salary amounts, declining rapidly when the restrictions on the outliers are removed.

5.1.2 Complete Cases

Downward biases are also evident for the universe of complete cases. For the mean and median the biases are somewhat smaller than those of the total universe (about 4 percent for both the mean and median). The downward bias in numbers of recipients, however, appears slightly larger and that of the variance only marginally smaller. In terms of aggregate wages, the complete cases were about 7 percent below the tax return aggregate. The correlation coefficient between the SIPP and tax return amounts was about 8 percentage points higher (.834) than for the overall universe.

5.1.3 Incomplete Cases

Aside from a slight upward bias in wage and salary recipients, the universe of incomplete cases displays much larger downward biases than those mentioned for the complete cases. Indicated in table A are SIPP mean and median amounts that are biased downward by 15 and 11 percent, respectively. Turning to the variance comparison one finds the situation relative to complete cases to be only somewhat worse, the SIPP variance being about 44 percent of the corresponding tax return wage variance.

An examination of the tax return summary measures for the complete and incomplete universes provides some solid evidence that respondents and nonrespondents differ when it comes to levels of wage and salary income. The mean amount and variance are much larger (an 18-percent higher mean and a variance that is 2 times larger) for the incomplete group consisting of imputed and partially interviewed cases.

Associated with these differences is the apparent failure to achieve one of the goals of the imputation process. That goal is to reduce the bias due to item nonresponse. The mean of the imputed survey wage and salary amounts for incomplete cases is about 5 percent higher than the mean for complete cases, but the tax return mean for the incomplete cases, as indicated above, exceeds the tax return mean for complete cases by 18 percent. Hence, less than one-third of the 18-percent gap, as measured by the mean, was filled by the imputation procedures.

5.2 Distributional Effects of Nonsampling Error

Comparing summary measures clearly indicates that measurement error is not evenly distributed. The distances between SIPP and tax return means and medians differ and large differences in the variances of their distributions are evident. In order to assess the net effect of these errors on the distribution of wages I have ranked cases separately based on the size of the SIPP and tax return wage and salary amounts and examined various characteristics of the wage deciles generated from this ranking process. Table B provides comparisons of decile "cutoffs" and decile shares (two commonly used measures of inequality, the Gini coefficient and the variance of the ln of wages, are also shown).

[10] If no adjustment is made for outliers the ratio of the SIPP variance to the tax return variance is only about .20.

Table A: Summary Comparisons of the Wage and Salary Data from SIPP and Matching Tax Returns for 1990.

(Matched married joint returns with wage and salary income from specified source)

ALL CASES

Measure	Tax Returns	SIPP	Percent Difference
Number of cases	5,558	5,540	-0.9
Mean amount	$43,630	$40,460	-7.3
Median amount	$37,640	$35,760	-5.0
Variance	1.933e9	9.016e8	.466
Maximum value	+$2,000,000	+$700,000	.350
Aggregate wages (in thousands)	$243,800	$224,140	-8.1

COMPLETE CASES

Measure	Tax Returns	SIPP	Percent Difference
Number of cases	4,409	4,326	-1.9
Mean amount	$42,060	$40,020	-4.3
Median amount	$37,190	$35,810	-3.7
Variance	1.446e9	7.874e8	.545
Maximum value	+$800,000	+$400,000	.500
Aggregate wages (in thousands)	$185,400	$173,100	-6.6

INCOMPLETE CASES

Measure	Tax Returns	SIPP	Percent Difference
Number of cases	1,179	1,214	3.0
Mean amount	$49,510	$42,010	-15.2
Median amount	$39,720	$35,490	-10.6
Variance	3.342e9	1.472e9	.440
Maximum value	+$2,000,000	+$700,000	.350
Aggregate wages (in thousands)	$58,400	$51,000	-12.6

5.2.1 Decile Cutoffs

Overall the SIPP appears to slightly overestimate levels of wages and salary income at the low end of the distribution, but underestimate them at other points. This can be seen by examining the cutoffs in table B. Here the SIPP cutoffs for the lowest two deciles are above the tax return cutoffs and then lower for all deciles beginning with the third. The underestimates in all cutoffs above the second decile are consistent in size, each falling between 4 and 5 percent.

The pattern of first overestimation, then under estimation described above for the total universe also holds true for the group consisting solely of complete cases. For this universe the range in underestimation in decile 3 and higher were found to range between 2 and 5 percent.

The decile cutoff data for incomplete cases, on the other hand, do not follow the pattern of initial overestimation followed by underestimation. For this group underestimation is found for all deciles. Underestimation is lowest for the lower deciles rising from about 4 percent for decile 1 and reaching 11 percent at the cutoff for the highest decile.

5.2.2 Decile Shares

Without exception the SIPP overestimates the share of wages received in all but the top decile. Significant underestimation occurs in the SIPP top decile reflecting the cumulative imbalance generated by overestimates at all lower groupings. Downward biases in the share received by the top decile are substantial, about 11 percent overall, 9 percent for complete cases, and 15 percent for the incomplete universe.

Table B: Income Cutoffs and Shares for Deciles of Wage and Salary Income for SIPP and Matching Tax Returns for 1990.

(Matched married joint returns with wage or salary income from specified source)

	REPORTING STATUS					
	Total		Complete		Incomplete	
Decile	Tax Returns	SIPP	Tax Returns	SIPP	Tax Returns	SIPP
1	$9,499	$10,462	$8,859	$10,240	$11,861	$11,275
2	18,684	18,902	17,847	18,672	21,287	19,593
3	25,542	24,975	25,040	24,968	27,262	24,986
4	31,888	30,942	31,452	30,242	33,356	30,022
5	37,637	35,759	37,185	35,805	39,716	35,488
6	43,617	41,488	43,197	41,778	45,097	40,286
7	50,260	48,171	49,893	48,673	51,978	46,523
8	59,552	56,845	59,005	57,416	62,550	55,676
9	75,510	71,585	74,241	71,495	80,358	71,726
Decile Shares						
Total	100.0	100.0	100.0	100.0	100.0	100.0
1	1.0	1.4	1.0	1.3	1.4	1.6
2	3.2	3.7	3.2	3.7	3.4	3.8
3	5.1	5.4	5.1	5.5	4.9	5.3
4	6.6	6.8	6.8	6.9	6.2	6.6
5	7.9	8.1	8.1	8.2	7.4	7.9
6	9.3	9.5	9.5	9.7	8.5	9.2
7	10.7	11.0	11.0	11.2	9.8	10.4
8	12.5	12.9	12.9	13.1	11.5	12.2
9	15.2	15.7	15.6	15.9	14.1	15.0
10	28.5	25.5	26.8	24.5	32.8	28.0
Gini	.390	.358	.381	.353	.417	.373
VarLn	1.180	.787	1.290	.822	.780	.651

5.2.3 Wage Inequality

Survey estimates for two widely used measures of wage inequality, the Gini coefficient and the variance of the natural log of wages are both biased downward. Overall, the biases are about 8 percent for the Gini and 33 percent for the variance of the log. This 33-percent understatement of the variance of the log is considerably less than that indicated earlier for the standard variance computation and reflects the "compressing" effects of the transformation.

5.3 Measures of Classification Error

Found in table C are distributions of error expressed in terms of the distance between the SIPP decile and the tax return decile. The table also shows the proportion of cases within each distance group by reporting status. More than half (53 percent) of all cases were classified into the correct decile based on their SIPP amount. This rate of accuracy is lower (40 percent) for the incomplete group where classifications depend on the amounts assigned in the SIPP imputation procedures. Another 33 percent of the SIPP cases were classified as either one decile lower or one decile higher than their tax return position. Combining these two groups yields a total of 86 percent of all cases classified within a distance of one decile of their tax return decile.

Classification accuracy is related to reporting status as shown in the right-hand portion of table C. The proportion of the total within each decile distance category attributable to incomplete cases rises as the distances between SIPP and tax return deciles increase.

Table C: Summary of Decile Classification Differences for SIPP and Tax Return Wage and Salary Income: 1990.

(Matched married joint returns with wage or salary income from both sources)

	REPORTING UNIVERSE					
	Total	Complete	Incomplete	Total	Complete	Incomplete
Distance in Deciles Between SIPP and Tax Return						
Number	5,425	4,248	1,177	5,425	4,248	1,177
Percent	100.0	100.0	100.0	100.0	78.3	21.7
Same Decile	53.3	56.7	40.4	100.0	83.6	16.4
1	33.0	32.6	34.1	100.0	77.3	22.7
2	8.2	7.3	12.0	100.0	68.9	31.1
3	2.5	1.7	5.4	100.0	52.9	47.1
4	1.5	0.8	3.9	100.0	39.3	60.7
5	0.7	0.5	2.2	100.0	50.0	50.0
6	0.4	0.2	0.8	100.0	42.9	57.1
7	0.3	0.2	0.6	100.0	52.9	41.1
8	0.2	0.1	0.3	100.0	33.0	37.0
9	(z)	0.0	0.2	100.0	0.0	100.0

(z) Less than .05 percent.

5.4 Dissecting the Error in Aggregate Wages

It was established in table A that the aggregate wage and salary income amount based on SIPP was about 8 percent lower than the aggregate from tax returns. Incomplete cases contribute a disproportionately large share of the net downward bias. While making up only about 21 percent of the cases, they account for 38 percent of the aggregate survey underestimate.

Even though the net effect of reporting error is clearly a downward bias, many SIPP cases have wage and salary amounts that exceed their matching tax return amounts. For both the complete and incomplete universes the proportion of cases having a higher SIPP than tax return amount was about 38 percent. Net overreporting amounted to about $16.3 million. The mean level of overreporting was about $5,800 for complete cases but nearly $14,000 of the incomplete universe. Cases with lower SIPP amounts exceed those with higher estimates in both the number and extent of error (difference between the SIPP and tax return amount). The 62-percent with lower SIPP amounts had a mean error of $8,000 for complete cases and $18,000 for incomplete cases.

Table D: Selected Measures of Error in the Wage and Salary Income Data from the 1990 SIPP Panel: 1990.

(Matched married joint returns with wage or salary income from both sources)

REPORTING STATUS

Measure of error	Total	Complete	Incomplete
Mean Difference	$-3,920	$-2,980	$-7,310
Median Difference	$-1,230	$-1,020	$-2,440
Mean relative difference(%)	30.0	34.2	15.1
Median relative difference(%)	-3.8	-3.4	-7.3
Mean absolute difference	$8,970	$6,980	$16,100
Median absolute difference	$3,660	$3,240	$6,090
Squared sum of differences (SSD)	4.664e12	2.095e12	2.569e12
Index of Inconsistency[1]	.442	.320	.653

[1] defined as (((sipp-tax return)^2/n))/var tax return)

Table E: Cumulative Distribution of Relative Absolute Differences Between SIPP and Tax Return Wage and Salary Income: 1990.

(Matched married joint returns with wage or salary income from both sources)

REPORTING STATUS

Percent Error	Total	Complete	Incomplete
Total	100.0	100.0	100.0
Less than 1 percent	3.1	3.4	1.9
Less than 2 percent	8.4	9.3	5.2
Less than 3 percent	14.1	15.7	8.5
Less than 4 percent	19.8	33.1	11.7
Less than 5 percent	25.1	27.8	18.5
Less than 10 percent	48.9	53.1	33.6
Less than 15 percent	62.1	66.6	45.7
Less than 20 percent	71.8	76.2	55.8
More than 20 percent	28.2	23.8	44.2

Errors in recipiency of wage and salary amounts serve to reduce rather than increase the downward bias in the survey aggregate. An upward bias is generated because the aggregate attributable to the false survey positives is more than twice as large as the aggregate tax return amount for those survey cases reporting no wage amount.

5.5 Looking at the Errors Themselves

To this point the main focus of the paper has been on the net effects of measurement error. The size and distribution of the errors themselves are also of importance in developing a full-screen picture of the error problem. This section is, therefore, devoted to profiling the errors.

5.5.1 Broad Measures

In attempting to profile the characteristics of the errors there are many statistics that could be examined. A number of these are included in table D and each seems to provide a somewhat different perspective. For example, the mean difference between SIPP and tax return wage and salary amounts was -$3,920. This measure reflects the net effect of both overreporting and underreporting. Shifting to a measure based on the absolute value of the difference, which eliminates the offsetting effect of over- and underreporting, the mean error appears more than twice as large at $8,970. Error measures for the incomplete universe show the significantly larger error problem associated with these cases.

Table F: Shares of Relative Absolute Differences by Decile for SIPP Wage and Salary Income: 1990.

(Matched married joint returns with wage or salary income from both sources)

	REPORTING STATUS		
	Total	Complete	Incomplete
Difference Decile			
1	0.2	0.1	0.3
2	0.5	0.4	1.0
3	0.9	0.8	0.7
4	1.3	1.1	2.5
5	1.8	1.6	3.4
6	2.5	2.1	4.5
7	3.3	2.8	5.9
8	4.4	3.7	8.0
9	6.7	5.4	12.0
10	78.4	90.0	60.7

5.5.2 Distribution of Error Size

One simple and very descriptive measure of error is the distribution of errors by their size (in this case relative absolute size). This measure can be found in table E. An examination of this table reveals that for about 25 percent of the cases the SIPP amount was within +-5 percent of the tax return amount and for about half of the cases the survey and tax return amounts differed by less than 10 percent. The agreement indicated by the complete cases is significantly higher than that shown by the incomplete cases.

5.5.3 Error Shares

Given some distribution of errors it seems legitimate and informative to have a measure of the concentration of such errors. Error concentrations were computed for the relative absolute difference between SIPP and tax return amounts. The results are presented in table F.

For the total universe a high degree of concentration is exhibited for this measures with nearly 80 percent of the aggregate error contributed by those cases in the top 10 percent of error distribution. Errors are even more concentrated in the top 10 percent of the complete cases with 90 percent of the aggregate error centered there. In contrast, the aggregate error for the incomplete universe is significantly less concentrated at the high end.

6. CONCLUSIONS AND RECOMMENDATIONS

The findings presented in paper are based on a linkage between wage and salary data collected in the 1990 SIPP panel and that reported by SIPP respondents on their federal individual income tax returns for 1990. This analysis of the linked data has indicated that the SIPP wage and salary data are biased downward when compared to information reported on tax returns. In terms of the total amount of wage and salary income received by married couples, this bias amounts to a net SIPP underestimate of about 8 percent.

This simple statistic, however, fails to reveal the complex nature of the measurement error problem. Four main dimensions of the error problem arise from this evaluation. The first dimension is one typified by "soft" errors, *i.e.* those attributable to some form of simple response error where the respondent reports an amount which is slightly smaller or larger than the "true" amount. The second dimension relates to more serious problems characterized by "hard" errors. Hard errors are those that result in a serious misclassification of the respondent's position within the wage and salary distribution. Such errors might result in a shift of 2 or more deciles in the distributional location of such cases. The third dimension of error is that related to missing data and the subsequent system for imputing values for the survey nonresponses. The final dimension of error is that resulting in the inadequate representation of the upper tail of the income distribution and the effect this problem has on measures of variance and wage inequality.

Most responses in SIPP suffer from the kinds of errors that can be thought of as soft errors. Indeed, the downward bias in the SIPP median wage and salary income is only 4 percent for those cases without nonresponse problems. Comparing tax returns and reported survey values shows that in over half of the cases the survey and tax return amount differed by less than 10 percent (and two-thirds by less than 15 percent). A measure of classification error based on distributional position shows that the survey-based decile and the tax-based decile were identical for 57 percent of the cases and that nearly 90 percent of all cases were within +-1 decile of that computed using the tax return amount.

The measure of error concentration indicates that much of the aggregate measurement error can be attributed to a small number of cases. For the most part, these are the cases exhibiting the hard error problems that result in significant classification error with respect to their position in the wage distribution. It was found that those cases within the highest error decile contributed about 90 percent of the aggregate amount of error. While this analysis made no attempt to examine the causes for measurement error, more resources should be allocated to this end. Some research is underway at the Bureau of the Census in the alternative measurement design program (Moore *et al.*). Hopefully, this will provide some insights into the problems identified here. In addition, it would be useful to examine the questionnaires for those cases displaying the largest errors. Such an investigation would be an inexpensive way to learn how these errors originated.

One of the most important findings contained in this paper is that regarding the level of error associated with missing data and subsequent imputation. Imputation for missing data is a major contributor to the overall error problem in SIPP. About 21 percent of the study universe had either 1) some survey nonresponse to wage and salary amounts or 2) some months in a total noninterview status (missing wave situation). This group contributed nearly 40 percent of the net 8-percent underestimate of aggregate wage and salary income. Using other measures of error, such as the sum of squared differences, indicates that nearly 90 percent of total error can be attributed to the subuniverse of incomplete cases. Based on these results, it appears clear that an investigation of the imputation system should be undertaken. Linkages between surveys and administrative records systems provide a unique environment in which imputation systems can be developed and evaluated.

Household surveys have traditionally had very great difficulty in providing accurate estimates of the upper tail of income distributions and SIPP is no exception. Comparison the SIPP and tax return variances for the wage and salary income distribution indicate that the SIPP distribution seriously understates the variance of wages and that this understatement is related mainly to the upper reaches of the distribution. The understatement of variance can be almost entirely attributed to SIPP's failure to capture the very highest portion of the wage distribution. For wage and salary levels below $200,000, the SIPP-based estimate of variance is biased downward by only about 10 percent. For amounts below $700,000 the downward bias rises to more than 50 percent and considering all cases without restriction the bias is 80 percent. Since the number of cases underlying this study is relatively small, further investigation of the downward bias in the SIPP variance is warranted. At the very least, one should compute the variance of the wage and salary income distribution derived from the Statistics of Income (SOI) public-use microdata file for 1990 in order to obtain a reliable estimate of the variance for all tax returns, not just those matched to the SIPP universe used in this study.

In addition to the continued research mentioned above, there are several other areas of research that should be initiated. First, this evaluation of data quality should be repeated based on a similar linkage between the SIPP and the total compensation data that will soon become accessible by the Social Security Administration. Linkage between SIPP and these data would permit a much more thorough examination of reporting errors than can be achieved using tax returns since the linkage would be on a person rather than a tax return basis. Thus, husbands and wives could be treated separately rather than as a unit. Use of this file would also eliminate the problem of incompatibility caused by the fact that the amounts reported on tax returns exclude deferred earnings since total compensation amounts reflect earnings levels for purposes of computing social security payroll taxes.

Second, it would be useful to extend the research presented here to determine the level of error observed for various population subgroups. For example, what age, work experience, education, class of worker, occupation, *etc.* classifications have the largest measurement error problems.

Third, research on post-stratification weighting based on tax return information should be renewed. Work several years ago at the Bureau of the Census showed that significant reduction in sampling variances and some reduction in bias could be achieved by using weighting controls derived from tax return information.

Finally, in perhaps a somewhat longer run project, an important unknown in the measurement error picture, that related to household noninterviews, should be explored. Currently, crude weighting adjustments are applied without any stratification for socio-economic status.

REFERENCES

Brownstone, D., and Valletta, R. (1992). Modeling measurement error bias in cross-section and longitudinal wage equations. Paper presented at the 1992 Bureau of the Census Annual Research Conference.

Coder, J.F. (1990). Exploring nonsampling errors in the wage and salary data from the March current population survey. Paper presented at the 1990 Allied Social Sciences Association/Society of Government Economists meetings.

Groves, R.M. (1989). *Survey Errors and Survey Costs*, New York, John Wiley and Sons.

Herriot, R.A., and Spiers, E.F. (1980). Measuring the impact on income statistics of reporting differences between the current population survey and administrative sources. *Studies from Interagency Data Linkages*, Report No. 11, U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, SSA Publication Number 13-11750, (March).

Lillard, L., Smith, J.P., and Welch, F. (1986). What do we really know about wages?" The importance of nonreporting and census imputation. *Journal of Political Economy*, 94, 3, Part 1, 488-506 (June).

Moore, J.C., Bogan, K.E. and Marquis, K.H. (1992). A cognative interviewing approach for the survey of income and program participation: development of procedures and initial test results. Paper presented at Statistics Canada Symposium 92 (November).

Scheuren, F.H., Oh, L., Vogel, L., and Yuscavage, R. (1981). *Studies from Interagency Data Linkages*, Report No. 10., U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, SSA Publication Number 13-11750 (January).

Scholz, J.K. (1990). The participation rate of the earned income tax credit. La Follette Institute of Public Affairs (August).

United States Bureau of the Census (1991). P-60 N0. 174, *Money Income of Households, Persons and Families in the United States: 1990* (August).

Vaughan, D.R. (9189). Reflections on the income estimates from the survey of income and program participation. *ORS Working Paper Number 39*, U.S. Department of Health and Human Services, Social Security Administration, Office of Research and Statistics (September).

SPECIAL INVITED LECTURE

Proceeding of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

ESTIMATORS FOR LONGITUDINAL SURVEYS WITH APPLICATION TO THE U.S. CURRENT POPULATION SURVEY

W.A. Fuller[1], A. Adam and I.S. Yansaneh

ABSTRACT

Estimation for surveys conducted on repeated occasions with partial overlap of sampling units is investigated. Time-in-sample effects are recognized in the construction of some procedures and implementation for surveys with a large number of characteristics is described. Alternative estimators of employment characteristics based on the U.S. Current Population Survey are compared.

KEY WORDS: Rotation scheme; Covariance structure; Best linear unbiased estimator.

1. INTRODUCTION

We shall consider estimation for surveys that are conducted at several points in time. Duncan and Kalton (1987) discuss different types of repeated surveys and the objectives of such surveys. We are interested in cases where repeated determinations are made on some elements of the sample, but not every element appears in the sample at every time point.

An early study describing the use of least squares to incorporate information from a previous occasion into the estimate of the current occasion is that of Jessen (1942). Patterson (1950) investigated estimation for rotating samples. This work was followed by a number of authors, including Eckler (1955), Rao and Graham (1964), Gurney and Daly (1965), Raj (1965), Singh (1968), Wolter (1979), Huang and Ernst (1981), and Kumar and Lee (1983). These authors treated the unknown quantities at each occasion as fixed parameters.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Smith (1978), and Jones (1979) considered estimation under the assumption that the underlying true values are the realization of a time series.

We will discuss estimation for the U.S. Current Population Survey. In this research, we develop a model for the covariance structure of observations made in the Current Population Survey, estimate the parameters of this model for two important characteristics of the labor force, and investigate alternative estimation procedures. We shall consider the unknown true values to be fixed parameters.

2. THE CURRENT POPULATION SURVEY

The U.S. Current Population Survey is a nationwide survey designed to produce estimates by states as well as for the nation. The sample is a stratified area sample with about 717 strata in 50 states. Of the strata, 384 contain more than one primary sampling unit (PSU). PSU's are areas of land and in the 384 strata, one PSU is selected for observation. In the terminology of the Census Bureau, the remaining 333 strata are self-representing primary sampling units. In a more standard textbook description, the primary sampling units in the 333 strata are subareas of the larger geographic subdivisions. These subareas are smaller units than the PSU's in the other 384 strata. The Current Population Survey is a large survey with about 57,000 households interviewed per month, and about 113,000 individuals interviewed per month.

[1] W.A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA 50011. U.S.A.

The survey is designed so that individuals respond more than one month. A particular group of individuals are brought into the sample, interviewed for four months, given an 8-month furlough, and then interviewed for four months more. After the second four months, the individuals rotate out of the sample forever. At any particular point in time, there are 16 groups under consideration, 8 of them are being interviewed, while 8 are on furlough. Of those being interviewed, one is being interviewed for the first time, one for the second, *etc*. As a result, the sample is balanced on the number of times that individuals have been interviewed. There are a number of complex operations in the data collection and estimation schemes, including, for example, a raking adjustment to population totals. We will not consider those operations in our discussion.

The basic data for our study were composed of two parts. The first is a set of 48 replicates for the 12 months of 1987. These replicates were constructed under a scheme designed by Fay (1989). Essentially, weights are assigned to the PSU's in a balanced way so that each set of weights give an unbiased estimate of the total. The squared difference between the estimate based on any of the replicates and the overall estimate is an estimate of 1/4 of the variance of the overall estimate. The replicate observations are for each of the 12 months and for each of the 8 time-in-sample group. There are 12 months for the 8 times-in-sample on 48 replicates. Thus, the 1987 data set contains 4,608 observations.

Table 1 contains a representation of the data. We call the columns of the table *streams*. The first entry for the first month is for individuals that are being interviewed for the first time. That is, A_{11} denotes the set of individuals being interviewed for the first time in month one. They are interviewed for the second time in month two. The first entry in the second stream is for individuals that are being interviewed for the second time in month one. If we move down the second column, those individuals that were being interviewed for the second time in month one are interviewed for the fourth time in month three. Then they rotate out and a new group comes in. In this case, it is Group E that is being interviewed for the fifth time in month four. Group E is interviewed for four months, then they rotate out and a new group comes in. In this case, it is Group F that is being interviewed for the first time in month eight. Thus, the rotation groups appear in a single stream. They rotate in and then out in accordance with the 4-8-4 rotation scheme.

Table 1: Data Arrangement for 1987 data.

Month	Streams							
1	$A_{1,1}$	$D_{1,2}$	$G_{1,3}$	$J_{1,4}$	$M_{1,5}$	$P_{1,6}$	$T_{1,7}$	$X_{1,8}$
2	$A_{2,2}$	$D_{2,3}$	$G_{2,4}$	$K_{2,5}$	$M_{2,6}$	$P_{2,7}$	$T_{2,8}$	$Y_{2,1}$
3	$A_{3,3}$	$D_{3,4}$	$H_{3,5}$	$K_{3,6}$	$M_{3,7}$	$P_{3,8}$	$U_{3,1}$	$Y_{3,2}$
4	$A_{4,4}$	$E_{4,5}$	$H_{4,6}$	$K_{4,7}$	$M_{4,8}$	$Q_{4,1}$	$U_{4,2}$	$Y_{4,3}$
5	$B_{5,5}$	$E_{5,6}$	$H_{5,7}$	$K_{5,8}$	$N_{5,1}$	$Q_{5,2}$	$U_{5,3}$	$Y_{5,4}$
6	$B_{6,6}$	$E_{6,7}$	$H_{6,8}$	$L_{6,1}$	$N_{6,2}$	$Q_{6,3}$	$U_{6,4}$	$Z_{6,5}$
7	$B_{7,7}$	$E_{7,8}$	$I_{7,1}$	$L_{7,2}$	$N_{7,3}$	$Q_{7,4}$	$V_{7,5}$	$Z_{7,6}$
8	$B_{8,8}$	$F_{8,1}$	$I_{8,2}$	$L_{8,3}$	$N_{8,4}$	$R_{8,5}$	$V_{8,6}$	$Z_{8,7}$
9	$C_{9,1}$	$F_{9,2}$	$I_{9,3}$	$L_{9,4}$	$O_{9,5}$	$R_{9,6}$	$V_{9,7}$	$Z_{9,8}$
10	$C_{10,2}$	$F_{10,3}$	$I_{10,4}$	$J_{10,5}$	$O_{10,6}$	$R_{10,7}$	$V_{10,8}$	$\Gamma_{10,1}$
11	$C_{11,3}$	$F_{11,4}$	$G_{11,5}$	$J_{11,6}$	$O_{11,7}$	$R_{11,8}$	$W_{11,1}$	$\Gamma_{11,2}$
12	$C_{12,4}$	$D_{12,5}$	$G_{12,6}$	$J_{12,7}$	$O_{12,8}$	$S_{12,1}$	$W_{12,2}$	$\Gamma_{12,3}$

3. COVARIANCE STRUCTURE OF BASIC ESTIMATORS

We began our investigation by postulating an analysis of variance model for these data. Let

$$y_{tjk} = \mu + u_j + \alpha_t + \tau_k + \gamma_g + \zeta_{tk} + \epsilon_{tkj}, \tag{1}$$

where y_{tjk} is the observation at the t -th time, or the t -th month, on the j -th replicate for the k-th time-in-sample, μ is the overall mean, u_j is a replicate effect, α_t is a month effect, τ_k is a time-in-sample effect, γ_g is a rotation group effect, and ζ_{tk} are interaction effects. The subscript g, denoting a rotation group, is completely determined by the month t and time-in-sample k. See Table 1.

Table 2 contains an analysis of variance constructed with the data according to the model. The month effects dominate. There are definitely different levels in unemployment and in Civilian Labor Force in different months. There are also definite time-in-sample effects. The sum of squares for rotation group effects is adjusted for months and times-in-sample. After the removal of rotation group effects, there remain 52 degrees of freedom from the original 95 degrees of freedom. The mean squares for these degrees of freedom are given on the line titled "Interactions".

We are willing to assume that replicates, by their construction, are nearly independent. The correlation between observations on the same rotation group means that the assumptions required for classical F-tests for this table are not satisfied. However, most would be willing to conclude that there are major effects in this set of data and these effects are dominated by the month effects. The time-in-sample effect is also very important.

Table 2: Analysis of variance for employed, unemployed, and Civilian Labor Force, 1987.

Source	d.f.	Mean Squares		
		Employed	Unemployed	CLF
Replicates	47	1.2134	0.1785	1.0762
Months	11	3553.2435	268.4974	2377.7240
Time-in-Sample	7	458.1149	75.4759	891.2340
Groups[1]	25	113.4492	20.7709	91.5742
Interactions	52	12.9719	6.7783	11.7550
Residual	4465	0.2458	0.0554	0.2112

[1] The groups mean square is adjusted for month and time-in-sample.

In order to estimate the covariance structure of these data, we examine a subset of the effects defined in our original model. Let r_{gjk} denote the sum of the replication effect and the epsilon effect of our original model. We use the index g for group identification rather than time. As mentioned earlier, knowing g and k is equivalent to knowing t and k. We write

$$r_{gjk} = u_j + e_{gj} + a_{gjk}, \tag{2}$$

where u_j is the replicate effect, e_{gj} is the permanent rotation group effect, and a_{gjk} is the transient rotation group effect. The replicate effect is a reflection of the difference among primary sampling units. For purposes of this model, this effect is assumed to be constant over time. For example, the mean over time of certain primary sampling units is larger than that of other primary sampling units. The e_{gj} is a similar effect for rotation groups. The long run average for certain rotation groups is assumed to be larger than that of other groups. The final component of our model, denoted by a_{gjk}, is included to capture the correlation of observations on the same rotation group that tends to decrease as the time between observations increases. Under our model, no matter when we observe the j -th replicate, we would get a u_j. No matter when we observe this particular group of individuals, we would get an e_{gj}. But there is also an effect associated with a particular group that dies out with time. That is, an individual has some overall propensity to be employed, for example. But if we observe that individual at two points close together in time, then the individual is more apt to be unemployed or employed at both of those observations. The a_{gjk} are used to represent the effect whose correlation declines as the time between observations increases.

We assume that the transient rotation group effect satisfies a third order autoregression,

$$a_{gjk} = \xi_1 a_{g,j,k-1} + \xi_2 a_{g,j,k-2} + \xi_3 a_{g,j,k-3} + b_{gjk}, \tag{3}$$

where

$$b_{gjk} \sim \mathrm{Ind}(0, \sigma_b^2).$$

Under our model, the correlation between observations at two points in time on the same rotation group is

$$\rho_r(h) = \frac{\sigma_u^2 + \sigma_e^2 + \rho_a(h)\sigma_a^2}{\sigma_u^2 + \sigma_e^2 + \sigma_a^2}, \tag{4}$$

where h is the distance apart of the two points and $\rho_a(h)$ is the autocorrelation of the a-effect. The variance of a randomly chosen observation is $\sigma_u^2 + \sigma_e^2 + \sigma_a^2$. The covariance between two observations in the same replicate but in different rotation groups is σ_u^2 for any h.

We can estimate the autocovariance for a rotation group. The autocovariance is an estimate of $\sigma_u^2 + \sigma_a^2 + \rho_a(h)\sigma_a^2$. That set of estimated covariances, alone, is not enough to separate the effects. The error line in the AOV of Table 2 estimates $\sigma_a^2 + \sigma_e^2$ and a separate AOV was used to estimate σ_u^2. Given estimates of these three quantities, we can solve for the other parameters.

Table 3: Average autocorrelations within a rotation group for 1987 CPS.

Lag	No. Obs.	Employed	Unemployed	Civilian Labor Force	Unemployment Rate
1	66	0.8068 (0.0062)	0.4979 (0.0136)	0.7876 (0.0068)	0.5187 (0.0132)
2	40	0.7332 (0.0106)	0.3788 (0.0199)	0.7197 (0.0111)	0.4019 (0.0195)
3	18	0.6856 (0.0182)	0.3230 (0.0312)	0.6668 (0.0192)	0.3484 (0.0306)
9	3	0.6732 (0.0461)	0.1566 (0.0832)	0.6377 (0.0504)	0.2034 (0.0818)
10	4	0.7191 (0.0354)	0.2691 (0.0686)	0.6187 (0.0452)	0.3159 (0.0665)
11	3	0.6038 (0.0536)	0.1401 (0.0830)	0.4910 (0.0614)	0.2138 (0.0814)
Ave. 9-11	10	0.6708 (0.0252)	0.1966 (0.0450)	0.5861 (0.0298)	0.2443 (0.0439)

Table 3 contains estimated autocorrelations for unemployment and the Civilian Labor Force. Autocorrelations are much higher for Civilian Labor Force than for unemployment. There are a limited number of rotation groups observed at lags 9, 10, and 11, so the average for these lags is presented in the table. The correlation is nearly 0.60 at lags 9 and 11 for Civilian Labor Force.

Table 4 contains the estimates of the parameters of the autoregressive process for a_{gk}. Table 5 contains the estimates of the other parameters of our model.

Figure 1 is a plot of the estimated autocorrelation for unemployed. This is the autocorrelation function for a particular rotation group. We only observed autocorrelations for periods zero through 3 and for periods 9, 10, and 11. All other autocorrelations given by the dots are estimated using our model. The squares on the figure are autocorrelations estimated by Breau and Ernst (1983) using data for 1976 and 1977. The estimated autocorrelations are very similar for the two periods.

Figure 1: Estimated Autocorrelation for Unemployed.

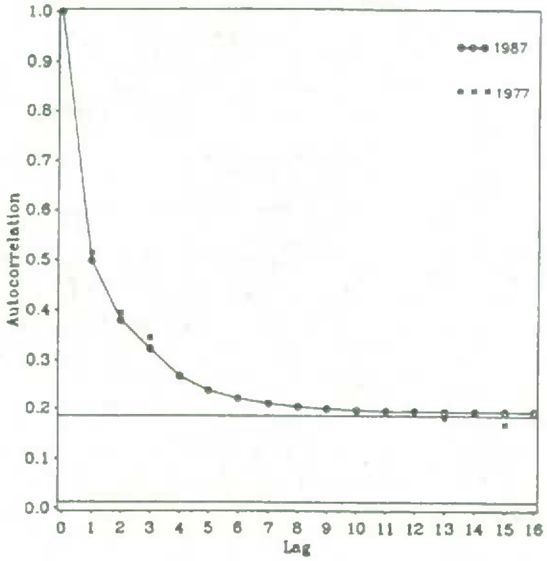


Figure 2: Estimated Autocorrelation for Civilian Labor Force.

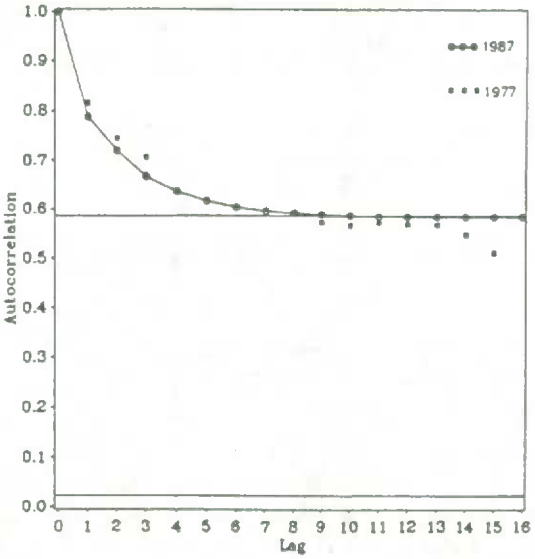


Figure 2 contains the analogous autocorrelation estimated for the Civilian Labor Force. Once again, the squares are estimates from Breau and Ernst based on 1976 and 1977 data. The Breau and Ernst estimates are a bit higher for short period autocorrelations and a bit lower for longer period autocorrelations.

Table 4: Estimates of parameters of transient processes a_{gjk}.

Model $a_{gjk} = \xi_1 a_{gj,k-1} + \xi_2 a_{gj,k-2} + \xi_3 a_{gj,k-3} + b_{gjk}$					
Characteristic	$\hat{\xi}_1$	$\hat{\xi}_2$	$\hat{\xi}_3$	$\hat{\sigma}_b^2$	$\hat{\sigma}_a^2$
Employed	0.40481	0.04270	-0.04945	0.06841	0.08263
Unemployed	0.33422	0.08452	0.05267	0.03831	0.04508
Civilian labor force	0.43415	0.11345	0.00318	0.06756	0.08962

Table 5: Estimates of σ_u^2, σ_e^2, and σ_a^2.

	Variance component			
Characteristic	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	Total
Employed	0.00552	0.16319	0.08263	0.25134
Unemployed	0.00065	0.01037	0.04508	0.05610
Civilian labor force	0.00499	0.12159	0.08962	0.21620

The lines in the figures identify the two permanent components. The bottom line is the line for σ_u^2 as a percent of the total variation. The next line is the sum $\sigma_u^2 + \sigma_e^2$ as a percent of the total. Under our model, even at long periods of time, there would be some correlation because of the permanent rotation group effect and the permanent primary sampling unit effect.

For Civilian Labor Force σ_u^2 and the $\sigma_u^2 + \sigma_e^2$ are a much bigger fraction of the total variation than for unemployed. Thus, the autocorrelation in Civilian Labor Force is sizable even at extremely long lags. The picture for employed is fairly similar to that for Civilian Labor Force. In fact, the long run correlation is higher for employed than for Civilian Labor Force.

The autocorrelation for unemployment rate is very similar to the correlation function for unemployed. The variation in the denominator, which is Civilian Labor Force, contributes little to the variation of the rate.

4. COMPARISON OF ALTERNATIVE ESTIMATORS

A simple estimator for the Civilian Labor Force at time t is obtained by averaging the estimates for the 8 rotation groups. We call this the direct estimator or the basic estimator and denoted it by $\bar{y}_{t0}$. Under our model, the variance of the direct estimator is

$$V\{\bar{y}_{t0.}\} = (64)^{-1}(64\sigma_u^2 + 8\sigma_e^2 + 8\sigma_a^2). \tag{5}$$

We assume that the rotation groups are independent within the primary sampling units and the primary sampling unit variation is reflected in σ_u^2, the replicate variance. The coefficients of variation for the direct estimators are about 0.3% for employed, about 1.6% for unemployed, and about 0.2% for Civilian Labor Force.

Table 6 gives the contribution to the variance of the direct estimator for the different components. The contributions are quite different for unemployed and for Civilian Labor Force. In the case of unemployed, about 9% of the variation is estimated to come from the replicate effect, while for Civilian Labor Force, the replicate contribution is about 16%. About 74% of the variance in unemployment comes from the transient rotation group effect while only 36% of the variance of Civilian Labor Force comes from the transient effect. The permanent rotation group effect is the most important component for the Civilian Labor Force. The permanent rotation group effect accounts for only 17% of the variance for unemployed.

Table 6: Variance of direct estimator based on eight streams.

Property	Employed	Unemployed	Civilian Labor Force	Unemployment Rate
Variance	9.2793	1.9408	8.0361	8.3994×10^{-5}
% due to $\hat{\sigma}_u^2$	15.22	8.60	15.89	8.56
% due to $\hat{\sigma}_e^2$	56.28	17.10	48.41	17.76
% due to $\hat{\sigma}_a^2$	28.50	74.30	35.70	73.68

The estimator presently used for the Current Population Survey is a weighted average of the direct estimator, a quantity that is the previous composite estimator plus an estimate of the change, and a third linear combination of the current individual rotation group estimates. The estimator is

$$\hat{\mu}_{tB.} = 0.6\bar{y}_{t0.} + 0.4(\hat{\mu}_{t-1,B.} + \hat{\delta}_{t,t-1} + 0.05[2^{-1}(y_{t01} + y_{t05}) - 6^{-1} \sum_{k=2}^{4} (y_{t0k} + y_{t,0,k+4})], \qquad (6)$$

where

$$\bar{y}_{t0.} = 8^{-1} \sum_{k=1}^{8} y_{t0k},$$

$$\hat{\delta}_{t,t-1} = 6^{-1} \left[\sum_{k=2}^{4} (y_{t0k} + y_{t,0,k+4}) - \sum_{k=2}^{4} (y_{t-1,0,k-1} + y_{t-1,0,k+3}) \right],$$

$\hat{\mu}_{tB.}$ is the estimator for period t, and $\hat{\delta}_{t,t-1}$ is an estimator of change constructed from the rotation groups observed in both period t and period $t-1$. The estimator used until 1985 contained only the first two terms. The third term in the estimator accomplishes several things. The third term is the difference between the average of the first and fifth rotation groups and the average of all other rotation groups. The term reduces the time-in-sample effects appearing in the original estimator. The first and fifth times-in-sample produce larger estimates of unemployed then do the other rotation groups. Therefore, the direct difference, $\delta_{t,t-1}$ is influenced by the fact that the first time-in-sample has a larger expected value than the second time-in-sample. The time-in-sample effects don't cancel in the difference estimate. The inclusion of the third term makes the expected value of the estimator closer to the expected value of the direct estimator. It also reduces the variance of the estimator relative to the two-part estimator used prior to 1985.

We now consider the best linear unbiased estimator of the current level. This estimator uses all data through time t to construct the best estimator at time t. In the tradition of the Current Population Survey, the estimator for time t will not be changed as more data become available.

In order to construct the best linear unbiased estimator, we show that it is not necessary to store all of the previous observations. However, all previous observations on any rotation group observed at time t are required. In the case of the Current Population Survey, there will be some observations as far as 15 months in the past that will appear in the estimator because if the rotation group is being observed for the last time, then that is the 16th month that that group has been associated with the Current Population Survey.

We illustrate the construction of the best linear unbiased estimator using a design in which 3 groups are observed at each time point. We assume that a group rotates in, is observed for 3 periods, and then rotates out forever. Under this simple design, the best estimators for $t-2$ and $t-1$ are used to construct the best linear unbiased estimator for time t. There are three types of observations at time t, those being observed for the first time, those for the second time, and those for the third time. It is assumed that groups are independent. Then the observation on the group being observed for the first time is independent of all other data. The observation for a group being observed for the second time is correlated with the preceding observation on that group. If we regress this second time period observation on the previous observation, we produce a deviation, denoted by w_{t2} that is uncorrelated with the previous observation. The linear combination created from the third time-in-sample observation, w_{t3} is independent of all earlier observations. There are five observations to be used in the estimator. It follows that the best linear unbiased estimator can be constructed from the linear model. In other words, all of the information up until the current time, relevant for estimating θ_t, is condensed in the two previous best estimators and the three transformed current observations. Thus, we have a linear model in $(\theta_t, \theta_{t-1}, \theta_{t-2})$, where, in terms of our earlier notation, $\theta_t = \mu + a_t$. Given the covariance matrix of the five estimators, we apply generalized least squares to construct our best estimator of the current level.

The linear model is

$$\begin{bmatrix} \hat{\theta}_{t-2} \\ \hat{\theta}_{t-1} \\ w_{t1} \\ w_{t2} \\ w_{t3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -b_{21} & 1 \\ -b_{32} & -b_{31} & 1 \end{bmatrix} \begin{bmatrix} \theta_{t-2} \\ \theta_{t-1} \\ \theta_t \end{bmatrix} + e_t,$$

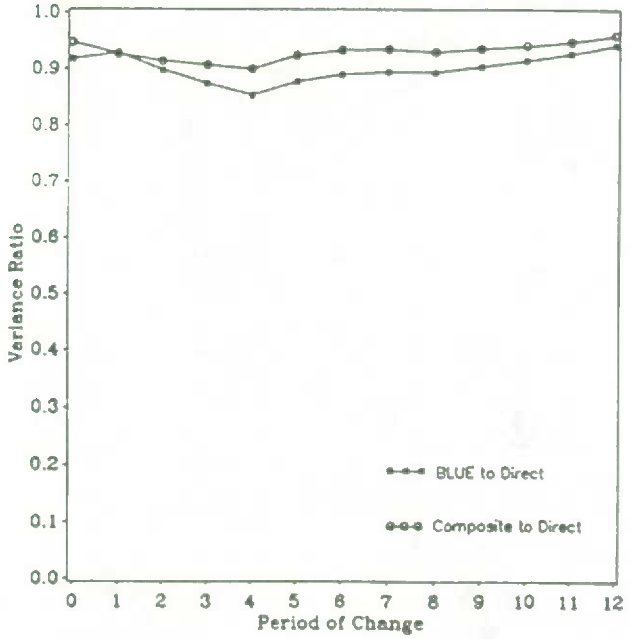
where, $w_{t1} = y_{t01}$, $w_{t2} = y_{t02} - b_{21} y_{t-1,0,1}$, $w_{t3} = y_{t03} - b_{32}\, y_{t-1,0,2} - b_{31}\, y_{t-2,0,1}$, e_t is the vector of differences between the observations and their expected values, and the b_{ij} are the population regression coefficients. The covariance matrix of $(\hat{\theta}_{t-2}, \hat{\theta}_{t-1})$ is known from the least squares fit constructed at time *t-1*. The vector (w_{t1}, w_{t2}, w_{t3}) is uncorrelated with $(\hat{\theta}_{t-2}, \hat{\theta}_{t-1})$ and has a diagonal covariance matrix determined by the covariance structure of the y_{t0j}.

In order to construct the best estimator at time t for the 4-8-4 rotation scheme, it is necessary to store 15 estimates for times $t-1, t-2, ..., t-15$, and to store 60 previous observations on the 15 rotation groups that have been observed previously. Although only 8 groups are observed at any point in time, a total of 16 groups are involved in the survey at that point. The variances of alternative estimators for unemployed, Civilian Labor Force and employed are given in Tables 7, 8, and 9, respectively.

The variances of alternative estimators for unemployed are compared in Figure 3. The line with the dots is the variance of the present composite estimator divided by the variance of the direct estimator for each of the periods of change, where zero denotes the estimator of current level.

The line with the squares is the variance of the best linear unbiased estimator (BLUE) divided by the variance of the direct estimator. The estimators that use previous information are only moderately superior to the direct estimator for unemployed. The present composite estimator of current level has a variance that is about 95% of the variance of the direct estimator and the variance of the least squares estimator is about 92% of the variance of the direct estimator. Both estimators utilizing previous information are about 7% superior to the direct estimator for one period change. The maximum gain is about 15% for the best linear unbiased estimator of four month change.

Figure 3: Ratios of Variances of Alternative Estimators to Variances of Direct Estimator for Unemployed.



The variance of alternative estimators of employed are compared in Figure 4.

Figure 4: Ratios of Variances of Alternative Estimators to Variances of Direct Estimator for Employed.

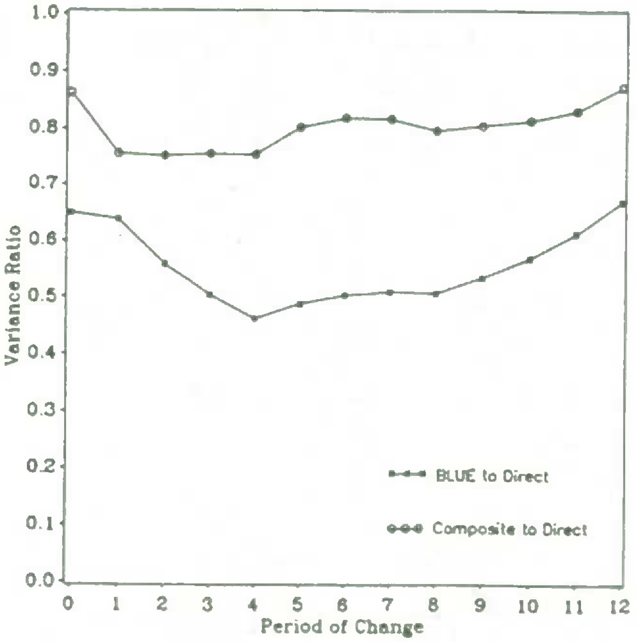


Table 7: Variances of alternative estimators of unemployed relative to the variance of the direct estimator of current level.

Period of Change	Direct Estimator	Present Composite	BLUE Current
0	1.000	0.947	0.918
1	1.155	1.070	1.073
2	1.490	1.361	1.338
3	1.686	1.528	1.473
4	1.830	1.645	1.562
5	1.830	1.691	1.606
6	1.830	1.708	1.628
7	1.830	1.710	1.636
8	1.830	1.701	1.634
9	1.787	1.671	1.614
10	1.745	1.641	1.595
11	1.703	1.614	1.578
12	1.660	1.593	1.564

The variance of the present composite estimator for the one period change for employed is about 86% of the variance of the direct estimator of one period change. The variance of one period change using the best estimator of current level is about 65% of the variance of the direct estimator of one period change. The best estimator for employed becomes even more superior relative to the direct and relative to the composite as the period of change increases. The variance of the best estimator of 4-period change is less than 50% of that direct estimator of change. The present composite estimator of 4-period change has a variance that is about 75% of the direct. At a period of 11, the current composite has a variance that is a about 84% of the variance of the direct while the best estimator has a variance that is about 68% of the variance of the direct.

Table 8: Variances of alternative estimators of Civilian Labor Force relative to the variance of the direct estimator of current level.

Period of Change	Direct Estimator	Present Composite	BLUE Current
0	1.000	0.868	0.704
1	0.697	0.538	0.474
2	1.086	0.828	0.652
3	1.410	1.076	0.774
4	1.688	1.283	0.859
5	1.688	1.364	0.914
6	1.688	1.392	0.946
7	1.688	1.391	0.962
8	1.688	1.362	0.963
9	1.565	1.277	0.942
10	1.445	1.191	0.921
11	1.324	1.111	0.903
12	1.203	1.052	0.887

The relative efficiency for the unemployment rate is very similar to that for unemployed. The Civilian Labor Force is somewhat similar to that for employed, but because the correlations are a bit higher for employed, the BLUE performs somewhat better for employed than the Civilian Labor Force.

Table 9: Variances of alternative estimators of employed relative to the variance of the direct estimator of current level.

Period of Change	Direct Estimator	Present Composite	BLUE Current
0	1.000	0.862	0.650
1	0.677	0.511	0.432
2	1.083	0.813	0.604
3	1.413	1.065	0.711
4	1.701	1.279	0.784
5	1.701	1.363	0.829
6	1.701	1.390	0.855
7	1.701	1.388	0.865
8	1.701	1.353	0.860
9	1.560	1.255	0.832
10	1.419	1.154	0.806
11	1.278	1.061	0.782
12	1.137	0.992	0.761

We also computed variances for linear estimators based on 12 months of data, 16 months of data, and 24 months of data. The maximum loss in efficiency from using 12 months of data relative to the use of the full past was 8% for the estimate of current level for employed. The maximum loss for the 16-month estimator was 3%. The maximum loss in efficiency from using 24 months of data relative to the use of the full past was 2%. Some estimators of long period change, such as year-to-year change had smaller variance when based on only 24 months of data.

The current rotation scheme for the Current Population Survey involves individuals being interviewed for 4 periods, being out of the sample for 8, and then back in for 4. The purpose of this design was to increase the efficiency of estimated year to year changes.

We compared estimates constructed with this rotation scheme to the rotation schemes in which individuals are in for 6 periods and then rotate out permanently and to the rotation scheme in which individuals are in for 8 periods and then rotation out permanently.

Table 10 contains the variances for alternative sampling estimation schemes for unemployed.

The efficiency of the estimator of current level of a scheme in which individuals are in continuously for 6 periods is slightly less than that of the 4-8-4 scheme for unemployed. Estimates for short period changes are slightly superior for the continuous 6, but estimates of long period changes are superior for the 4-8-4 scheme. A scheme in which individuals are in for 8 continuous periods and then out is also considered. For unemployed, there is a bit of a loss at current level for the continuous 8 relative to the 4-8-4 and then gains on an order of 5% for short period changes. The 4-8-4 is superior to continuous 8 for longer period changes, particularly year-to-year changes.

Table 10: Variances of alternative estimators of unemployed; variance of direct estimator of current level equals one.

Quantity estimated	Present composite estimator	Best estimator 4-8-4	Best estimator 8-in-then-out	Best estimator 6-in-then-out
Current level	0.947	0.918	0.944	0.938
1-period change	1.070	1.073	1.003	1.051
2-period change	1.361	1.338	1.250	1.312
3-period change	1.528	1.473	1.372	1.443
4-period change	1.645	1.562	1.473	1.543
5-period change	1.691	1.606	1.533	1.607
6-period change	1.708	1.628	1.577	1.655
7-period change	1.710	1.636	1.612	1.686
8-period change	1.701	1.634	1.642	1.705
9-period change	1.671	1.614	1.663	1.719
10-period change	1.641	1.595	1.678	1.727
11-period change	1.614	1.578	1.688	1.733
12-period change	1.593	1.564	1.696	1.737
12-period average	0.255	0.249	0.301	0.266
Change in 12-period averages	0.273	0.262	0.372	0.359

Table 11: Variances of alternative estimators of Civilian Labor Force; variance of direct estimator of current level equals one.

Quantity estimated	Present composite estimator	Best estimator 4-8-4	Best estimator 8-in-then-out	Best estimator 6-in-then-out
Current level	0.868	0.706	0.796	0.783
1-period change	0.538	0.474	0.430	0.470
2-period change	0.828	0.652	0.589	0.651
3-period change	1.076	0.774	0.709	0.786
4-period change	1.283	0.859	0.793	0.883
5-period change	1.364	0.913	0.858	0.962
6-period change	1.392	0.946	0.913	1.032
7-period change	1.391	0.961	0.963	1.088
8-period change	1.362	0.962	1.010	1.133
9-period change	1.277	0.941	1.049	1.170
10-period change	1.191	0.920	1.083	1.199
11-period change	1.111	0.902	1.111	1.223
12-period change	1.052	0.886	1.135	1.242
12-period average	0.369	0.346	0.448	0.396
Change in 12-period averages	0.259	0.206	0.393	0.413

There is a similar picture for Civilian Labor Force. There is a loss at current level for the continuous 8 relative to the 4-8-4. Under the continuous scheme, we are dealing with 8 groups at any point in time whereas, under the 4-8-4 scheme, we're dealing with 16 groups. There is some additional information in the additional 8 groups. On the other hand, the change estimates up through about 6 are superior for the continuous 8 scheme.

Table 12: Variances of alternative estimators of employed; variance of direct estimator of current level equals one.

Quantity estimated	Present composite estimator	Best estimator 4-8-4	Best estimator 8-in-then-out	Best estimator 6-in-then-out
Current level	0.862	0.653	0.761	0.759
1-period change	0.511	0.432	0.395	0.434
2-period change	0.813	0.604	0.559	0.619
3-period change	1.065	0.710	0.669	0.747
4-period change	1.279	0.783	0.731	0.829
5-period change	1.363	0.828	0.782	0.901
6-period change	1.390	0.854	0.828	0.970
7-period change	1.388	0.863	0.874	1.026
8-period change	1.353	0.858	0.960	1.071
9-period change	1.255	0.830	0.960	1.108
10-period change	1.154	0.803	0.993	1.139
11-period change	1.061	0.779	1.021	1.165
12-period change	0.992	0.758	1.046	1.186
12-period average	0.369	0.326	0.440	0.394
Change in 12-period averages	0.248	0.162	0.365	0.403

Estimates of change after period 8 are superior for the 4-8-4 scheme and the estimate of year-to-year change is about 20% superior for the 4-8-4 scheme. For employed, the 4-8-4 dominates the continuous 6 procedure.

Under the present estimation scheme, the same coefficients are applied to the direct estimates to construct the estimators of employed as are used to construct the estimates of unemployed. This means that the estimates are internally consistent. The expected value of each estimator is exactly the same linear combination of month effects and time-in-sample effects.

However, because the autocorrelation structure is different for employed than for unemployed and because the coefficients, approximately, minimize the variance for unemployed, the estimates for employed are inefficient.

The following procedure could be used to increase the efficiency of the estimates of employed.

1. Construct estimates of Civilian Labor Force that are best subject to the restriction that the expectation of the estimator is the same linear combination of time-in-sample effects as is the expectation of the estimator of unemployed.

2. Using the estimates of Civilian Labor Force and unemployed, construct weights for the current observations that reproduce the estimates of unemployed and Civilian Labor Force.

The conceptually optimum estimator for a characteristic *y* would use the past information on all characteristics. Operationally, this is not feasible. Our investigation has established that the cross correlations between unemployed and Civilian Labor Force are small. The first six cross correlations are estimated at less than 0.10 in absolute value. It follows that estimates based only on the past values of the characteristic being estimated, are nearly optimum for these two characteristics.

We investigated the behavior of time-in-sample effects by constructing linear combinations of the 8 basic time-in-sample estimates for each time point that are linear functions only of the population time-in-sample effects. Variation over the eleven years, 1980-1990, in the time-in-sample contrasts are approximately equal to the estimated variances of the contrasts. Therefore, the data give no reason to reject the hypothesis that time-in-sample effects were constant during that time period.

Despite this result, many practitioners would not be willing to adopt the model of constant time-in-sample effects for a period of many years. We present the variance of estimates constructed to have the same expectation as that of the direct estimator. However, estimators could be constructed with an expectation of any linear combination of time-in-sample effects. For example, one could construct an estimator with the same expectation as the present composite estimator.

Table 13: Variances of linear estimators with time-in-sample effects; variance of direct estimator of current level equals one.

Period of Change	Unemployed		CLF	
	BLUE 24	Recursive 36	BLUE 24	Recursive 36
0	0.928	0.923	0.763	0.733
1	1.089	1.075	0.490	0.480
2	1.348	1.342	0.682	0.663
3	1.487	1.479	0.816	0.789
4	1.578	1.569	0.911	0.877
5	1.623	1.613	0.974	0.934
6	1.646	1.635	1.015	0.970
7	1.656	1.644	1.039	0.987
8	1.655	1.642	1.047	0.990
9	1.635	1.622	1.035	0.972
10	1.617	1.603	1.023	0.954
11	1.602	1.587	1.014	0.939
12	1.589	1.573	1.009	0.926

Results for two estimators are given in Table 13. The estimator denoted by "BLUE 24" is the linear estimator that is based on 24 months of data and is best for current level. The estimator denoted by "Recursive 36" was constructed in the same manner as the recursive estimator discussed earlier. Time-in-sample effects were included in the model as fixed effects that are constant over time. However, the variance of the time-in-sample effects in the "covariance matrix", used to update the estimates, was held at the level for an observation period of 36 months. Thus, the estimator is similar to an exponentially smoothed estimator in that the effect of observations in the past dies out as the distance increases.

The restriction that the expectation of the linear estimator be the same as that of the direct estimator causes the variances in Table 13 to be slightly larger than the variances of the corresponding estimators of Table 10 and Table 11.

Restricting the recursive estimator to have the same expectation as the direct estimator and using the equivalent of 36 observations to estimate the time-in-sample effects increases the variance of the estimator by about 0.5% for all change estimators of unemployed, relative to the best estimator with no time-in-sample restrictions. If only 24 months of data are used in the estimation, the variance can be as much as 1.5% larger than the unrestricted estimator for unemployed.

For the Civilian Labor Force, the restricted estimator based on 36 period time-in-sample effects is 1.3% to 4.5% less efficient than the unrestricted estimator. The estimator of year-to-year change using only 24 months of data is about 14% less efficient than the unrestricted estimator for year-to-year change.

5. SUMMARY

We identified three sources of variation in the estimates of labor force characteristics. The permanent effects of rotation groups and of primary sampling units are more important for Civilian Labor Force and employed than for unemployed. The transient effect of rotation groups is more important for unemployed than the Civilian Labor Force. It follows that use of past information gives bigger gains in the estimation of Civilian Labor Force than in the estimation of unemployed. A full least squares estimation procedure using all past information is considerably superior to the present composite estimator for Civilian Labor Force. Only very small gains relative to the present procedure are possible for the estimates of unemployed. A least squares weights procedure is described that can be used to produce internally consistent estimators when different linear combinations of past observations are used to estimate unemployed and employed. For labor force characteristics, the inclusion of time-in-sample effects in the estimation would have a modest effect on the efficiency of estimators.

ACKNOWLEDGEMENTS

This research was partly supported by the U.S. Bureau of Labor Statistics and the U.S. Bureau of the Census through contract J-9-J-8-0082 and joint statistical agreements JSA 91-1 and JSA 91-21.

REFERENCES

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

Bailar, B.A. (1978). Rotation sampling biases and their effects on estimates from panel surveys. In N. Krishnan Namboodiri, Ed. *Survey Sampling and Measurement*, 385-407. Academic Press, New York.

Blight, B.J.N., and Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B35, 61-66.

Breau, P., and Ernst, L.R. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

Cochran, W.G. (1977). *Sampling Techniques, Third Edition*, John Wiley, New York.

Duncan, G.J., and Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.

Eckler, A.R. (1987). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-685.

Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 495-500.

Fay, R.E. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 212-217.

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.

Gurney, M., and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.

Huang, L.R., and Ernst, L.R. (1981). Comparison of an alternative estimator to the current composite estimator in the current population survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 303-308.

Jessen, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

Jones, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.

Jones, R.H. (1970). Recursive estimation of a subset of regression coefficients. *Annals of Mathematical Statistics*, 41, 688-691.

Kumar, S., and Lee, H. (1983). Evaluation of composite estimation for the Canadian Labor Force Survey. *Survey Methodology*, 9, 2, 178-201.

Odell, R.L, and Lewis, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *The Journal of the Royal Statistical Society, Series B*, 12, 241-255.

Plackett, R.L. (1950). Some theorems on least squares. *Biometrika*, 37, 149-157.

Raj, D. (1965). On sampling over two occasions with probability proportionate to size. *Annals of Mathematical Statistics*, 36, 327-330.

Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

Sallas, W.M., and Harville, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 869.

Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F., and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.

Smith, T.M.F. (1978). Principles and problems in the analysis of repeated surveys. In N. Krishan Namboodiri, Ed. *Survey Sampling and Measurement*, Academic Press, New York.

Tam, S.M. (1986). Optimal prediction in stochastic regression models with application to the analysis of repeated surveys. *Australian Journal of Statistics*, 28, 345-353.

Wolter, K. (1979). Composite estimation in finite populations. *Journal of the Americal Statistical Association*, 74, 604-613.

CLOSING REMARKS

Proceedings of Statistics Canada Symposium 92
Design and Analysis of Longitudinal Surveys
November 1992

CLOSING REMARKS

G.J. Brackstone[1]

That brings us to the end of Symposium 92. We have heard a large number of very good papers, from Graham Kalton's comprehensive overview of the issues on Monday morning to Wayne Fuller's informative presentation on longitudinal estimation this afternoon.

We have covered a wide variety of methodological issues and a range of applications. Without attempting to summarize, let me pick out a few points that struck me as important during the Symposium. In the very first session we heard that longitudinal surveys tend to be complex, but we were urged to keep them as simple as possible. That is good advice. The challenge of longitudinal surveys is complicated enough, let's not add unnecessary complexity. Related to that was an exhortation always to keep focused on the primary objectives - also good advice. There is a tendency to want to keep adding one more requirement to a longitudinal survey given what appears to be low marginal cost. But sooner or later these added requirements distort the design and add complexity that can detract from the primary objectives.

We heard that the richness of longitudinal databases present special challenges in the use of these data. Firstly, the full value of these databases cannot be realized through more and more complex cross-tabulations; other methods of analysis reflecting the time dimension of the data are required. Secondly, there is a challenge of how to make such databases available to analysts while, on the one hand, protecting confidentiality and, on the other hand, not destroying their richness.

Finally, we heard a good deal about administrative records and their importance in developing longitudinal data, whether as a direct source of data, a supplement to survey data, or as a source of data evaluation.

It is for you to judge whether the Symposium has been a success. I hope everyone has heard or learned something from this Symposium that will be of importance to take home with them for use in their continuing work. Certainly in terms of attendance, this Symposium has been successful. We had 420 registrants. Fortunately, they did not all attend all the sessions or we would have been very squeezed. Registrants came from nine different countries.

We will again be producing Proceedings of this Symposium which will be sent to all registrants from outside Statistics Canada. Within Statistics Canada we will ensure that copies are available at the divisional level. Proceedings of previous Symposia and extra copies for this Symposium can be ordered.

I would like to acknowledge and thank a number of people who have helped this Symposium to run smoothly. First and foremost, the Organizing Committee of John Armstrong, Nancy Darcovich and Pierre Lavallée deserve our thanks. They planned an interesting and balanced program and succeeded in attracting many of the most experienced practitioners and knowledgeable theoreticians in this area. Only those of you who have had the experience will truly appreciate the amount of work that goes into planning such an event. Furthermore, they are not able to fully benefit from the results of their labours because of the need to deal with difficulties arising during the Symposium itself.

I would like to thank again our co-sponsors for this event, the Laboratory for Research in Statistics and Probability of Carleton University and the University of Ottawa, and the Environmental Health Directorate of

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26-J, R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

Health and Welfare Canada. Also I must thank Dan Krewski and Avi Singh for having each organized one of the sessions of the Symposium.

A number of people have assisted with the administrative and hospitality arrangements: Hélène St-Jean helped in many aspects of the organization; Suzanne Bonnell, Carole Jean-Marie, Carmen Lacroix, Christine Larabie, and Lynn Savage helped during the Symposium itself. We would especially like to thank Carolyn Zirbser for her many hours of preparing materials and making arrangements for Symposium 92.

Let me add my thanks to our translators for their excellent service, and to all presenters, session chairpersons, and participants for helping to make this Symposium run smoothly.

Next year we will be breaking with tradition. For our 10th Symposium we are going to Buffalo, New York. We are co-sponsoring the International Conference on Establishment Surveys to be held June 27-30, 1993 in Buffalo and this will replace our regular Symposium here in Ottawa. I urge you to support this Conference which will deal with survey methods for businesses, farms and institutions. Further information is available in the brochures on the Information Table.

So that concludes our proceedings. Thank you for coming and bon voyage.