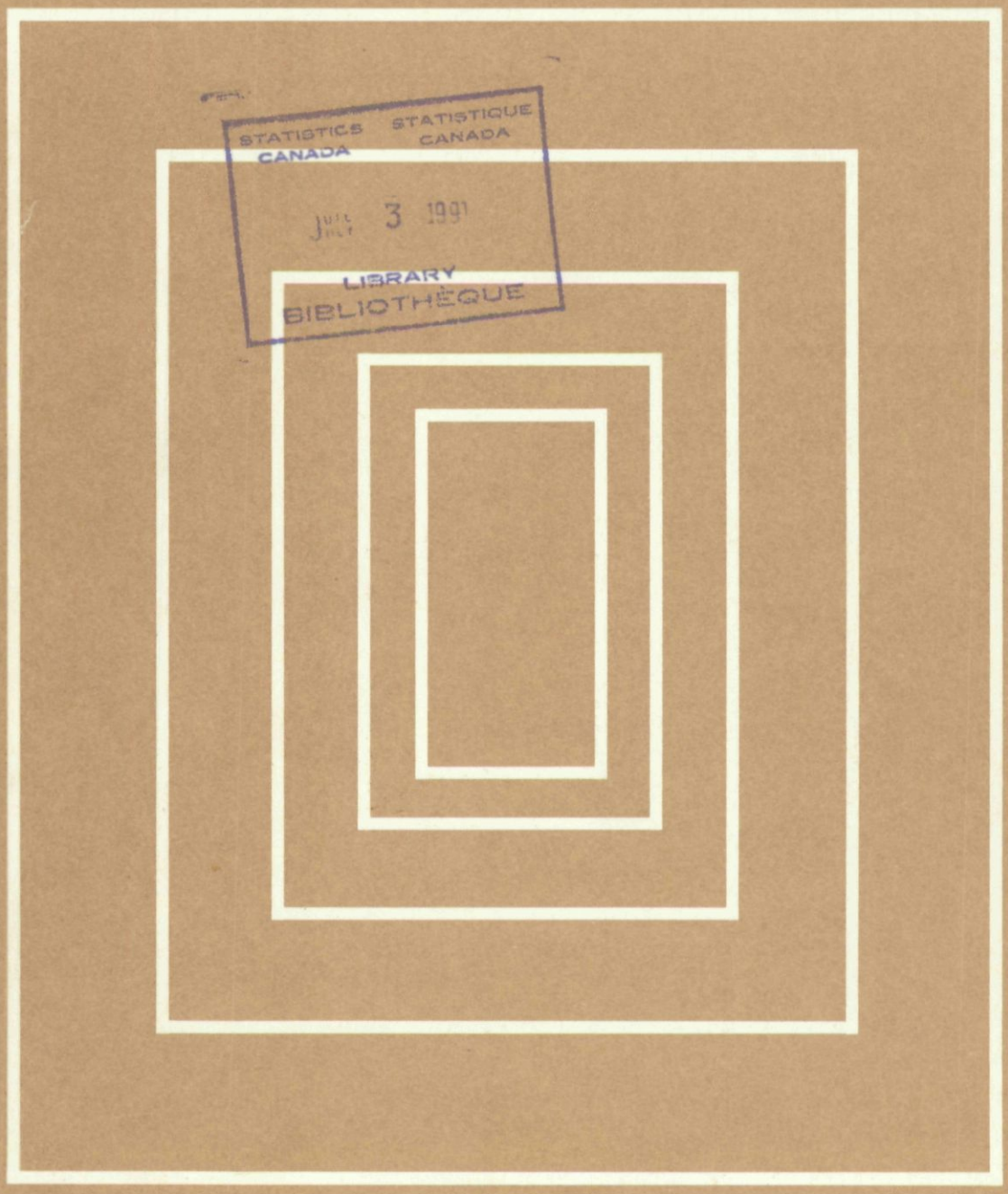
Catalogue 12-001

# Survey Methodology

A Journal of Statistics Canada

June 1991 Volume 17 Number 1





Statistics Canada Statistique Canada

Canada

Statistics Canada
Social Survey Methods Division

# Survey Methodology

A Journal of Statistics Canada

June 1991 Volume 17 Number 1

Published under the authority of the Minister
of Industry, Science and Technology



June 1991

Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

# SURVEY METHODOLOGY

## A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.



---

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

# SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 17, Number 1, June 1991

## CONTENTS

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 1–2
Statistics Canada



# In This Issue

Conducting a survey using traditional telephone methods is relatively inexpensive and maintains direct, albeit remote, contact between interviewer and respondent. These two characteristics of telephone surveys – low cost and human contact – explain to a large extent why such surveys became popular. The extensive research that took place during the last decade is summarized in the book *Telephone Survey Methodology* edited by R. M. Groves *et al.* (1988) and in Volume 4, Number 4 of the *Journal of Official Statistics*. This issue's special section, devoted to data collection and capture, presents further developments, with the telephone playing a major role in both old and new ways.

In the first paper of the special section, Werking and Clayton look at research on telephone collection methods done during the past seven years at the U.S. Bureau of Labor Statistics (BLS). They show how the research has led to the decision to implement computer assisted telephone interviewing (CATI) and touchtone data entry (TDE) in the Current Employment Statistics Survey during 1991 and 1992. The authors also mention a new technology, voice recognition, as an alternative to TDE.

Phipps and Tupek discuss one particular area of the above research at the BLS, namely, measurement errors in TDE surveys. They conclude that entering extra or incorrect digits has the most impact, but that this can be reduced via editing, particularly when longitudinal data are available. They also suggest ways of improving TDE systems.

The Mitofsky-Waksberg random digit dialing (RDD) procedure is a commonly-used method for selecting households. Brick and Waksberg discuss an inefficient operational feature of the procedure and propose a modified method. They study its statistical properties and provide guidelines for choosing between the original and modified variants of the procedure.

Bethlehem and Keller discuss Blaise, an innovative software package developed at the Netherlands Bureau of Statistics. They show how Blaise is used as a tool to integrate various steps in survey processing, including data collection and data entry. As the use of portable computers for sample surveys increases, tools such as Blaise will become indispensable.

Like most statistical agencies throughout the world, Statistics Canada has studied increased and improved telephone usage for survey taking. Recent research and testing for household surveys is summarized by Drew in the last paper of the special section. The implications of the results for the redesign of the Labour Force Survey are also presented.

For sampling on successive occasions, a normal linear regression model can be used. Bellhouse obtains the marginal and conditional likelihoods for the correlation matrix of this model. Applications of these likelihood methods are given, both for the case of simple random sampling and for more complex designs.

Schiopu-Kratina and Srinath describe the methodology of the Survey of Employment, Payroll and Hours (SEPH) conducted by Statistics Canada. SEPH is a large monthly establishment survey which uses a rotating sample. The determination of monthly sample size is quite complex due to continuous changes in the population. Some possible simplifications of the design are mentioned.

Kott shows how some techniques from design-based sample survey theory, namely the inclusion of sampling weights and the mean square error (mse) estimator based on linearization, can be used in the estimation of a system of linear equations. He shows that the use of sampling weights may be preferable when the existence of missing regressors is likely. Furthermore, the mse estimator based on linearization is nearly unbiased for many error structures.

The need to make information available quickly, combined with the need for accurate estimates based on all available data, leads to a process of revision of the national accounts estimates. Biggeri and Trivellato review recent developments in the analysis of the reliability of national accounts estimates that are subsequently revised. An empirical analysis using data from Canada, Italy and the United States is also presented.

The Editor

As we were going to press, we learned that Dr. M.N. Murthy, Director of the Applied Statistics Centre, Madras, India passed away on 2 April, 1991. Dr. Murthy made important contributions to survey methodology and was the author of a well-known textbook, *Sampling Theory and Methods*. During his career he worked at the Indian Statistical Institute and the UNO-sponsored Statistical Institute for Asia and the Pacific. He was a Fellow of the ASA, the Royal Statistical Society and the International Statistical Institute. We were fortunate to have him serve on our editorial board. He will be missed by his colleagues and former students around the world. He is survived by his wife Vyjayanthi and daughter Shashi.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 3–14
Statistics Canada



# Enhancing Data Quality Through the Use of Mixed Mode Collection

GEORGE S. WERKING and RICHARD L. CLAYTON[1]

ABSTRACT

A chronic problem in the preparation of time critical estimates is the significant limitations inherent in data collection by mail. To address this issue, the U.S. Bureau of Labor Statistics has conducted an extensive 7 year research effort into the use of computer assisted telephone interviewing (CATI) and the computer assisted self interviewing (CASI) methods of touchtone and voice recognition self-response. This paper will summarize some of the significant results of this research covering both performance and cost data. The paper will conclude with a discussion of a large scale implementation program of these techniques for a monthly sample of 350,000 establishments.

KEY WORDS: Employment statistics; Revisions; CATI; Touchtone collection; Voice collection; Cost analysis.

## 1. INTRODUCTION

### 1.1 Employment Statistics in the United States

On the first Friday of each month, the U.S. Bureau of Labor Statistics releases data on the United States' employment situation for the previous month. On release day, the Commissioner of Labor Statistics appears before the Joint Economic Committee of Congress and provides a detailed analysis of the current month's data and trends; at the same time, the data are made available to the news media and the financial and business communities. This closely watched set of statistics is the earliest indicator available on the previous month's economic activity and is used as a major gauge of the health of the U.S. economy. The data in the release cover employment, hours and earnings by detailed industry which are derived from the Bureau's 350,000 unit monthly establishment survey – the Current Employment Statistics (CES) survey – along with labor force and unemployment data which are derived from the Bureau's 60,000 unit household survey – the Current Population Survey (CPS).

The establishment survey data have many important economic uses. Due to the CES survey's size and timeliness in conjunction with the importance of the basic payroll statistics which it collects, the CES monthly estimates are not only used as principal economic indicators by themselves but they are also included in the development of many of the Nation's other major economic indicators including: Personal Income for the Gross National Product, the Index of Leading Economic Indicators, the Index of Coincident Indicators, the Industrial Production Index, Real Earnings measures and Productivity measures. While the timeliness and accuracy of the CES statistics are essential in analyzing the current economic conditions in the United States, the CES survey has had to rely on a mail data collection process since its inception in the early 1900s. This collection process results initially in the publication of "preliminary" estimates for higher level aggregates using only the sample returns received to date followed by "final" estimates two months later which use the full sample. The process of producing both preliminary and final estimates for a given month periodically yields a

[1] G. Werking and R. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Room 2089, 441 G Street, N.W., Washington, D.C., USA, 20212.

substantial revision to the initial estimates. These revisions not only affect the basic CES statistics but also the other statistics which make use of CES estimates as input. To address this issue, the Bureau initiated a research program into automated telephone data collection approaches with the objective of substantially reducing the size and frequency of large revisions in the preliminary estimates.

This paper provides an overview of the Bureau's 7 year research program into automated telephone collection techniques and summarizes some of the most significant results. The following sections describe the CES survey process; discuss the research program evaluating Computer Assisted Telephone Interviewing (CATI), Touchtone Data Entry (TDE) and Voice Recognition (VR) data collection methods; detail some of the significant research results covering both performance and cost data; and conclude with a discussion of the large scale implementation program of these methods in the CES survey.

### 1.2 Current Employment Statistics Survey

The CES survey, with 350,000 units, is the largest monthly sample survey in the United States. It is conducted by the Bureau as a Federal-State cooperative program under which the Bureau specifies the survey's sample design and operational procedures while each State conducts all data collection and edit reconciliation activities. The Bureau produces and publishes extensive monthly industry detail at the 2, 3 and 4 digit industry levels for the Nation as a whole while each State produces monthly State and area (270 Metropolitan Statistical Areas) estimates.

The CES estimates are widely regarded as highly accurate economic statistics. Once each year, complete (or universe) employment counts for the previous year become available from the Unemployment Insurance tax records; these counts are used to annually benchmark (realign) the CES sample estimates to these universe counts. The annual benchmark process yields more accurate current monthly estimates along with providing an annual estimate of overall survey error. The average difference in the CES final sample estimate versus the complete universe count over the past 5 years is under 0.2% with 4 years in the 1980s when the difference was approximately zero. While the CES final monthly estimates are regarded as highly accurate relative to the universe counts; the preliminary monthly estimates, which are based on approximately 50% of the mail sample returns, have been periodically subject to large revisions when compared to the final estimate that is available 2 months later. Over the years, some improvement in reducing the size of the monthly revisions has been made; however, periodic large revisions have been viewed as a byproduct of conducting a large decentralized mail data collection process.

The decade of the 1980s brought about a number of changes for the CES program which would significantly alter the urgency and options for resolving the monthly revision issue. The 1980s created a far more quality conscious user constituency and while the CES products had not necessarily deteriorated, the CES users' expectations on quality and "fitness for use" had greatly increased. Much of this new way of thinking is directly attributable to the efforts of Deming, Juran and others on the subject of quality management. The 1980s also saw a much greater focus on the uses and the importance of the CES payroll statistics in assessing the current health of the U.S. economy, but with the rise in the use and the visibility of the CES statistics came a corresponding user frustration with monthly revisions. The 1980s also ushered in some dramatic new technological breakthroughs, most notably in microcomputers. This new technology offered survey agencies many new opportunities for improving data collection control and quality that included: Computer Assisted Telephone Interviewing, Touchtone Data Entry, Voice Recognition, Computer Assisted Personal Interviewing and FAX. Several of these methods would ultimately offer options to significantly improve timeliness and quality at an equivalent or reduced ongoing program cost.

The 1980s saw the Bureau shift from experimental research in the CES survey to full production testing of some of the most advanced state-of-the-art automated collection techniques then available, with major implementation of these techniques scheduled for 1991.

## 2. CES RESEARCH PROGRAM

### 2.1 Research Goals

In the early 1980s, the Bureau began an extensive 7 year research effort into the causes of late response and alternative collection methods which could significantly increase response rates for the preliminary estimates. The focus of the survey research centered around obtaining answers to three basic questions:

- Are data available at the establishment in time to respond by the publication deadline for the preliminary estimates?
- Are there data collection methods which can ensure an 80-90% response rate under these tight time constraints?
- Can the cost of these data collection methods be controlled at about the same level as the current mail collection costs?

At the conclusion of the research program, a mixed mode CATI/TDE collection approach emerged which satisfied the response rate and cost constraints for the survey. The following sections provide a brief description of these personal computer (PC) based data collection methods, the research tests, the response rate results and the cost analysis. Further details on these tests are documented in the research papers listed in the references. Additionally, recent results on measurement error for Touchtone collection are presented in a paper by Phipps and Tupek, this issue.

### 2.2 Data Collection Methods

The CES survey has a very limited data collection time period available to meet the preliminary estimates publication deadline. The CES survey's reference date is the payroll period containing the 12th of the month; thus, there are only 2 1/2 weeks available to collect, keypunch, edit, tabulate, validate and publish the data. In order to meet these tight time constraints, a collection method must be able to obtain the required data as soon as they become available within the establishment. The four data collection methods studied are described in turn below.

**Mail** - The CES questionnaire is a single page mail-shuttle form which provides space for the employer to record 12 months of data. The employer receives the questionnaire in the mail each month on or about the 12th of the month (*i.e.*, the survey reference date) and subsequently fills in the row of data items corresponding to the current month. There are five basic data items collected: all employees, women worker employment and production (or nonsupervisory) worker employment, hours and earnings. Once completed, the employer mails the form back to the State agency where it is keypunched and edited. The form is then filed so that it can be mailed back to the employer for the collection of the next month's report. As indicated earlier, this process currently yields a 50% response rate in the 2 1/2 weeks available for the preliminary estimates.

**Computer Assisted Telephone Interviewing** - Under CATI collection, the employer is mailed the CES questionnaire once at the beginning of the year and retains it for recording each month's data throughout the year. Each month as the payroll data become available, the employer fills

in the data items for that month and waits for the prearranged CATI call from the State agency. When the State agency calls, the data are collected under CATI, edited and a time for the next month's collection call is arranged.

**Touchtone Date Entry** - Under TDE reporting, the employer does the same activities as under CATI except instead of waiting for the State agency's CATI call, the employer now calls an 800 telephone number connected to the touchtone PC located at the State agency. The employer then touchtone enters the data items following the prompts in the automated CES interview. As each data item is entered by the employer, it is read back for respondent verification.

**Voice Recognition** - VR data reporting is identical to touchtone collection except the employer no longer needs to have a touchtone phone. The employer now reads the data as they appear on the form and the voice PC translates and reads back the data to the employer for verification. The VR system is speaker independent and accepts continuous speech; it recognizes the digits 0 through 9 and "yes" and "no".

### 2.3 Research Tests

The Bureau began developing a PC-based CATI system in 1983 for use in a two State test that began in 1984 (Figure 1). The CATI system developed by the University of California at Berkeley was selected for the test and was subsequently used throughout the research effort. A small random sample of 200 units was selected in each State and collection procedures and systems were refined over the next 7 years. The initial research tests were highly successful in the response rates they achieved and the tests were expanded to 9 States in 1986 and then to a total of 14 States in 1988. The composition of the test sample was also changed in 1986. Instead of selecting random samples of the full CES sample, the subsequent research tests focused only on random samples of habitually late CES respondents (*i.e.*, those units which had a response rate of under 20% for the preliminary estimates publication deadline). Thus, the success of the new collection methods of CATI and TDE was measured in terms of their ability to move samples of reporting units with a 0-20% preliminary estimates response rate to a stable ongoing 80-90% response rate. By the end of the CATI research phase in 1990, the Bureau was collecting over 5,000 units monthly under CATI and had conducted well over a quarter of a million CATI interviews.

While CATI was proving to be highly successful in improving response rates, it also became clear by 1985 that ongoing CATI collection would be more expensive than the existing mail collection. At this time, a separate path of research was begun on how to reduce the cost of CATI, while still maintaining the high monthly response rates which it was achieving. While improvements were made in reducing the length of time required for a CATI interview, it was a new alternative PC-based telephone reporting method which would offer dramatic reductions in the collection costs of CATI.

By 1985, many U.S. banks were operating a version of touchtone entry verification for check cashing at drive-in windows. The Bureau identified a PC-based touchtone reporting system suitable for survey research testing and by 1986 was conducting a small two State test of this technique for collecting data. TDE was not viewed as a direct replacement for mail nor as a competitive method to be tested against CATI. CATI's role was to take habitually late responders and turn them into timely responders through personal contact and an educational process, while TDE's role was to take these timely CATI responders and maintain their response rates at the same high level, but at a greatly reduced unit cost. Over the 5 years of data collection, TDE has also proven to be a very successful and reliable method of telephone data collection. The research phase for TDE is now also being concluded with over 5,000 units continuing to report monthly under TDE across 14 States; in total the Bureau has collected over 100,000 schedules using this new automated reporting method.

As a natural follow-up to TDE, the Bureau is currently conducting several small research tests of a new Voice Recognition reporting system. Preliminary results for VR reporting have replicated the same high monthly response rates achieved under TDE, but with the important advantage that respondents find VR reporting more natural and generally prefer it over TDE. At this time, the cost of the VR hardware is approximately 15 times that of TDE; however, within several years as the initial costs of VR drop, this collection method should become a viable replacement for TDE.

### 2.4 Research Results

Over the past 7 years, the Bureau has been able to establish that payroll data is available in most firms prior to the publication deadline for the preliminary estimates and that CATI collection has the ability to take traditionally late mail responders (*i.e.*, 0-20% response rate for preliminary estimates) and within 6 months turn them into timely responders with response rates of 82-84% (Figure 1). These response rates have been remarkably stable over the years as the CATI sample has been expanded from 400 units to 5,000 units and the number of participating States increased from 2 States to 14 States. The research results indicate that the data do exist at most establishments in time to meet the publication deadline and that CATI collection can raise mail response rates by 60-80% for these late respondents and this rate can be maintained in the targeted range of 80-90% over long periods of time. The principal limiting factor in a respondent's ability to make the publication deadline was found to be the length of the firm's pay period (Figure 3). Employer pay periods are generally weekly, biweekly, semi-monthly or monthly. Weekly and semi-monthly payrolls can almost always be collected in time for publication with biweekly pay periods available most of the time; however, most monthly payroll systems close out well after the publication deadline. Monthly payrolls have been one of the largest factors in limiting the CATI response rates to the 82-84% range.

Several other important results have come out of the CATI research. Under CATI, approximately 60% of the respondents will have their data available on the prearranged date for the first call with the remaining 40% using the first call as a prompt call (Figure 1). This rate has varied little across States or over the years of testing. A small test is scheduled to be conducted to see if an advance postcard notice to the respondent shortly before the prearranged CATI contact date will significantly limit the number of callbacks required.

| | | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 |
|---|---|---|---|---|---|---|---|---|
| Mail | Resp. Rates | 47% | 47% | 48% | 49% | 49% | 51% | 52% |
| CATI | Units | 400 | 400 | 2000 | 3000 | 5000 | 5000 | 5000 |
| | Resp. Rates | 83% | 84% | 82% | 84% | 83% | 84% | 82% |
| | % Call Back | 44% | 42% | 40% | 41% | 42% | 41% | 41% |
| | Av. Minutes | 5.6 | 5.6 | 5.0 | 4.8 | 4.4 | 3.5 | 3.8 |
| TDE & Voice | Units | | | | 400 | 600 | 2000 | 5000 |
| | Resp. Rates | | | | 78% | 80% | 84% | 82% |
| | % Call Back | | | | 45% | 45% | 43% | 40% |
| | Av. Minutes | | | | 1.8 | 1.8 | 1.7 | 1.7 |

**Figure 1.** Research Summary

The average time for a CATI interview depends on the number of items to be collected, the time efficiency of the interview instrument, and the experience of the data collector. The average time for a CATI call (Figure 1) was reduced by one-third as the CATI instrument was streamlined and interviewers became more experienced. Another very important concern in the testing was the effect of CATI on sample attrition. There was some concern that employers would not want to be constantly bothered by telephone contacts and would drop out of the program. However, the sample attrition rate for CATI was about one-third of that for mail with almost no loss of large reporters under CATI. In summary, CATI appears to have come close to maximizing the achievable response rate for the preliminary estimates while also enjoying broad support from the respondents.

Due to the increased cost associated with CATI collection, the Bureau initiated research into touchtone collection. During the 4 years of testing, TDE has demonstrated the ability to take timely CATI reporters having 82-84% response rates and maintain these high rates under completely automated TDE reporting (Figure 1). The importance of this result lies in the cost savings under TDE collection versus CATI collection. One of the major concerns for TDE collection was that, unlike CATI where respondent contacts are scheduled throughout the day, TDE respondents might tend to call during the same time period thus generating busy signals and require an excessive number of touchtone PCs to handle peak load reporting. Fortunately, this was not the case, and while the touchtone PCs are on-line 24 hours a day, most calls are relatively uniformly distributed between 8am and 5pm (Figure 4). TDE respondents tend to require the same proportion of prompt calls as CATI respondents - approximately 40%. Methods are currently also being tested to reduce the TDE prompting workload. One major advantage for the respondent is that TDE collection requires only one-half of the time of a CATI interview with the average TDE interview lasting only 1 minute and 45 seconds. Additionally, touchtone phones are widely available at most establishments; current estimates indicate that over 80% of employers could report under touchtone data collection. While TDE reporting offers many advantages to the survey agency, its strongest feature is respondent acceptance; respondent reaction to touchtone reporting has been very positive due to its speed and convenience for the respondent.

One general observation concerning the development of a CATI research program is that it is not critical which CATI hardware or software system is used during the research phase as long as it is reasonably flexible for change. The final results from testing may suggest very different CATI requirements for production implementation than those originally required for the research program. The most important and time consuming activity is the development and refinement of the methods and procedures for respondent contact. Once effective methods and procedures are developed, the requirements for the "right" system become more obvious.

### 2.5 Cost Analysis

With the performance testing and respondent acceptance for CATI and TDE proving to be highly successful, the final phase of research shifted to analyzing the transitional costs of CATI and the ongoing costs of TDE collection.

The major "labor" and "non-labor" cost categories were studied for mail, CATI, and TDE collection (Figure 2). The study looked not only at estimates of current cost, but also at projected costs over the next 10 years using the current rate of increase for the major cost items. Since CATI was to play only a 6 month transitional role (*i.e.*, moving late responding mail units to on-time responding CATI units) prior to conversion to ongoing TDE, the major focus of the cost analysis was on the cost tradeoffs between ongoing mail versus ongoing TDE data collection.

| Cost Category | Mail | CATI | Self-response (TDE & VR) |
|---|---|---|---|
| **LABOR** mail out | ↗ | | |
| mail return | ↗ | | |
| data entry | ↗ | ↗ | |
| edit and edit reconciliation | ↗ | ↗ | ↗ |
| nonresponse followup | | | ↗ |
| **NON-LABOR** postage | ↗ | | ↗ |
| telephones | | ↘ | ↘ |
| microcomputers | | ↘ | ↘ |

**Recent Annual Price Change Factors**

| | | |
|---|---|---|
| Labor | +5.7% | ECI, State and Local Government |
| Postage | +5.0% | U.S. Postal Service |
| Telephone | −1.7% | CPI-U, Intrastate toll calls |
| Microcomputers | −19.5% | PPI Experimental Price Indexes (16 bit computer) |

**Figure 2.** Data Collection Costs (arrows show direction of recent price change)

For the labor categories, the monthly mailout, mailback, check-in and forms control operations of mail were replaced by a single annual mailout-only operation under TDE, thus eliminating a large monthly clerical operation in the States. The batch keypunching, keypunch validation, and forms control operations under mail were completely eliminated under TDE, where the respondent touchtone enters the individual firm's data and validates each entry. Another major quality and cost-efficiency advantage of TDE was that procedures for telephone nonresponse follow-up became far more feasible under TDE than mail. Under TDE, an accurate up-to-date list can be generated of respondents who have not yet called in their data, this list can then be used to conduct brief telephone prompting calls. Under mail, telephone follow-up activities of "apparent" nonrespondents were awkward since the State staff did not know whether the respondent's form was not yet completed, currently in the mail, in the State check-in process or at keypunch; in addition, respondents who had recently sent their form tended to resent the additional reminder for an activity that they perceived as completed. Due to the voluntary nature of the program and the uncertainty of a respondent's response status, telephone prompting under mail was only used for critical (large employers) units.

There were no significant cost savings made for edit reconciliation as the number of edit failures under TDE remained at about the same level as under mail. This was also true for postcard reminders where the number of postcards used under mail collection for late respondents was approximately the same as the number used under TDE, where respondents received an "advance" postcard notice to touchtone their data by the due date.

In the non-labor categories, the cost of postage under mail (currently 58 cents per unit) is replaced by the cost of a telephone call and the amortized cost of the TDE machine (together currently 46 cents per unit). Postage is a continually increasing cost with an annual price increase of approximately 5% (Figure 2). The rising cost of postage is driven by annual labor cost increases (+5.7%) and by fuel costs (also generally increasing) with labor accounting for over 80% of total postage costs. In contrast, under TDE the cost of telephone calls has been decreasing in recent years (−1.7%), along with the cost of microcomputers (−19.5%).

Excluding the additional new requirement of full telephone nonresponse prompting for TDE, there are demonstrable cost savings in shifting from mail to TDE. Perhaps more importantly, under a 10 year projection of future costs of these two collection methods (Figure 5), these savings grow substantially. Attempts will be made to redirect future cost savings from TDE to help offset the full nonresponse prompting activities.

There are several major conclusions concerning TDE reporting which have emerged from the performance and cost analysis review. (1) The traditional view that mail is the least expensive collection option available to statistical agencies is no longer true. The major technological breakthroughs of the 1980s in automated telephone collection can not only reduce the collection cost below that of mail but can also improve timeliness and control over the collection process. Additionally, over the next 5-10 years, the cost of mail will become even less cost competitive with these high-technology/low-labor collection approaches due to the increasing labor and postage costs associated with mail. (2) The transition of respondents from mail to TDE appears to cause very little disruption to monthly reporting. In the Bureau's follow-up interviews with respondents who were converted to TDE, results have shown that respondents have very little trouble adapting to this new method of reporting. Virtually all respondents completed their first month TDE report accurately and without assistance, with many respondents commenting on the ease of reporting under TDE. (3) TDE can be viewed as a reliable replacement method for survey data collection; over the past 4 years of collection there have been no major equipment failure problems or disruptions of the collection process. Minor equipment problems have been easily resolved using a back-up PC when required. In addition, future back-up protection for the State TDE collection process will involve the use of a call forwarding option to reroute calls to a central site should major problems occur at the State.

## 3. IMPLEMENTATION

### 3.1 Major Issues

By the end of 1989, the Bureau had completed a very successful research program and had sustained the high performance levels over 7 years. However, there is a significant difference between the completion of successful research and full scale implementation of new methods. While over 10,000 units were being actively collected under these new techniques, these units represented under 3% of the CES sample. Proposed collection changes for a monthly sample of 350,000 units which has been collected for well over half a century under a decentralized State mail collection environment requires not only a very strong demonstrable user need but also broad-based support at national, regional and State levels.

As it turned out, the user need had begun to change in the early 1980s. During the 1980s, the U.S. economy experienced the longest sustained peace-time growth period in its history, with over 19 million jobs created, and unemployment rates at their lowest levels since the early 1970s. By the mid 1980s, economic policy was firmly focused on establishing non-inflationary economic growth. The monthly CES employment growth and wage data were being closely

monitored for signs of wage-induced inflationary pressures resulting from strong job growth during a period of low unemployment. With this greatly increased use and visibility of the monthly data, came a corresponding user frustration with the periodic large revisions to the preliminary estimates.

While monthly revisions to the preliminary estimates had always been a part of the CES survey process and even though the size of these revisions had been reduced over the years, large revisions of over 100,000 in the preliminary monthly employment estimates of over-the-month change were now being viewed as unacceptable. These user demands for greater accuracy in the preliminary estimates would lead the Bureau to develop proposals for the implementation of automated CATI and TDE collection methods into the U.S. government's largest monthly survey. While user demand is critical, major changes of this magnitude could not be undertaken without full support at the State level where data collection actually occurs. One guideline which remained constant throughout the research program was that the collection system was ultimately the States' data collection system and therefore must be designed to integrate well into their survey environment and create as minimal an organizational impact as possible. To that end, the CATI and TDE systems remained open for change throughout the research program. As many State suggestions and requirements as possible were taken into account with each new release of the systems. The success of much of the development work can be credited to the resiliency and endurance of the 14 research States as they made constant recommendations for improvements in systems and procedures. In the end, the CATI interview instrument had moved from an awkward simulated household survey type interview approach to a fast and efficient "screens" and "windows" approach well suited for capturing and editing longitudinal economic data. Thus, at the conclusion of the research phase, the systems and procedures were well tested and refined across a wide range of States. This approach to testing brought with it a strong sense of confidence in the methods and the systems at the State level. This would prove to be essential for the Bureau's proposed quick production implementation timetable of these state-of-the-art collection methods.

### 3.2 Approach and Impact

The main focus of the implementation proposal was the control of revisions in the preliminary estimates. Over the past 5 years, approximately 40% of all revisions were over 50,000 with 13% of the revisions exceeding 100,000 (*i.e.*, large revisions) (Figure 6). The goal of the implementation study was to identify a minimum set of late responders which, if obtained by the publication deadline, would control the size of revisions to what was considered to be in an acceptable level (under 50,000 revision in the over-the-month employment change). While one obvious approach was to convert all 175,000 late respondents to the new collection methods, this approach was considered to be lengthy and costly. While there was a need to responsibly control the size of revisions (*i.e.*, not necessarily completely eliminate all revisions), there was also a corresponding need to resolve this problem in as timely a fashion as possible (*i.e.*, convert the smallest number of units necessary to control revisions to under the 50,000 level).

Establishment surveys, unlike household surveys, generally have differential weighting for individual units with very large units being "certainty" units in the sample design. In the CES sample design, units with 100+ employees make up only 20% of the sample (*i.e.*, 75,000 units), but account for over 83% of the unweighted sample employment. These units tend to have a much lower response rate for the preliminary estimates so that if the late respondents' employment trend differs from the early respondents, these units can create a substantial revision in the sample estimates. Revision impact studies were conducted to assess the affect of large employers 100+ on the preliminary estimates. To test the impact of large employers,

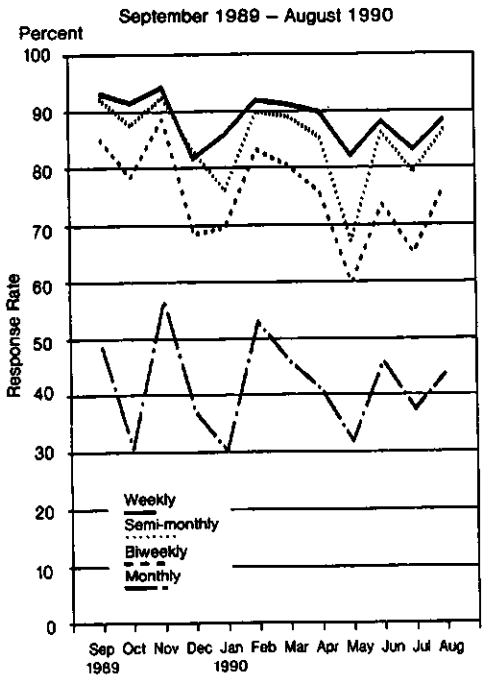



**Figure 3.** CES CATI First Closing Performance by Length of Pay Period

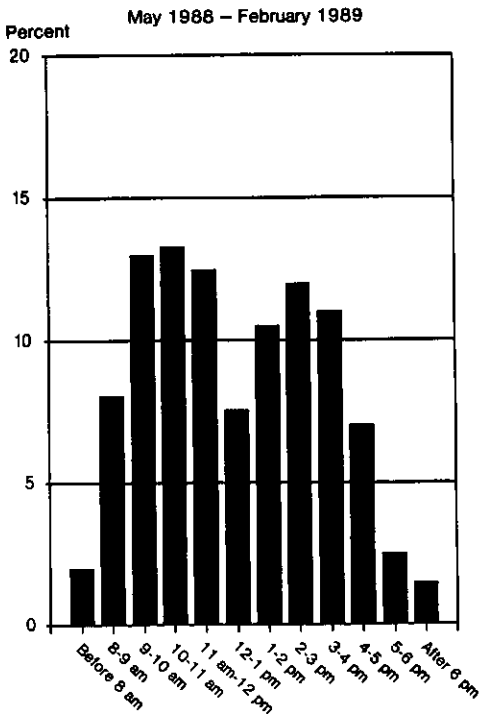



**Figure 4.** Touchtone Data Entry Distribution of TDE Calls by Time of Day

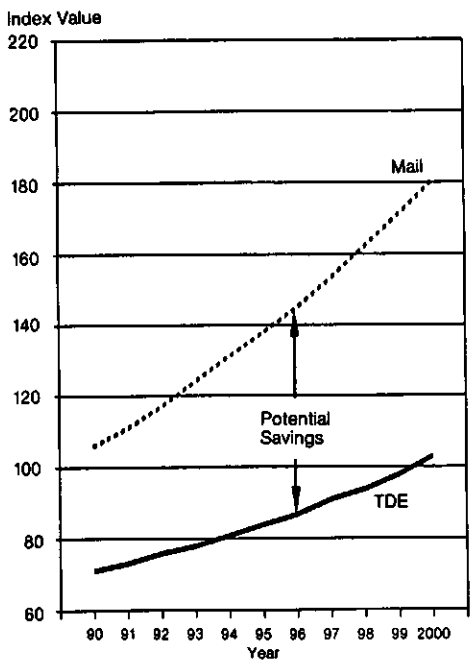



**Figure 5.** Estimated Unit Costs by Mode: 1990-2000

Note: Not official BLS projections

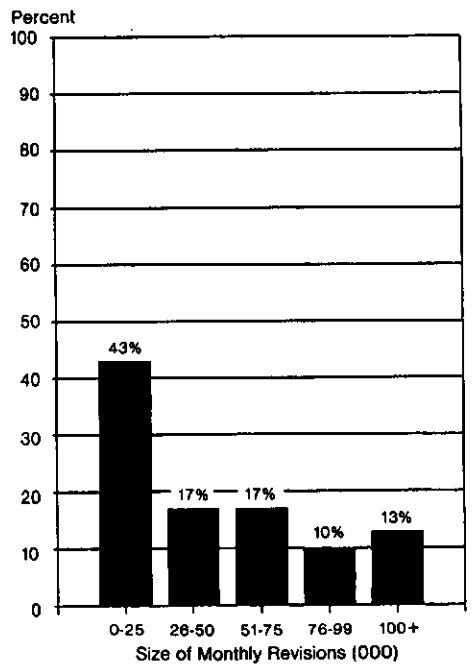



**Figure 6.** Distribution of Magnitude of Revisions (1985-1989)

late respondents in size class 100+ were included in the original sample used for the preliminary estimates and the estimates were recalculated. These new estimates were then compared to the original preliminary estimates to determine the impact of 100+ employers on revisions. The results indicated that from one-half to two-thirds of the revision was attributable to these units. These studies were repeated over several months with similar results. Applying these projected reduction rates in revision size to revisions over the past 5 years resulted in over 97% of all revisions being below the 50,000 level as compared to the current level of only 60%. This greatly reduced targeted sample size for conversion to CATI/TDE provided for an accelerated implementation time schedule consistent with controlling conversion costs to the minimum level necessary to protect against large revisions in the preliminary estimates. The Bureau will begin the implementation of CATI/TDE collection methods in 25 States in 1991 with planned expansion to all States in 1992. With the implementation of the new collection methods, the Bureau will be able to help resolve one of the most difficult and visible quality issues affecting the CES user community.

## 4. SUMMARY

The decade of the 1980s has brought about many changes for survey agencies. Some of the changes can be viewed in terms of our accomplishments made over the decade while others are more subtle and need to be viewed in terms of the changes in the survey environment in which we operate.

The 1980s created a far more quality conscious user constituency which is quick to identify and point out our product limitations. While our products may not have deteriorated, our users' expectations on quality and "fitness for use" have greatly increased. This is an issue which we as statistical agencies must be able to respond to in order to maintain our credibility with the user community. The 1980s also ushered in dramatic new technological breakthroughs most notably in microcomputers. The new technology has offered survey agencies many new opportunities for improving data collection control and quality including: CATI, CAPI, TDE, VR and FAX. Some of these options offer improved quality and control at lower ongoing costs. The decade of the 1990s may well offer even greater opportunities for using technology to improve our data collection timeliness and quality at lower costs.

As we look at the status of our statistical programs, we often find very rigid environments. The data collection approach for our surveys often date back to their inception. Our data collection cost assumptions and cost studies are usually well outdated and often simplistic in approach. Since data collection generally represents the largest part of a survey's cost, it is usually well entrenched in the agency's organizational structure and can be quite difficult to restructure in order to accommodate large scale change. It is within this survey environment that we will face the major challenges and opportunities of the 1990s.

The challenge for statistical agencies in the 1990s will be threefold:

- to be responsive to the changing quality needs of our users;
- to attempt to have our research stay up with the rapid change of technology and automated data collection approaches; and perhaps more importantly
- to continue to find ways to incorporate successful research into our ongoing programs.

These challenges will determine the cost and quality competitiveness of our programs and our agencies in the future.

## REFERENCES

CLAYTON, R.L., and HARRELL, L., Jr. (1989). Developing a cost model of alternative data collection methods: MAIL, CATI and TDE. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

CLAYTON, R.L., and WINTER, D.L.S. (1990). Speech data entry: Results of the first test of voice recognition for data collection. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

JURAN, J.M., GRYNA, F.N., Jr., and BINGHAM, R.S., Jr., eds. (1979). *Quality Control Handbook*, Third Edition. McGraw Hill.

GROVES, R.M.J., *et al.*, eds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.

OFFICE OF MANAGEMENT AND BUDGET (1988). *Quality in Establishment Surveys*. Statistical Policy Working Paper 15.

OFFICE OF MANAGEMENT AND BUDGET (1990). *Computer Assisted Survey Information Collection*. Statistical Policy Working Paper 19.

PONIKOWSKI, C., and MEILY, S. (1988). Use of touchtone recognition technology in establishment survey data collection. Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida.

WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., and ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.

WERKING, G.S., and TUPEK, A.R. (1987). Modernizing the Current Employment Statistics Program. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 122-130.

WERKING, G., TUPEK, A., and CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 349-362.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 15–26
Statistics Canada



# Assessing Measurement Errors in a Touchtone Recognition Survey

POLLY A. PHIPPS and ALAN R. TUPEK[1]

ABSTRACT

Electronic data collection utilizing touchtone recognition is in place for a monthly establishment survey at the Bureau of Labor Statistics. The Touchtone Data Entry (TDE) system features digitized phrases requesting respondents to answer questions using the numeric keypad of a touchtone telephone. TDE has substantial implications for lowering survey costs; many labor intensive activities are eliminated. However, little is known about measurement errors associated with this mode of data collection. This study assesses TDE mode error using three sources of data, which allow for analyses of errors associated with selected aspects of the human-machine interface. In addition, instrument design issues associated with mode error are addressed. We conclude by extending the implications of our findings to other surveys.

KEY WORDS: Mode of data collection; Human-machine interface; Computer-assisted self interviewing.

## 1. INTRODUCTION

The U.S. Bureau of Labor Statistics (BLS) issues monthly employment estimates for the United States from a survey of 350,000 business establishments. This survey, the Current Employment Statistics (CES) survey, provides one of the earliest monthly measures of U.S. economic health. However, the preliminary estimates from the survey are released with data from only about one-half of the business establishments in the survey. Revised estimates are produced two months after the initial press release. The low response rate for the initial press release can result in large revisions to the estimates. The BLS began investigating the use of automated collection techniques in 1983 to increase the timeliness of response and reduce the potential for large revisions.

The CES survey has traditionally been collected by mail through state employment security agencies. Research tests conducted between 1984 and 1986, involving the replacement of mail collection with computer-assisted telephone interviewing (CATI), have shown CATI to be an effective means for improving the timeliness of response (Werking, Tupek, Ponikowski and Rosen 1986). Average response rates under CATI collection have been between 85 and 90 percent for preliminary estimates, compared to 45 to 50 percent with mail collection. While CATI collection has been effective in improving the timeliness of response, the cost of full CATI collection in the CES survey cannot be absorbed within the survey's current budget. Research has been conducted since 1986 on touchtone data entry (TDE) to develop an alternative collection method with the performance gains of CATI, but with a lower unit cost (Ponikowski, Copeland and Meily 1989). Further discussion on the use of CATI and TDE collection for the CES survey is provided in the paper by Werking and Clayton, this issue.

Recent tests of the TDE system provide data on the timeliness of response, the cost of collection, and edit failure rates. BLS tests show that TDE collection with CATI back-up is as timely and effective as CATI collection (Werking, Tupek and Clayton 1988). In addition,

[1] Polly A. Phipps, Office of Employment and Unemployment Statistics, Bureau of Labour Statistics, Room 2821, 441 G Street N.W., Washington, D.C. 20212; Alan R. Tupek, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Room 2919, 441 G Street N.W., Washington, D.C. 20212.

TDE lessens survey costs considerably. Current cost estimates indicate that the monthly unit costs for TDE are approximately 30 percent less expensive than mail collection, while the monthly unit costs for CATI collection are 20 percent more expensive than mail (Clayton and Harrell 1990). Collection by TDE is now expanding in the CES, beginning with the establishments with the greatest employment. Research continues in determining ways to improve the touchtone system, even though respondents' acceptance of touchtone collection is highly favorable.

The purpose of the current research is to identify respondents' problems using the touchtone system and to measure errors associated with the mode. Results indicate where improvements are needed to reduce respondent problems as well as errors. Section 2 provides background information on TDE and the operations of the BLS TDE system. In the third section we address the potential errors due to the TDE mode of data collection, which requires a human-machine interface. In section 4 we describe the three sources of data collected for the study. A record-check survey, machine-recorded data and a respondent debriefing survey are used in the analysis of problems and errors. Section 5 includes the methods, analyses and findings from each of the data sources. Finally, in section 6 we provide an overall assessment of measurement errors due to this mode of collection, suggestions for improving the system, and implications for other surveys.

## 2. TOUCHTONE DATA ENTRY

The primary reason to consider collection by touchtone self-response is to reduce the cost of collecting data by CATI, while maintaining the timeliness and quality of CATI collection. The CES survey seemed to be a good candidate for touchtone collection, since only five or six numeric data items are collected each month. The data items include: all employees, women workers, production workers, production-worker payroll, production-worker hours, and for some industries, over-time hours or employee commissions. Establishments are asked to report totals for each data item for the pay period which includes the 12th of the month. TDE respondents are also asked to report their establishment identification number and the month for which they are providing data.

The features of the BLS touchtone system include the ability to:

- detect legitimate establishment respondents based on a match to a file of establishment numbers;
- vary the set of questions depending on the industry of the establishment;
- read back all responses for respondent confirmation using a computer simulated (digitized) voice (respondents are requested to enter "1" to confirm their answer or "0" to reenter their answer);
- wait two seconds for respondents to begin entering their answer, and wait two seconds between digits before interpreting data entry as complete (data entry is also assumed complete if the entire field length is filled - example: the data item "month" has a two-digit field);
- repeat each data item question up to three times (for identification number, month and all employees) or request the respondent to confirm that they have no answer for the question (for all other data items), if a respondent does not confirm their answer or if no answer is provided in the two seconds after the question is read;
- store the date, start and end time of each call, and all data items (Werking *et al.* 1988).

Respondents are mailed instructions on touchtone data entry. The instructions give direction on how to use the system and examples of the computer and respondent interaction, such as:

| **Computer:** | **Your Response:** |
|---|---|
| Enter all employees | For 25 employees, press 2 and 5 |
| You entered 2, 5 | Press 1 to confirm, 0 to reenter. |

The instructions also include the optional use of the "#" sign to indicate completion of data entry for an item. The use of the "#" sign reduces the time of the interview. In addition, before reporting their data respondents can call in to try out the system using a special test identification number.

Touchtone respondents are contacted by a telephone interviewer during their first month on the system to determine any problems they may have and to provide guidance, if necessary. Respondents receive a postcard reminder each month and a prompt call if they do not self-report by a specified date. The prompt call asks the respondent to telephone into the touchtone system as soon as possible. Data are not usually collected during the prompt call.

## 3. MEASUREMENT ERROR IN A HUMAN-MACHINE INTERFACE

Respondent use of a system such as TDE to answer survey questions has little precedent. However, touchtone recognition is widely used in such procedures as electronic banking and customer-controlled telephone services. While these services may save time and expense, they have a potential to alienate users. Problems and errors can originate with the system, task, or respondent. System problems primarily generate nonresponse error, while measurement error is related to the task and respondent performance.

While not directly related to surveys, the human-factors literature suggests several interrelated factors that may contribute to performance errors in a human-machine interface. First, respondents may not be familiar or comfortable with the technology. Waterworth (1984) suggests that the language of the human-machine interface is different than human communication as actions are performed in an order reflecting computer program logic. Since the ability to think in a way that parallels the logic is not a minor exercise, those with limited experience may have difficulty understanding the task and using the system. Second, synthetic speech is more difficult to understand than natural speech and places greater processing demands on working memory (Schwab, Nusbaum and Pisoni 1985). Thus, comprehension and memory problems associated with the mode may cause errors.

Synthetic speech includes both digitized speech, where a human voice is sampled, digitally encoded, and stored, and rule-based synthesized speech, generated using text as input (Marics and Williges 1988). The TDE system utilizes digitized speech, which is less difficult to understand than rule-based synthetic speech. However, comprehension problems occur with digitizing, as it introduces distortion into original speech (Cox and Coope 1981). Research shows that the understanding of synthetic speech may improve with training. In an experiment on perception of synthetic speech, Schwab and colleagues (1985) found that training with synthetic speech increases perception performance. Thus, comprehension may improve with exposure to and experience with the system. Another factor that may affect comprehension is the pace of the system. Marics and Williges (1988) found that the rate of the synthesized speech significantly affects speech intelligibility, as measured by transcription errors and response latency. However, subjects who received contextual information prior to listening to the speech had fewer transcription errors.

Thus, potential errors in the human-machine interface can occur from lack of experience with the technology and task, and from comprehension and memory problems associated with voice clarity and pace. Yet these problems are surmountable, as the evidence indicates that experience and training can improve performance.

## 4. DATA

There were several objectives we considered in measuring TDE problems and mode error, and determining what data to use or collect. First, it was necessary to identify if and where problems were occurring. Second, we felt respondents should identify and interpret problems, but we also wanted measures independent of respondent assessment. Third, we needed to address problems and errors associated with the task and comprehension, including the possible improvement of respondent performance over time.

We decided to assess TDE problems and mode error using three different data sources, which have in common approximately 465 Pennsylvania business establishments. These establishments reported their monthly survey data by TDE to the Automated Collection Techniques (ACT) Laboratory at the BLS national office in Washington, D.C. A small number of the establishments began reporting by TDE to the ACT Lab in April, 1989. Others were added monthly through November of 1989. Most of the establishments continued reporting to the ACT Lab through April of 1990. The majority of these establishments moved from mail to TDE reporting.

The first source of data has two components. One is the TDE data recorded by machine from April to December, 1989. The other component is the same data recorded by establishments on a survey form. All respondents receive a yearly survey form on which they are requested to record their data for each month. Mail respondents fill in the form each month and mail it to the state employment security agency. The agency records the data, then returns the form by mail for next month's collection. CATI and TDE respondents are sent the survey form, but they do not return it. However, we sent a request to the TDE respondents to return their 1989 survey form, and obtained a 96 percent return rate. We then compared the TDE and form data, identifying discrepancies between the two. The TDE and survey-form data includes 1,930 observations across a nine-month period. Since establishments were phased into TDE slowly, the number of observations per establishment varies. The data cover approximately 75 establishments for 6-9 months, 200 establishments for 4-6 months and 190 establishments for 2-3 months. We refer to these data as the record-check data.

The second source of data includes machine-recorded information on respondent performance during the TDE telephone call. The TDE instrument was reprogrammed in January 1990 to automatically count and record the number of times a question was repeated due to nonresponse (question repeat), the number of times a respondent reentered data (data reentry) for each question, and the number of times an establishment called and hung up before entering data. Unfortunately, only the questions asking for the month and all employees total could be explicitly separated into question repeat and data reentry. For the data items including women and production workers, payroll and hours, we had to combine repeats and reentrys, due to the structure of the original computer program. We refer to these data as the machine-recorded data.

The third source of data is a telephone debriefing survey, conducted from January to April of 1990 with the Pennsylvania establishments on their experiences with the TDE system. Approximately 411 business establishment respondents completed the interview, an 88 percent response rate. The questions covered such topics as voice quality, pace of interview, task problems, use of systems features, adequacy of instructional materials, and a system rating.

## 5. RESULTS

### 5.1 Record-Check Data

When we requested TDE establishments to return their survey forms, our first question was: how many respondents really used the forms? We speculated that one source of mode error was respondents who did not complete the form for use when entering their TDE data, which would increase demands on their memory. Those who did not complete the form might be more likely to enter and/or verify incorrect data. Thus, the request for the survey forms indicated that respondents were to return the form regardless of whether they completed it or not. However, of the 96 percent that returned their survey forms, only one establishment mailed in a blank form; all others sent in completed forms. While nonrespondents may work from memory, most of the respondents had completed their forms, giving us reason to believe memory problems due to lack of form use were not a major source of errors.

When comparing the data received by TDE with that on survey forms, we identified and coded discrepancies. The data on the survey form are those we would have received and used if the respondents were reporting by mail. The results are shown in Table 1. The first type of discrepancy occurred when the TDE data indicated there was no response for a data item, but there was a response on the establishment's survey form. This item nonresponse accounted for the greatest number of discrepancies, 82 out of 177, and was quite evenly spread across the applicable data items.

There was a pattern to the item nonresponse by month and establishment. The item nonresponse rate was 40 percent higher in the first month an establishment reported by TDE. In addition, some establishments had more difficulty than others, indicated by two or more nonresponses. Nearly half of all item nonresponse occurred in 18 establishments at or close to the time they began responding by TDE. This indicated problems existed with first-time use of TDE that might decrease with experience. Since the problem was concentrated in a small group, we believe it reflected lack of familiarity with automated processes. The remaining item nonresponse had no identifiable patterns; our suspicion was that some establishments simply missed the item, possibly due to office distractions, and continued on with the next question.

**Table 1**

Record-Check Data - Number and Type of TDE Discrepancies

| | |
|---|---|
| TDE item nonresponse | 82 |
| 1-2 few/too many digits | 18 |
| Slipped on keypad | 17 |
| Dis/confirm - "1", "0" error | 14 |
| Form corrected, not TDE | 12 |
| No apparent error reason | 26 |
| Other reasons | 8 |
| Total | 177 |

The second type of discrepancy was entering extra digits or, in a few cases, entering too few digits, which accounted for 18 of the 177 errors. This was specifically a problem associated with entering the payroll data item, where four respondents tried to enter cents instead of rounding to the nearest dollar. Several of the same respondents appeared to enter a half hour, 50, for production-worker hours rather than rounding. In the third type of discrepancy, the TDE numbers were nearly the same as those on the form, but one number off. The number entered incorrectly indicated a potential task problem in that the respondent may have had their fingers slide over on the keypad to the number directly on the side or below the correct digit. This accounted for 17 discrepancies. The fourth type of discrepancy occurred primarily for the all employee data item. There were eight establishments who had a "1" entered for this item in the TDE data, but had a larger employment number on their survey form. We speculate that respondents entered "1" twice when confirming the previous question on month.

Finally, there were a few respondents who had corrected data on their survey form, but not on TDE. There were other discrepancies which we could not explain. In addition, several respondents transposed their numbers or were off one category, accounting for the "other" reasons. For most of the errors, it was difficult to specifically ascertain if they were caused at the time of data entry, or not clearly comprehending the question or numbers being read back for verification. We suspected the former, but only for the second discrepancy, adding too many digits, could we really rule out comprehension problems.

The error rates for the survey items, ranging from 1.2 to 2.5 percent, are shown in Table 2. The all-employee, women-worker and production-worker questions have a lower percentage of errors than payroll and hours. This is not surprising since payroll and hours worked are usually four to six digits, compared to two to three digits for the other items. Thus, longer strings of numbers cause more difficulty. This may be related to difficulties entering the data, lack of respondent motivation in correction, or problems remembering longer strings of numbers during validation.

Table 3 shows the potential effect of the discrepancies on the CES data items, calculated by taking the sum of the difference between the values in the TDE system and the form, then dividing by the sum of the values on the form. The CES Survey uses a link-relative estimator for published estimates. The estimates in Table 3 do not take into consideration this estimator. However, the estimates in Table 3 provide an indirect measure of TDE mode error on survey estimates. None of the error is significantly different from zero at the five percent level. However, the potential for mode error appears to be more serious for production workers, payroll and production-workers hours. In this study, the number of production workers are overestimated by 7.3 percent, payroll by 7.3 percent, and hours by 4.4 percent.

Nearly all the discrepancies would have failed the edit parameters used in the CES survey and been corrected. The resultant effect of the discrepancies after edit corrections is zero for

**Table 2**

Record-Check Data - Number of Discrepancies and Percent Error by Data Item

| | All Employees | Women Workers | Production Workers | Payroll | Production Hours | Total |
|---|---|---|---|---|---|---|
| Discrepancies | 23 | 29 | 28 | 48 | 49 | 177 |
| % Error | 1.2 | 1.5 | 1.5 | 2.5 | 2.5 | 1.8 |
| (SE) | (.2) | (.3) | (.3) | (.4) | (.4) | (.3) |

(N = 1,930 for each item).

**Table 3**
Record-Check Data - TDE Mode Error for Data Items Before and After Edit Corrections

| | All Employees | Women Workers | Production Workers | Payroll | Production Hours |
|---|---|---|---|---|---|
| % Mode error, before edit corrections | 0.0 | 0.5 | 7.3 | 7.3 | 4.4 |
| (SE) | (.4) | (.3) | (5.2) | (3.8) | (3.7) |
| % Mode error, after edit corrections | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| (SE) | (0.0) | (0.0) | (0.0) | (0.0) | (0.0) |

(N = 1,886-1,930 for each item).

all data items, as shown in Table 3. Examples of the large discrepancies include a respondent who incorrectly entered payroll for the number of production-workers, increasing production workers by over ten thousand, and two respondents who incorrectly entered the number of production-worker hours for number of production workers, raising the latter by several thousand. Payroll and hours have similar gross discrepancies, including the respondents who put in cents, instead of rounding for payroll.

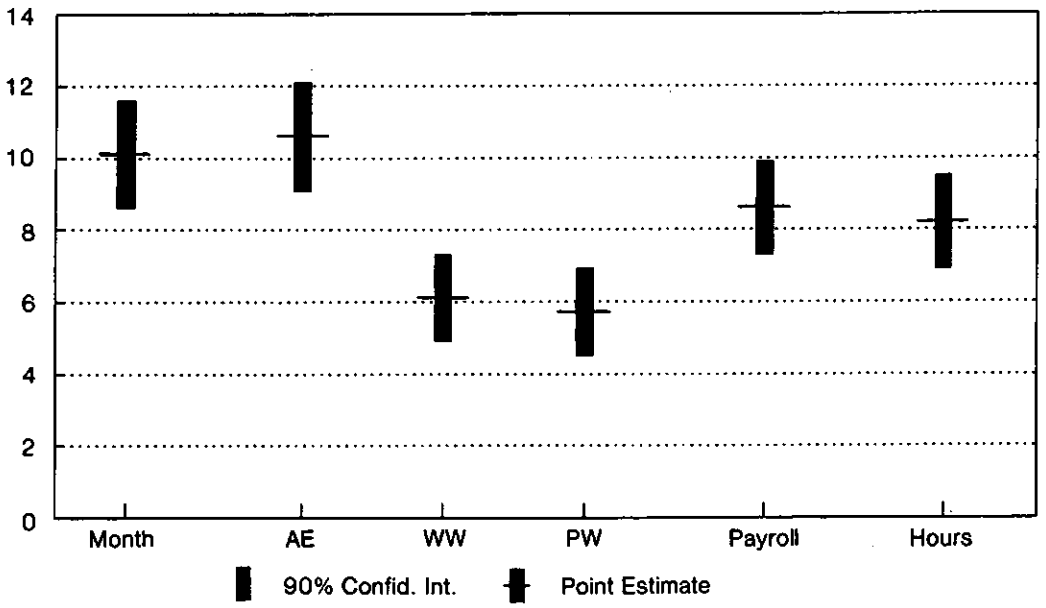
### 5.2 Machine-Recorded Data

The touchtone system provides a tool for assessing difficulties which respondents have with this mode of collection. TDE can record the number of times respondents reenter their data, and how often the question is read back a second or third time before they respond. Information can also be kept on those respondents who hang-up before entering data.

The machine-recorded data were collected for a total sample of 1,203 observations over a three-month period in 1990. There were approximately 474 unique respondents, many of whom provided data for two or three months. There were few differences in the machine-recorded data by month, so all data are presented for the three months combined.

Figure 1 provides, for each data item, the percent of calls for which the question was stated to the respondent more than once. The question could be stated a second time if the respondent does not answer in two seconds (repeat), or if the respondent fails to confirm his or her answer by entering "1," after it is read back (reenter). The figure indicates that the first two questions on the month and all employees, and the payroll and hours questions have higher rates of repeating and reentries by respondents than other data items. The higher rates for the first two questions - each over ten percent - may be due to respondents needing a few questions to orient themselves to the system. The payroll and hours questions generally have the greatest number of digits, so we suspect that data entry errors are more likely to occur, causing the question to be reread and the answer to be reentered.

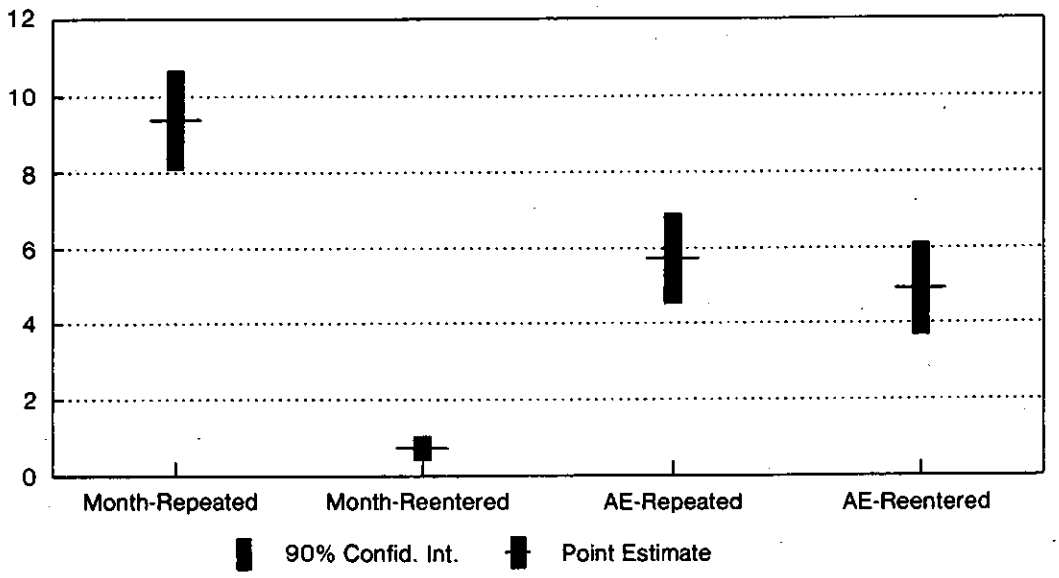
Figure 2 provides data for the first two questions on month and employment for repeated questions after no answer (repeat) and after lack of confirmation (reenter), separately. It was not possible to acquire data separately for the other data items. The CES touchtone system requires respondents to enter at least their report identification number, the month and employment. The system will accept item nonresponse for the other data items. The mandatory entering of month and employment allowed the separation of repeated questions after no answer versus after a respondents lack of confirmation of the answer provided. Of respondents with problems




AE = All Employees
WW = Women Workers
PW = Production Workers

(N = 1,203 per item)

**Figure 1.** Machine Recorded Data - Percent of Questions Repeated/Reentered




AE = All Employees
Repeated = No Answer Within 2 Seconds
Reentered = Correction of Answer

(N = 1,203 per item)

**Figure 2.** Machine Recorded Data - Percent of Questions Repeated Versus Reentered for 1st Two Data Items

on the month question, almost all were due to repeats, that is, two seconds passed without a response. On the other hand, problems with the employment question were almost evenly split between repeats after no previous answers and reentries after lack of confirmation of the previous answer.

Only two percent of the calls received by TDE each month included just a report identification number. During these calls, respondents had simply hung up, or could have been cut off the system.

In addition, the TDE system records all calls received that include at least the report identification number, month and employment. Using the TDE component of the record-check data discussed earlier, we identified respondents with more than one call during a month and coded reasons for the call-backs. In all, about four percent of the respondents called the system more than once in a given month. Most of these respondents provided data items which were not supplied on the initial call (2%). An additional one percent provided corrections to some data items in addition to new data items. Many of these respondents appeared to have had problems with entering the data the first time. Another one percent of the respondents called back only to provide corrections to data items previously supplied or provided identical data. These calls were often several days later, possibly implying that new data had been obtained from their records. In the case of the identical data, respondents may have forgotten whether or not they had previously reported their data. The system currently accepts the data with the latest date and time, although analysts are provided a list of respondents with duplicate records for review, and if necessary, correction.

A common reason for callback seemed to be related to the "enter 1" to confirm after each data item is entered. Many of the respondents who corrected their data had "1" in the data field prior to the callback, and some other response afterwards. Callbacks were twice as common with first-time respondents on the touchtone system than for respondents who were "experienced" users.

A few respondents called in their data three times for a given month, and one respondent called in data four times. These respondents seemed to be having difficulties with the system, but finally reported all of their data correctly.

Overall, these data suggest that respondents are having some difficulties with the system (more than they admit to during the respondent debriefing interview). Some steps could be taken to help alleviate some of the problems. These include providing more time for respondents to answer, providing better instructions, and trying to improve the confirmation of data entry method. In addition, being able to go back to a data item might solve some of the problems.

### 5.3 Respondent Debriefing Survey

BLS interviewers conducted a telephone debriefing survey with TDE respondents during 1990. Given the human-factors literature discussed earlier, some of the questions focused on understanding and pace of the digitized voice. The results from the machine-recorded data showed a substantial number of repeats and retries of questions, thus, questions were developed to address that topic. In addition, respondents were asked to rate the TDE system and answer questions relating to systems design.

The results from the debriefing survey are presented in Table 4. Respondents expressed little difficulty in comprehending the digitized voice. About 97 percent said the voice was very understandable, and all respondents indicated that it was easy to understand the numbers as the voice read them back for confirmation. During the first two months of the survey, we asked respondents about the pace of the interview. Most of the respondents said the pace was about right (88%), although about ten percent felt it was too slow. Our suspicion throughout the

**Table 4**
Results of Debriefing Survey*

| | |
|---|---|
| Voice understandable | 97% |
| Easy to understand #'s read back | 100% |
| Pace about right | 88% |
| Never reentered numbers | 60% |
| Never repeated questions | 83% |
| Never had poor telephone connection | 93% |
| Used speed enhancement feature | 63% |
| Instructions adequate | 98% |
| TDE experience very favorable | 93% |

* N = 411, except for the pace and question repeat items. Approximately 177 respondents were asked about pace and 209 were asked about the repeating of questions.

study was that voice comprehension was much less a problem than were difficulties carrying out the task. While it was difficult to separate out the two in the record-check data, the debriefing interviews lend support to our suspicion.

For task difficulties, 60 percent of respondents said they never had to reenter numbers, while most of the others indicated they had to reenter numbers sometimes. When asked the reasons for reentering numbers, a majority indicated they had accidently entered a wrong number. Others said they did not have enough time, were distracted, or entered their numbers too fast. In the latter several months of interviewing, respondents were asked about the repeating of questions (without reentering data). About 83 percent of the respondents said they never found it necessary to repeat questions. Of the 17 percent who repeated questions, the majority said they were distracted, while others said they did not have enough time.

Most respondents had little difficulty with telecommunications failure, as 93 percent said they never experienced a poor telephone connection when using TDE. Of the respondents who did get a poor connection, most said it happened only once. A large number of the respondents, 63 percent, used the pound sign, a feature of the system designed for speeding up the reporting of data.

Nearly all respondents said the instructions sent to them as they began TDE were adequate. Overall, respondents seemed satisfied with the TDE system – approximately 93 percent rated their experience using TDE as very favorable.

## 6. DISCUSSION

The data show few serious problems with the TDE mode of data collection. Record-check data indicate some item nonresponse error, which is associated with first-time users. Entering additional or incorrect digits appears to be the most serious problem affecting the data items. However, in a panel survey, longitudinal edit checks could reduce this error, as could logical edit checks in all surveys. In addition, the rounding of data needs to be addressed in respondent instructions. Both the record-check and machine-recorded data show that there are more difficulties with longer strings of numbers, probably in both entering data and verifying

incorrect data. The latter could indicate difficulty remembering longer number chains during verification, as comprehension of numbers appeared to be good, *i.e.* respondents said they easily understood numbers being read back for confirmation.

Record-check data show that establishments may have carried over their confirmation of the month into the all-employee question. In addition, the machine-recorded data indicate respondents often do not respond to the month question the first time it is asked. Since respondents appear to be using their survey forms as they enter data, it is likely that moving from the identification number at the top of the form, to the month and data items further down the form, they require extra time to locate themselves. This problem could be solved by placing all information that needs to be entered in one location on the survey form. This might reduce the number of question repeats for the "month" item and potentially lower costs by reducing the length of calls. Question repeats for other items might be reduced by giving respondents more time to respond, since they report they were distracted from the task. However, since most respondents feel the pace of the system is about right, and many are using the speed enhancement feature, adding more time could cause frustration. Probably little can be done to reduce the number of reentries, as respondents indicate they have entered a wrong number and need to correct it.

The data show that errors are reduced with experience. This indicates that a panel survey may be best for this mode of data collection. For surveys requiring numeric or yes/no responses, we believe touchtone also has great potential. The errors are not extremely serious, and respondents rate their experiences using TDE very favorably. TDE may be particularly attractive to business respondents, who can call at convenient times, rather than be interrupted by telephone calls requesting data. However, for some surveys, self initiation and the lack of human contact may be problems which would contribute to nonresponse error.

Although respondent acceptance of touchtone collection is very favorable, there are some steps which can be taken to make the system better. These include:

- giving respondents enough time to key enter their data, especially for the first few questions and those which have a long string of digits,
- investigating ways to improve the confirmation of data items, and
- providing longitudinal edit checks to detect reporting of dollars and cents and other gross errors. The edits could be built into the TDE system with appropriate questions/probes to respondents to correct or confirm their answers.

BLS has used touchtone collection with one other survey. This survey was a small sample follow-up of business establishments who had participated in a Survey of Employer Drug Assistance Programs in 1988. The follow-up survey in 1990 was intended to determine if any substantial changes had occurred in the percentage of establishments providing employer drug assistance programs over the past two years. These establishments were mailed a short survey questionnaire requesting numeric or yes/no answers and encouraged to report their data by touchtone telephone. At the end of the first several weeks of the survey, approximately 20 percent of the establishments had reported their data by touchtone, and an equal amount by mail. TDE was not used after nonresponse follow-up activities began - about two weeks after the initial mailout. The remaining data were collected by telephone (CATI).

We believe that other surveys with time dependent data can take advantage of the time and keypunch savings of touchtone data collection. The mode may communicate the importance of timeliness to the respondent. This paper indicates that measurement errors are controllable using touchtone collection.

Given the timeliness and lower costs of touchtone data collection, we expect it will be used more extensively in the future. We know of two current projects testing touchtone recognition in a survey setting. Statistics Canada is testing a touchtone data collection system for the Survey of Employment, Payroll and Hours. In addition, a touchtone system for a survey of AT & T customers is being developed at Bell Laboratories (Wendler 1990).

BLS is also experimenting with the use of voice recognition technology for data collection in the CES survey (see Winter and Clayton 1990). While touchtone telephones are increasingly available, we estimate that between 10 to 20 percent of our respondents do not have touchtone telephones. Once speaker-independent voice recognition technology reaches an acceptable level for the ten digits needed to report CES data, we expect users will prefer it over touchtone collection. Further work on measurement errors associated with voice recognition technology needs to be undertaken.

## ACKNOWLEDGEMENTS

The authors thank Darrell Philpot and Henry Chiang for their assistance with this research.

## REFERENCES

CLAYTON, R., and HARRELL, L.J., Jr. (1990). Developing a cost model for alternative data collection methods: Mail, CATI and TDE. Presented at the Annual Meeting of the American Statistical Association, Anaheim, California.

COX, A.C., and COOPE, M.B. (1981). Selecting a voice for a specified task: The example of telephone announcements. *Language and Speech*, 24, 233-243.

MARICS, M.A., and WILLIGES, B.H. (1988). The intelligibility of synthesized speech in data inquiry systems. *Human Factors*, 30, 719-732.

PONIKOWSKI, C.H., COPELAND, K.R., and MEILY, S.A. (1989). Applications for touch-tone recognition technology in establishment surveys. Presented at the American Statistical Association Winter Conference, San Diego, California.

SCHWAB, E.C., NUSBAUM, H.C., and PISONI, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.

WATERWORTH, J.A. (1984). Interaction with machines by voice: A telecommunications perspective. *Behaviour and Information Technology*, 3, 163-177.

WENDLER, E.R. (1990). Respondent-initiated computer-directed surveys. Presented at the Annual Conference of the American Association of Public Opinion Research, Lancaster, Pennsylvania.

WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., and ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.

WERKING, G., TUPEK, A., and CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 4, 349-362.

WINTER, D.L.S., and CLAYTON, R.L. (1990). Speech data entry: Results of the first test of voice recognition for data collection. Bureau of Labor Statistics, Washington, D.C.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 27–41
Statistics Canada



# Avoiding Sequential Sampling with Random Digit Dialing

J. MICHAEL BRICK and JOSEPH WAKSBERG[1]

ABSTRACT

The Mitofsky-Waksberg procedure is an efficient method for selecting a self-weighting, random digit dialing (RDD) sample of households. The Mitofsky-Waksberg procedure is sequential, requiring a constant number of households be selected from each cluster. In this article, a modified Mitofsky-Waksberg procedure which is not self-weighting or sequential is described. The bias and variance for estimates derived from the modified procedure are investigated. Suggestions on circumstances which might favor the modified procedure over the standard Mitofsky-Waksberg procedure are provided.

KEY WORDS: Random digit dialing; Telephone sampling; Cluster sampling; Trimming.

## 1. INTRODUCTION

The Mitofsky-Waksberg procedure for selecting random digit dialing samples of households (Waksberg 1978) is frequently used for sample selection in telephone surveys. As described in the Waksberg paper, it is an efficient method of producing a self-weighting sample, that is, one in which all telephone households have the same probability of selection (except for households with more than one telephone number). The efficiency is due to the sharp reduction in the proportion of nonhousehold telephone numbers that have to be dialed in order to identify sample households.

The Mitofsky-Waksberg procedure is a two-stage sample design. In the first stage, a sample of clusters is chosen where the clusters consist of blocks of 100 telephone numbers, or multiples of such blocks. The clusters (or blocks of 100 telephone numbers) are first selected with equal probability. One telephone number is chosen at random in each cluster and dialed. If the number is that of a household, the cluster is retained. Otherwise, it is rejected. The second stage is the selection of households within the retained sample clusters. For the self-weighting feature of the sample to apply, a constant number of households per cluster is required. Some organizations (including Westat Inc.) generally go a little further and specify a constant number of interviewed households per cluster (or screened households if the first part of data collection is screening). The rationale is that substituting another randomly selected household within the same cluster for each nonrespondent is a reasonable way of reducing nonresponse bias.

There is an awkward operational feature to this system. It sometimes takes a fairly large number of callbacks to determine whether or not a telephone number is residential, particularly for numbers that repeatedly ring with no answer. Even more are needed to learn which households cooperate. Such determinations must be made for an initially selected sample to ascertain which clusters require more telephone numbers to achieve the desired cluster size and how many telephone numbers have to be added. In effect, a sequential scheme is necessary for each cluster, where all previous cases need to be cleared up before it is known whether the sample needs to be increased. This process is particularly inconvenient when there is a tight time schedule for data collection.

Several attempts to modify the Mitofsky-Waksberg method have been proposed which reduce or eliminate the sequential features of the plan. Potthoff (1987) developed a generalization

[1] J. Michael Brick and Joseph Waksberg, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850, U.S.A.

of the Mitofsky-Waksberg technique in which $c$ telephone numbers are chosen per cluster in determining whether to retain the cluster, whereas Mitofsky-Waksberg use only one. A self-weighting sample is achieved by having clusters in which only one of the $c$ telephone numbers dialed continue with a sampling plan that includes having a fixed number of households per cluster, and the remaining clusters having a fixed number of telephone numbers. The latter group of clusters does not require a sequential approach. Potthoff reports that in practice most clusters will fall into the second class so that the sequential operations, although not eliminated, are sharply reduced.

Lepkowski and Groves (1986) describe a sampling method in which blocks of telephone numbers with more than a trivial number of telephone numbers listed in directories (and other sources, if available) are selected with probability proportionate to the number of listed numbers. Blocks of numbers which contain zero or very few telephone numbers are sampled through the Mitofsky-Waksberg procedure. Sudman (1973) had previously proposed sampling blocks of numbers with probability proportionate to the numbers listed in directories, but without making any provision for empty blocks (which could have unlisted numbers). Drew and Jaworski (1986) describe an RDD survey carried out in Canada in which purchased counts of residential numbers (both published in directories and nonpublished) were used as measures of size. Since the counts were considered as virtually complete, there was no need to sample empty blocks. As far as we are aware, there is no way of getting virtually complete counts of residential numbers in the U.S.

Neither the Potthoff nor the Lepkowski-Groves sample design completely eliminates the need for a sequential process, although both appear to reduce the portion of the sample which requires it. There are also some other disadvantages to the two procedures. The Potthoff technique appears to be rather complex – as far as we know it has not been used much for RDD surveys. For national surveys, the Lepkowski-Groves technique involves the purchase of a directory list covering the total U.S. and processing it to obtain measures of size. Such commercial lists are available, but they are expensive. Furthermore, a number of recent reports indicate the percentage of all residential numbers that are listed in directories is not very high, and is rapidly decreasing. Tucker (1989) describes an analysis of listed numbers in a group of U.S. counties and cities which shows listing rates varying from 48 to 62 percent. An article by Linda Piekarski (1990) states that if the rate of increase of unlisted numbers continues at the current level, "as many as 62% of the nation's households may be unlisted by the year 2000." The measures of size thus are probably only moderately correlated with the actual number of households in a working block.

Waksberg has suggested an alternative modification of the Mitofsky-Waksberg procedure (Waksberg 1984) which completely eliminates the need for sequential sampling. Westat has used this method in a large number of studies using RDD. Cummings (1979) had previously used the same procedures as a result of an error in implementing the Mitofsky-Waksberg procedure. Cummings did not recognize its usefulness in avoiding sequential sampling and did not explore its features for use in other surveys. We describe the method and its mathematical and statistical properties.

## 2. ALTERNATIVE METHOD OF ESTABLISHING CLUSTER SIZES WITH MITOFSKY-WAKSBERG TECHNIQUE

As indicated earlier, the Mitofsky-Waksberg technique requires a constant number of sample residential numbers per cluster (or block of numbers) to produce a self-weighting sample. The alternative that is proposed is to use a constant number of telephone numbers per cluster for

the sample ($K$). The first stage of selection is unchanged. (The first-stage selects clusters with probability proportionate to the number of households.) With a constant number of telephone numbers per cluster, the sample numbers can be designated in advance eliminating the sequential process. We note that followup effort is still necessary to determine which sample telephone numbers are residential, both in the first and in the second stages of sampling. However, this has to be done for a fixed set of telephone numbers. A sequential process is not involved.

The alternative procedure does not produce a self-weighting sample. Since the first stage is selected with PPS, the probability of a cluster being selected is $r\ N_i/100$ where $r$ is the sampling rate for selection of the clusters, that is, the first stage selection rate, and $N_i$ is the number of residential numbers in the $i$th cluster. The weight should be proportional to $N_i^{-1}$, but since $N_i$ is not known, it is taken to be proportional to $n_i^{-1}$, the number of sample households in the cluster.

This modification of the Mitofsky-Waksberg method has good features for survey operations. It is simple. The sample can be virtually preselected and no costly control operations are needed. Although weighting is required, the weights are directly available from the sample data, and they can be mechanically produced without any extensive professional oversight.

There are, however, some serious problems. First, there is a bias when $N_i^{-1}$ is estimated by $K/100n_i$ where $K$ is the number of telephone numbers selected per cluster (a constant number in all clusters). The bias is fairly small, but it does exist. It cannot be eliminated or reduced by minor modifications of the weights, such as using $1/(n_i + t)$ instead of $n_i^{-1}$, with "$t$" denoting a fixed constant. Secondly, the introduction of variable weights increases the variances of the estimates substantially. (The increase is not so much caused by the weights as by the fact they reflect variable probabilities of selection.) Finally, the modification loses one of the useful features of the Mitofsky-Waksberg method – the ability to fix the exact sample size desired. The Mitofsky-Waksberg method's use of a constant number of households per cluster means that any desired sample size can be obtained by selecting a sample with the appropriate number of clusters. With the modification, the sample size becomes a random variable, which generally will not be exactly equal to the desired sample size. Although the deviations are usually small, the ability to achieve exact target sizes is useful when contracts or budget commitments require the survey organization to satisfy exact target requirements. We discuss these issues in Sections 3 and 4.

Before going on to a discussion of the variances and biases, it is useful to examine the distribution of cluster sizes in the U.S. Tables 1 to 3 show estimates of such distributions prepared from data reported in two large national U.S. surveys conducted via RDD by Westat Inc. Both of these surveys used the modification of the Mitofsky-Waksberg procedure described above. The sample for the survey summarized in Table 1 was selected in 1986 and consisted of 2,427 clusters (retained after first-stage sampling) with 15 telephone numbers per cluster, or 36,405 total numbers. There were 18,756 completed screeners, 2,396 refusals, 1,727 nonresponse for other reasons, and 13,526 nonresidential or nonworking numbers, ring no answers, and cases that could not be classified. The analysis is restricted to the 18,756 completed cases. The data in Tables 2 and 3 are based on a 1989 sample of 1,000 clusters with 30 telephone numbers per cluster or 30,000 telephone numbers, of which 19,586 were residential with screeners completed. Table 2 shows the distribution of the 15,030 completed cases and Table 3 shows the distribution of the 19,586 residential numbers found in the 1,000 clusters. The cluster weights shown are expressed as $\bar{n}/n_i$ where $\bar{n}$ is the average number of households per cluster. It seems useful to express them in this form since they then show the deviations from a self-weighting sample. The design effects only account for the increased variances arising from variable sampling fractions. They do not include effects of other aspects of the sample design.

**Table 1**

Number of Completed Screeners per Cluster in 1986 Survey
(Based on sample of 2,427 clusters with 15 telephone numbers per cluster)

| Number of Completes per Cluster | Average Cluster Weight[1] | Household Distribution | | | Cluster Distribution | | |
|---|---|---|---|---|---|---|---|
| | | Frequency | Percent | Cumulative Percent | Frequency | Percent | Cumulative Percent |
| 0 | xx | 0 | 0 | 0.0 | 62 | 2.6 | 2.6 |
| 1 | 7.93[2] | 54 | 0 | 0.3 | 54 | 2.2 | 4.8 |
| 2 | 3.97[2] | 106 | 0.6 | 0.9 | 53 | 2.2 | 7.0 |
| 3 | 2.64 | 258 | 1.4 | 2.2 | 86 | 3.5 | 10.5 |
| 4 | 1.98 | 440 | 2.3 | 4.6 | 110 | 4.5 | 15.0 |
| 5 | 1.59 | 810 | 4.3 | 8.9 | 162 | 6.7 | 21.7 |
| 6 | 1.32 | 1,290 | 6.9 | 15.8 | 215 | 8.9 | 30.6 |
| 7 | 1.13 | 1,960 | 10.5 | 26.2 | 280 | 11.5 | 42.1 |
| 8 | 0.99 | 2,656 | 14.2 | 40.4 | 332 | 13.7 | 55.8 |
| 9 | 0.88 | 2,862 | 15.3 | 55.6 | 318 | 13.1 | 68.9 |
| 10 | 0.79 | 2,990 | 15.9 | 71.6 | 299 | 12.3 | 81.2 |
| 11 | 0.72 | 2,717 | 14.5 | 86.1 | 247 | 10.2 | 91.4 |
| 12 | 0.66 | 1,548 | 8.3 | 94.3 | 129 | 5.3 | 96.7 |
| 13 | 0.61 | 780 | 4.2 | 98.5 | 60 | 2.5 | 99.2 |
| 14 | 0.57 | 210 | 1 | 99.6 | 15 | 0.6 | 99.8 |
| 15 | 0.53 | 75 | 0 | 100.0 | 5 | 0.2 | 100.0 |
| Total | | 18,756 | 100.0 | xx | 2,427 | 100.0 | xx |
| Mean cluster size[3] | | | 7.93 | | | | |
| Design effect[4] | | | 1.31 | | | | |

[1] The cluster weight is the mean cluster size (*i.e.*, 7.93) divided by the number of completes in the *i*-th cluster.
[2] Trimming the weights would bring these weights down to 3.
[3] The mean cluster size is the average over the 2,365 clusters with one or more completed screeners.
[4] The design effect is reduced to 1.12 if the maximum weight is 3.

It should be noted that Table 1 is based on a sample of 15 telephone numbers per cluster and Tables 2 and 3 used 30 telephone numbers per cluster. Estimates of the percent residential in a cluster based on 15 telephone numbers will, of course, be subject to a higher sampling error than an estimate based on 30 telephone numbers. However, the number of clusters used in Table 1 was more than twice those in Tables 2 and 3 which should largely offset the effect of the different cluster sizes.

There are two differences between Tables 2 and 3. One is that Table 2 shows the distribution of completed screeners (as does Table 1) while Table 3 is based on all sample households. The use of only completed cases in Table 2 reduces the estimate of the average number of households per cluster and shifts the entire distribution. In addition, it introduces more variability to the estimates of the distribution shown because the distributions reflect sampling errors of both the distribution of households per cluster and the distribution of nonresponse rates per cluster. The second difference is that Table 2 (and Table 1) is expressed in terms of the number of cases per cluster and Table 3 shows the distributions by the percentage of residential numbers per cluster. It was convenient to express Table 3 in that form for analyses described later in this report.

One other feature of the percentages shown in Tables 1 to 3 should be noted. They reflect the size distributions of clusters which fell into the sample, not the distribution of clusters in the U.S. The use of probability proportionate to size sampling results in an oversampling of clusters with a high proportion of residential numbers and an underrepresentation of clusters with a small number. It is possible to convert the distribution from one that represents the sample to one that represents the population by multiplying each percentage by the cluster weights and computing the percentage distribution of the resulting figures. Since the weights are exactly proportional to the reciprocal of the number of completes per cluster, it turns out that converting the household distribution so that it represents the distribution in the population produces the percentages shown in the cluster distribution. The cluster distribution in the sample is thus the same as the household distribution in the population.

We show distributions of both all-sample households and completed cases because both are of interest to researchers. The Table 3 data have been used for the analyses in Sections 3 and 4.

**Table 2**

Number of Completed Screeners per Cluster in 1989 Survey
(Based on sample of 1,000 clusters with 30 telephone numbers per cluster)

| Number of Completes per Cluster | Average Cluster Weight[1] | Household Distribution | | | Cluster Distribution | | |
|---|---|---|---|---|---|---|---|
| | | Frequency | Percent | Cumulative Percent | Frequency | Percent | Cumulative Percent |
| 0 | xx | 0 | 0 | 0.0 | 8 | 0.8 | 0.8 |
| 1 or 2 | 7.57[2] | 6 | 0 | 0.0 | 3 | 0.3 | 1.1 |
| 3 or 4 | 4.33[2] | 37 | 0.2 | 0.3 | 10 | 1.0 | 2.1 |
| 5 or 6 | 2.75 | 126 | 0.8 | 1.1 | 22 | 2.2 | 4.3 |
| 7 or 8 | 2.02 | 403 | 2.7 | 3.8 | 53 | 5.3 | 9.6 |
| 9 or 10 | 1.59 | 688 | 4.6 | 8.4 | 72 | 7.2 | 16.8 |
| 11 or 12 | 1.32 | 1,325 | 8.8 | 17.2 | 115 | 11.5 | 28.3 |
| 13 or 14 | 1.12 | 1,987 | 13.2 | 30.4 | 147 | 14.7 | 43.0 |
| 15 or 16 | 0.98 | 2,636 | 17.5 | 50.0 | 170 | 17.0 | 60.0 |
| 17 or 18 | 0.85 | 2,692 | 17.9 | 65.9 | 154 | 15.4 | 75.4 |
| 19 or 20 | 0.78 | 2,387 | 15.9 | 81.8 | 123 | 12.3 | 87.7 |
| 21 or 22 | 0.70 | 1,673 | 11.1 | 92.9 | 78 | 7.8 | 95.5 |
| 23 or 24 | 0.64 | 816 | 5.4 | 98.3 | 35 | 3.5 | 99.0 |
| 25 or 26 | 0.55 | 254 | 1.7 | 100.0 | 10 | 1.0 | 100.0 |
| 27 or 28 | xx | 0 | 0 | 100.0 | 0 | 0 | 100.0 |
| 29 or 30 | xx | 0 | 0 | 100.0 | 0 | 0 | 100.0 |
| Total | xx | 15,030 | xx | xx | 1,000 | xx | xx |
| Mean cluster size[3] | | | | 15.11 | | | |
| Design effect[4] | | | | 1.33 | | | |

[1] The cluster weight is the mean cluster size (*i.e.*, 15.15) divided by the number of completes in the *i*-th cluster.
[2] Trimming the weights would bring these weights down to 3.
[3] The mean cluster size is the average over the 992 clusters with one or more completed screeners.
[4] The design effect is reduced to 1.12 if the maximum weight is 3.

**Table 3**

Proportion of Residential Numbers per Cluster in 1989 Survey
(Based on sample of 1,000 clusters with 30 telephone numbers per cluster)

| Proportion of Residential nos. per Cluster | Average Cluster Weight[1] | Household Distribution | | | Cluster Distribution | | |
|---|---|---|---|---|---|---|---|
| | | Frequency | Percent | Cumulative Percent | Frequency | Percent | Cumulative Percent |
| 0 | xx | 0 | 0.0 | 0.0 | 5 | 0.5 | 0.5 |
| .001 to .049 | 21.76[2] | 5 | 0.0 | 0.0 | 3 | 0.3 | 0.8 |
| .05 to .099 | 8.70[2] | 18 | 0.1 | 0.1 | 6 | 0.6 | 1.4 |
| .10 to .149 | 5.22[2] | 41 | 0.2 | 0.3 | 9 | 0.9 | 2.3 |
| .15 to .199 | 3.73[2] | 48 | 0.2 | 0.6 | 8 | 0.8 | 3.1 |
| .20 to .249 | 2.90 | 53 | 0.3 | 0.8 | 7 | 0.7 | 3.8 |
| .25 to .299 | 2.37 | 144 | 0.7 | 1.6 | 16 | 1.6 | 5.4 |
| .30 to .349 | 2.01 | 178 | 0.9 | 2.5 | 17 | 1.7 | 7.1 |
| .35 to .399 | 1.74 | 408 | 2.1 | 4.6 | 34 | 3.4 | 10.5 |
| .40 to .449 | 1.54 | 459 | 2.3 | 6.9 | 34 | 3.4 | 13.9 |
| .45 to .499 | 1.37 | 840 | 4.3 | 11.2 | 56 | 5.6 | 19.5 |
| .50 to .549 | 1.24 | 1,040 | 5.3 | 16.5 | 63 | 6.3 | 25.8 |
| .55 to .599 | 1.14 | 1,926 | 9.8 | 26.3 | 107 | 10.7 | 36.5 |
| .60 to .649 | 1.04 | 2,126 | 10.9 | 37.2 | 109 | 10.9 | 47.4 |
| .65 to .699 | 0.97 | 3,255 | 16.6 | 53.8 | 155 | 15.5 | 62.9 |
| .70 to .749 | 0.90 | 2,610 | 13.3 | 67.1 | 116 | 11.6 | 74.5 |
| .75 to .799 | 0.84 | 3,022 | 15.4 | 82.6 | 126 | 12.6 | 87.1 |
| .80 to .849 | 0.79 | 1,556 | 7.9 | 90.5 | 61 | 6.1 | 93.2 |
| .85 to .899 | 0.75 | 1,458 | 7.4 | 98.0 | 54 | 5.4 | 98.6 |
| .90 to .949 | 0.71 | 399 | 2.0 | 100.0 | 14 | 1.4 | 100.0 |
| .95 to .999 | xx | 0 | 0.0 | 100.0 | 0 | 0.0 | 100.0 |
| Total | xx | 19,586 | 100.0 | xx | 1000 | 100.0 | xx |
| Mean cluster size[3] | | | 19.68 | | | | |
| Design effect[4] | | | 1.28 | | | | |

[1] The cluster weight is the mean proportion in a cluster (*i.e.*, 0.653) divided by the proportion of residential numbers in the $i$-th cluster.
[2] Trimming the weights would bring these weights down to 3.
[3] The mean cluster size is the average over the 995 clusters with one or more residential numbers.
[4] The design effect is reduced to 1.12 if the maximum weight is 3.

## 3. VARIANCE IMPLICATIONS OF THE MODIFIED MITOFSKY-WAKSBERG METHOD

In the standard Mitofsky-Waksberg method the variance of a sample estimate is dependent upon the number of households selected per cluster and the homogeneity of the households within and between clusters. The variance for a cluster sample can be written as the variance for a simple random sample multiplied by $[1 + \rho(\bar{n} - 1)]$, where $\rho$ is intraclass correlation and $\bar{n}$ is the average number of households per cluster. Since telephone clusters are often related to geographic areas and tend to be somewhat homogeneous, selecting a large number of households per cluster can be inefficient.

When the modified Mitofsky-Waksberg method is used, another source of variance is introduced because the number of households selected per cluster is allowed to vary from cluster to cluster. As pointed out in Section 2, the denominator of the second stage probability of selection does not cancel with the number of households in the cluster (which is proportional to the probabilities in the first stage) and the overall probabilities of selecting households vary from cluster to cluster.

The variability among clusters in the overall household sampling rates causes the variances of the estimates to be larger than those in the standard Mitofsky-Waksberg method where each household has the same probability of selection. Methods for estimating the increase in the variance of an estimate arising from unequal probabilities of selection are discussed by Kish (1965) and by Waksberg (1973). A simple approximation to the variance of an estimate under an unequal weighting scheme (where the weights do not reflect variable sampling rates in strata deliberately chosen to reduce sampling variances) is the sampling variance which would occur with a self-weighting sample multiplied by a variance inflation factor (VIF), given by VIF = {1 + Relvar(weights)}. We will use this approximation below to investigate the variance implications associated with the modified Mitofsky-Waksberg method.

The relative variance of the weights was computed by partitioning the process into two components. First, the mean and variance of the weights were computed conditioned upon sampling from a truncated (since zero households cannot be obtained if the cluster is sampled in the first stage) hypergeometric distribution, defined by the household density in the cluster and the cluster sample size. The unconditional mean and variance of the weights were then computed by integrating over the distribution of households in the sampled clusters shown in Table 3. The distribution of households in the sample is critical in the evaluation of the VIF.

The natural weight assigned to a household in the modified Mitofsky-Waksberg is proportional to $n_i^{-1}$, where $n_i$ is the number of households observed in sample cluster $i$. This weight can vary by factors which range from as little as $1/K$ to 1, where $K$ is the number of telephone numbers selected in a cluster. The average weight is roughly $1.5/K$, since about 65 percent of numbers in the sampled clusters are residential.

If the number of telephone numbers sampled per cluster is between 5 and 30, then the increase in variance due to the weighting is about 30 percent. The VIF decreases slightly as the number sampled per cluster increases beyond 30, reaching approximately 17 percent when all the numbers in the cluster are sampled.

The VIF or the relative variance of the weights is a function of the distribution of the number of households across clusters and random sampling variability within the clusters. This decomposition is made explicit by expressing the variance of the weights as the sum of the mean of the conditional variance of the weights and the variance of the conditional mean of the weights, where the conditioning is with respect to the household density of the cluster.

When the cluster sample size is small, the mean of the conditional variance is the dominant component of the overall variance. As the cluster sample size increases, the variance of the conditional mean becomes more dominant. This is why the relative variance of the weights, shown in the first row of Table 4, is not a monotonic function of the cluster sample size.

**Table 4**

Approximate Variance Inflation Factors (VIF) for Modified Mitofsky-Waksberg Random Digit Dialing Samples

| Weight | Cluster Sample Size ($K$) | | | | |
|---|---|---|---|---|---|
| | 5 | 10 | 30 | 60 | 100 |
| $1/n_i$ | 1.31 | 1.34 | 1.29 | 1.23 | 1.17 |
| $1/(n_i + .5)$ | 1.18 | 1.21 | 1.20 | 1.18 | 1.16 |
| $1/(n_i + 1)$ | 1.12 | 1.15 | 1.16 | 1.15 | 1.14 |
| $1/(n_i + 2)$ | 1.07 | 1.09 | 1.11 | 1.12 | 1.13 |

### Variances Using Different Weights

Weights other than ones proportional to the inverse of the number of households were also examined to determine their impact on the bias and variance of the estimates. Many of the alternative weights studied were derived from variance stabilizing transformations suggested for binomial variables.

Of all the alternatives examined, the estimators with the best bias and variance properties involved simple adjustments of the natural weight. In particular, adding a small constant to the observed number of households (estimators of the form $(n_i + t)^{-1}$ where $t$ is .5, 1, or 2) resulted in reducing the increases in variance due to differential weighting. The addition of the constant reduces the range of the weights by cutting the values of the largest weights while only slightly modifying the weights for clusters where more households are found in the sample.

Table 4 shows the VIF for the estimators of the form $(n_i + t)^{-1}$ for different numbers of telephone numbers sampled per cluster. The table also is based on the household and cluster distributions shown in Table 3. It is clear from the table that a substantial reduction in the variance due to unequal weighting can be achieved by using $(n_i + 1)^{-1}$, rather than the natural estimator. This is especially true for RDD designs which sample 30 or fewer telephone numbers per cluster. The increase in variance due to differential weighting for $(n_i + 1)^{-1}$ is only 16 percent when 30 numbers are selected per cluster as opposed to a 29 percent increase when the natural estimator is used.

### Variances with Trimmed Weights

A practice that is often used to mitigate the variance inflation associated with varying weights is the truncation of very large weights. This truncation, or trimming of weights, is usually fixed at a weight above which relatively few observations are found. In many Westat RDD samples, weights that exceed two or three times the mean weight have been truncated. For this research, we have examined weights truncated at about 3 times the mean weight. For samples of 10 per cluster, the weights were truncated at 2 times the mean weight because so few observations are affected otherwise.

Table 5 shows the VIF for the estimators for different cluster sample sizes when the weights are trimmed at three times the mean weight for $n_i^{-1}$. The VIF's for samples of 5 per cluster are not given because the truncation point in samples of this size is nearly at unity, the largest possible weight.

**Table 5**

Approximate Variance Inflation Factors (VIF) for Modified Mitofsky-Waksberg Random Digit Dial Samples with Trimmed* Weights

| Weight | Cluster Sample Size ($K$) | | | |
|---|---|---|---|---|
| | 10 | 30 | 60 | 100 |
| $1/n_i$ | 1.12 | 1.11 | 1.09 | 1.09 |
| $1/(n_i + .5)$ | 1.11 | 1.10 | 1.09 | 1.09 |
| $1/(n_i + 1)$ | 1.09 | 1.10 | 1.09 | 1.09 |
| $1/(n_i + 2)$ | 1.07 | 1.09 | 1.09 | 1.08 |

* All weights trimmed at 3 times the mean weight, except samples of 10 trimmed at 2 times the mean.

The tabled values show trimming substantially reduces adverse impact of the differential weights on the variance of the estimates. The most dramatic reduction is for the natural estimator; its VIF is reduced by over 50 percent by the use of trimming. The VIFs for the other estimators are improved somewhat, but the reductions are less striking since they already had smaller VIF's than the natural estimator. Trimming has the potential of introducing biases which may counteract the advantage in variance reduction. Biases are discussed in Section 4.

**Variances with Augmented Sampling**

A third approach to reducing the variability of the weights is the use of augmented sampling. Large weights occur when the number of households identified in the cluster is small relative to the expected number of households per cluster. To reduce the chance for this happening, an augmented sampling procedure can be used. If the number of households in a cluster is smaller than a fixed number (say less than one third of the mean number per cluster), then the sample size in the cluster can be doubled or increased by some other amount.

This procedure could be iterated to insure that the number of households per cluster reaches a specified limit or until all numbers in the cluster are used. The obvious disadvantage of this iterative plan is that it requires monitoring sample yield by cluster and the very fact that it is sequential. Another disadvantage of the method is that it results in sampling more telephone numbers from clusters that have a lower household density (the ones most likely to need augmentation), hence reducing productivity.

Despite the operational shortcomings of the augmented sampling approach, we did a limited examination of the method. Since the results for the augmented sample approach was not better than trimming the weights, this method is not discussed further.

## 4. BIAS IMPLICATIONS OF THE MODIFIED MITOFSKY-WAKSBERG METHOD

The increase in variance is just one of the consequences of using the modified Mitofsky-Waksberg method of sampling. Another important feature of the method is the bias in the resulting estimates. If a fixed sample size is selected in a cluster and no weight adjustment is made, the variance of the estimates are not increased but the bias has the potential of being very large.

The unbiased weight ($W_u$) for the modified method is

$$W_u = \frac{100}{rN_i} \times \frac{100}{K},$$

where the terms are as defined above. The problem is that $N_i$ is unknown and does not cancel with the second stage term, as it does in the standard Mitofsky-Waksberg method. Weights are therefore introduced in an effort to reduce the bias.

We refer to the estimator which uses a weight of $n_i^{-1}$ as the natural estimator because $n_i/K$ is an unbiased estimator of $N_i/100$ in sampling from a binomial or hypergeometric distribution. (We use the weight of $n_i^{-1}$ although the weight is actually $K/100n_i$. Since $K/100$ is a constant, the relationship among the weights are not affected by using $n_i^{-1}$.) This weight appears to be the natural estimator despite the fact that $n_i^{-1}$ is not unbiased for $N_i^{-1}$ unless all 100 numbers are selected in a cluster. The bias of $n_i^{-1}$ is discussed in literature; for example, see the discussion on stratification after sampling in Hansen, Hurwitz and Madow (1953). No simple unbiased estimator, certainly none of the form $(n_i + t)^{-1}$, is likely to exist for all possible cluster sample sizes.

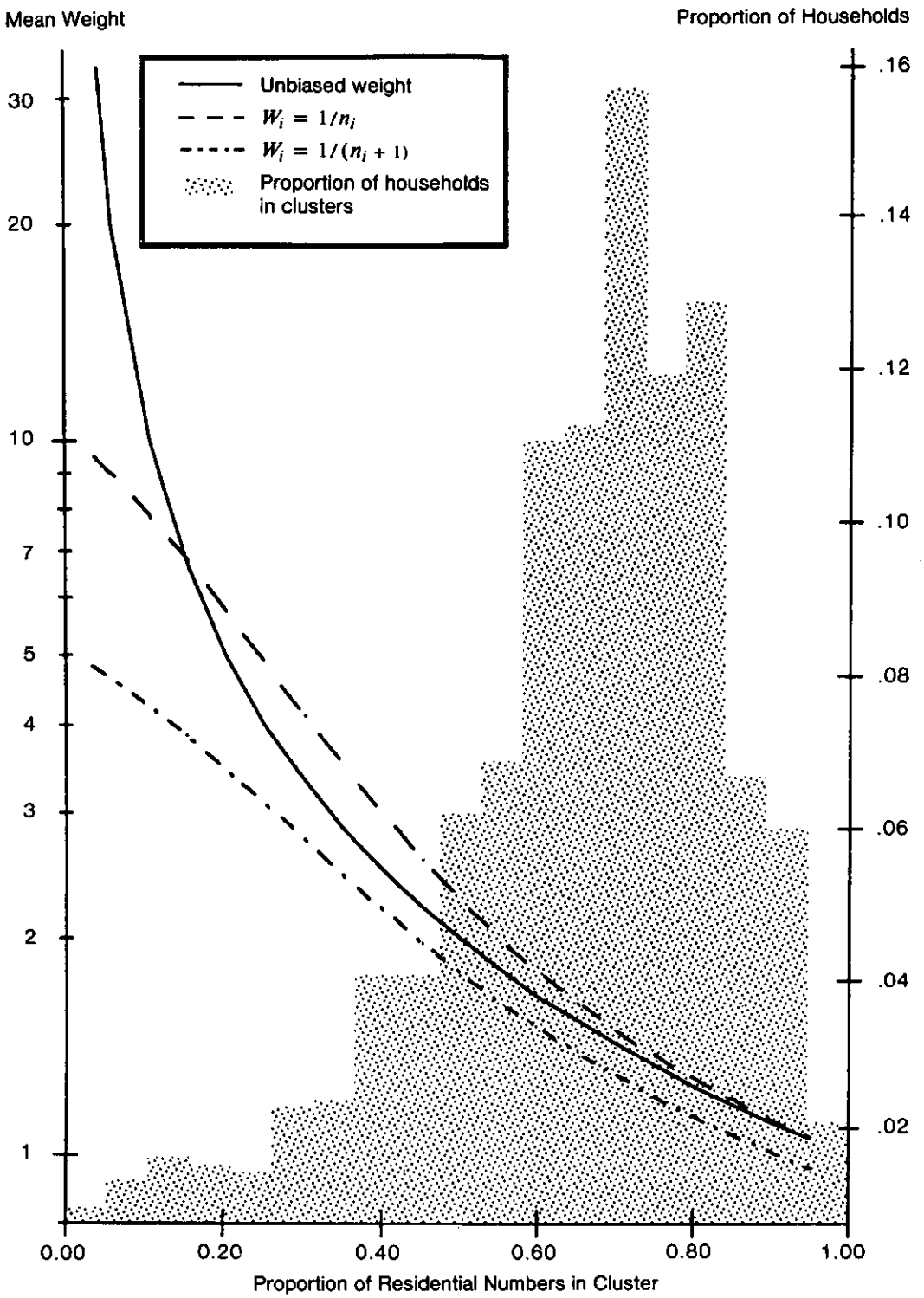



**Figure 1.** Mean Weights of Estimators Conditional on the Proportion Residential with Shaded Histogram of Proportion of Households in Cluster

One of the ways to examine the potential bias is by comparing the expected value of the estimators (the mean weight using estimators of the form $(n_i + t)^{-1}$) with the unbiased weight, $W_u$. Since both the unbiased weight and the expected value of the estimators are functions of $N_i$, we will begin by investigating these quantities conditioned on $N_i$.

Figure 1 shows the graph of the unbiased weight and the mean weights, using the estimators $n_i^{-1}$ and $(n_i + 1)^{-1}$, when there are $K = 10$ telephone numbers selected per cluster. The constant cluster sampling rate, $r$, has been omitted from all of the weights. A logarithmic scale has been used for the mean weights because of the range in $W_u$.

The graph clearly shows that the biggest differences between $W_u$ and the mean weights for the two estimators are found when $N_i/100$ is small. Once the residential density exceeds 20 percent when $(n_i + 1)^{-1}$ is used, and 10 percent for $n_i^{-1}$, the differences are relatively minor. The graph shows that the weight $(n_i + 1)^{-1}$ is always smaller than $W_u$, but this will not be true if poststratification is used. Poststratified weights are not used in the graph because poststratification really operates on the unconditional weights rather than the conditional weights shown here. The unconditional bias is addressed below.

The shaded histogram in the figure shows the distribution of households from Table 3. It has been overlaid to illustrate the fact that the large differences in weights occur in clusters which account for a very small fraction of the sampled households.

**Bias in Sample Size and Bias in Estimates**

In nearly all RDD surveys, including those using the Mitofsky-Waksberg sample design, poststratification of the sample to known totals of persons or households is used. One of the prime reasons for using poststratification is to adjust the estimates to the levels associated with all persons, not just those in households with telephones. Massey and Botman discuss this and other benefits of poststratification in RDD surveys in Chapter 9 of Groves *et al.* (1988).

Regardless of the reasons for using it, poststratification results in estimates that are equal to known totals irrespective of the weights applied to the individual households. Since this bias, which can be considered as bias in sample size, is always zero, it is difficult to find a single statistic that measures unconditional bias directly. To attack this problem, we will examine the relative contribution to the bias in sample size over the range of household density values.

The following steps were taken to compute a measure of this contribution to bias in sample size. First, the different weighting functions or estimators were computed using the empirical household density shown in Table 3. Then, the estimates were poststratified to equal unity and the contribution to the total was computed for different values of $N_i/100$. Finally, the relative bias in sample size was defined as the difference between the contribution to the total from the particular estimator and the contribution from the total using $W_u$ as the weight.

This measure thus takes into account both the difference in the weights for fixed values of $N_i$ and the distribution of households across all the values of $N_i$. Thus, sampled households from clusters with values of $N_i$ that are rare will not contribute heavily to the relative bias in sample size even if they are associated with large differences in weights.

To illustrate these computations, Figure 2 shows the relative bias in sample size for some estimators for samples of 30 numbers per cluster. One of the estimators uses the unadjusted weight, *i.e.*, the weight is a constant for all households regardless of the number of households identified in a cluster. The relative bias in sample size for the estimator with unadjusted weights is much larger than when other weights are used. The unadjusted weight has relative biases in sample size that range from about $-2$ percent to $+3$ percent.

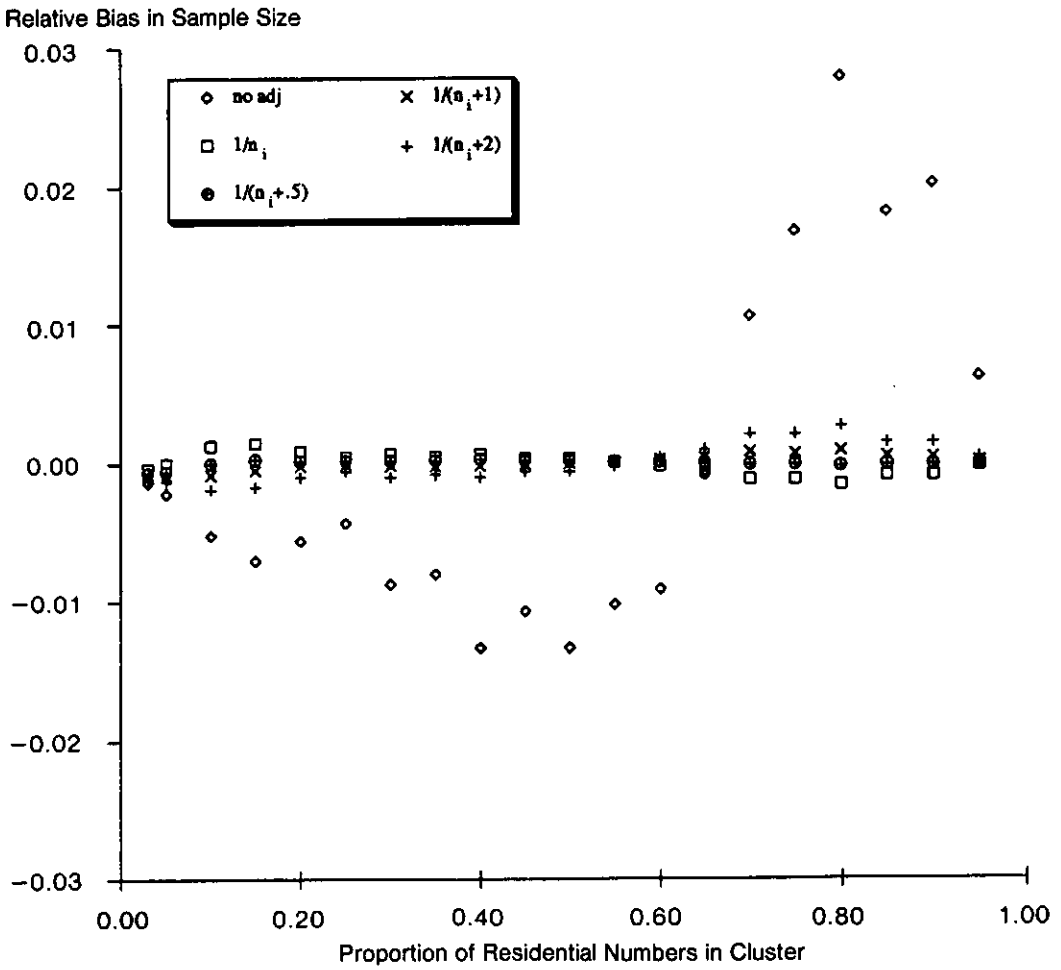



**Figure 2.** Relative Bias from Sample Size for Samples of 30 Cluster

The size of the bias in the estimate of a characteristic is bounded by the size of the bias in sample size. In other words, the relative bias in the estimate can be no larger than the relative bias in the sample size. For almost all characteristics, this upper bound will not be attained. The upper bound is only attained when the characteristic and the residential density are perfectly correlated. Very high correlations are not likely in national samples, but might be more feasible in samples in restricted geographic areas.

It can be seen that there are patterns in the biases; for example, the unadjusted estimator is uniformly too low in low proportion residential clusters and too high in clusters with a high proportion of households. When there are differences in the characteristics between low and high density clusters, the biases can be quite serious. The bias in estimates resulting from using unadjusted weights can be seen for some characteristics in Table 1 in Cummings (1979). The biases are not very large, but appropriate weighting will effectively eliminate them.

In general, the relationship between the estimate and the number of households in a cluster will be unknown and inconsistent across all the characteristics to be estimated. Therefore, a reasonable practice is to choose an estimator that has a relative bias in sample size that is small across the range of values of $N_i$. If the relative bias in sample size for a set of the estimators is small, then the choice of estimators can be dictated by variance considerations.

**Biases Using Different Weights**

The relative bias in sample size were computed using different estimators for samples of 5, 10, 30, and 60 telephone numbers per cluster. The relative bias in sample size is negligible for the cluster sample sizes of 30 and 60 numbers, except when the unadjusted weights are used. Any of the adjusted estimators could be used for cluster samples sizes of this size without incurring biases in the estimates. When 10 numbers are selected per cluster, all of the weights except the unadjusted one again perform reasonably well. The bias performance of $(n_i + .5)^{-1}$ is especially encouraging.

For the smallest cluster size studied, 5 numbers per cluster, the potential for bias is somewhat greater. The natural weight, $n_i^{-1}$, has a somewhat lower bias in sample size than the weight $(n_i + .5)^{-1}$, but the difference is not very large. The relative bias in sample size for both of these weights is always less than one percent. For residential densities between about 45 percent and 80 percent the bias is positive and elsewhere it is negative. This pattern might be problematic only for the few characteristics that are very highly correlated with residential density.

**Biases with Trimmed Weights**

The introduction of trimming can produce significant biases, depending on the relationship between the characteristics being estimated and the weights which are being trimmed. In some applications, the bias associated with trimming may limit the amount of trimming that can be applied and, hence, its usefulness for variance reduction.

The relative bias in sample size was also computed for cluster samples of 10, 30 and 60 numbers and the weights trimmed at about 3 times the mean weight. The trimming for samples of 10 numbers per cluster was done at a factor of 2 rather than 3 as described previously.

The difference between the relative bias in sample size for the trimmed and untrimmed weights is largely inconsequential for all cluster sample sizes and most values of $N_i/100$. The only noticeable difference occurred when the residential density is under about 10 to 15 percent. There is a slightly greater potential for bias in these regions. However, the relative bias in sample size for the trimmed weights is still much less than one percent at all residential density values.

## 5. CONCLUSION

The standard Mitofsky-Waksberg method is an effective method of producing a self-weighting, RDD sample of fixed size. However, the sequential monitoring of the number of cases per cluster is an awkward operational feature of this method. One of the consequences of the sequential monitoring of caseloads by cluster is that it is difficult to complete data collection in a tight time frame. The data collection period has to be flexible enough to allow for obtaining the appropriate number of cases in each cluster. The more extensive data collection period and the monitoring of caseloads also result in increasing costs. Another problem with the sequential operations is that the requirement for frequent monitoring of the caseloads can lead to frustration arising from complications of combining sample selection and data collection operations.

The modified Mitofsky-Waksberg approach eliminates the sequential nature of the design and, with it, the need to monitor the work by cluster. A fixed number of telephone numbers are assigned to each sampled cluster in the modified method. Therefore, the costs associated with monitoring caseloads and a longer data collection period are not incurred. However, the modified Mitofsky-Waksberg method does introduce new components of bias and variance into the estimates. These statistical concerns should be addressed before the modified approach is used.

Specific recommendations on when the standard or modified Mitofsky-Waksberg method should be used are difficult to formulate since they depend upon circumstances which vary from survey to survey. Guidelines for choosing between the methods are suggested below.

A simple rule is that for surveys which require either very tight controls on sample size or a nearly self-weighting sample, then the standard Mitofsky-Waksberg approach is advisable. Even though the sample size in the modified method can be estimated relatively precisely, some variation, especially because of uncertainty of the nonresponse rates, can be expected. A self-weighting sample, which is not achieved when the modified Mitofsky-Waksberg method is used, also has some advantages in simplifying standard statistical analysis.

Since the costs for standard and modified methods are different, it would be very useful to have cost-variance models to help evaluate the two methods. Unfortunately, the differences in costs of the standard and modified methods are not easy to quantify. In fact, the lack of reasonable cost models is a major and pervasive problem that limits the ability to establish optimal survey design.

Because of lack of reasonable cost-variance models, we suggest some conditions in which one approach might be favored over another. One of the conditions that favors the modified approach is a relatively brief interview length. As the interview becomes longer, the cost savings associated with the modified method is likely to become smaller relative to the increases in variances of the estimates.

The length of the interview is particularly important for surveys which screen households to find those with particular characteristics. For example, some RDD surveys screen households and only interview if a member is in a particular target group. In these situations, the screening interview is often very brief. The modified Mitofsky-Waksberg approach may be very beneficial. Surveys in which households are screened also tend to have large cluster sample sizes, and this improves the performance of the modified procedure. When 10 or more numbers are selected per cluster (equivalent to about 6 households per cluster), the biases in the estimates under the modified Mitofsky-Waksberg approach are virtually inconsequential and the increases in variance with trimming are only about 10 percent. Samples of 10 or more numbers per cluster are frequently acceptable for screening purpose although such large cluster sizes are typically inefficient for the interview sample, even when the intraclass correlation is small.

Based on these factors, a general guideline is that the modified Mitofsky-Waksberg method can be recommended when households within the clusters must be screened. More specifically, the modified method with trimmed weights should be considered if the following conditions exist: (1) Ten or more numbers are sampled per cluster; and (2) the total cost for the modified method is at least 10 percent less than the standard method, or the data collection period is relatively short. If both of these conditions are not met, then the choice between methods must be made on evaluations of other survey requirements.

When the cluster sample size is less than 10, the bias and variance arising from the use of the modified Mitofsky-Waksberg method are more serious concerns. Any characteristics correlated with the proportion of residential numbers in a cluster could be affected with a cluster sample size this small. Also, the variance of the estimates with the modified method will be

20 to 30 percent larger than with the standard method since trimming is not very effective with small sample size. Therefore, in most surveys with sample sizes of less than 10 numbers per cluster, the problems of implementing the standard method should be quite serious before a decision is made to abandon it and use the modified method.

## ACKNOWLEDGEMENT

The author would like to thank the referees. Their comments were very helpful in improving the presentation of the paper.

## REFERENCES

CUMMINGS, K.M. (1979). Random digit dialing: A sampling technique for telephone surveys. *Public Opinion Quarterly*, 233-244.

DREW, J.D., DICK, P., and SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-127.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., and WAKSBERG, J. (editors) (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, 2. New York: John Wiley and Sons, 138-139.

KISH, L.(1965). *Survey Sampling*. New York: John Wiley and Sons, 429-430.

LEPKOWSKI, J.M., and GROVES, R.M. (1986). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 73-98.

PIEKARSKI, L. (1990). Working block density declines. *The Frame*, a publication of Survey Sampling Inc.

POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.

SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.

TUCKER, C. (1989). Characteristics of commercial residential telephone lists and dual frame designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-137.

WAKSBERG, J. (1984). Efficiency of alternative methods of establishing cluster sizes in RDD sampling. Unpublished Westat Inc. memorandum.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subjects of the population. *Proceedings of the Social Statistics Section, American Statistical Association*.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 43–56
Statistics Canada



# The Blaise System for Integrated Survey Processing

JELKE G. BETHLEHEM and WOUTER J. KELLER[1]

ABSTRACT

Application of recent developments in computer technology allow national statistical offices to produce high quality statistics in an efficient way. At the Netherlands Central Bureau of Statistics (CBS) an increasing use is made of microcomputers in all steps of the statistical production process. This paper discusses the role of software and hardware in data collection, data editing, tabulation, and analysis. To avoid the negative effects of uncontrolled de-centralized data processing, the importance of integration is stressed. This makes the statistical production process easier to manage, and moreover it increases its efficiency. The Blaise System, developed by the CBS, is discussed as a data processing tool that encourages integration. Using a description of the survey questionnaire, this system is able to automatically generate various computer programs for data collection (CAPI or CATI), or data entry and data editing (CADI). The system can also create interfaces to other packages. Particularly, the link between Blaise and the internally developed packages Bascula (for weighting) and Abacus (for tabulation) is described. In this way the Blaise System controls and co-ordinates, and therefore integrates, a large part of the survey process.

KEY WORDS: Integration; Survey processing; CAPI; CATI; Microcomputers; Decentralization; Standardization.

## 1. INTRODUCTION

The Netherlands Central Bureau of Statistics (CBS) makes an increasing use of microcomputers in survey data processing. The introduction of microcomputers has a considerable impact on the way the work of the statistical office is carried out. Subject matter statisticians become increasingly aware of the potential of the new technology, and consequently use it more and more in their daily work.

This paper discusses the role of the new automation technology in data collection, data editing, tabulation, and analysis. We will stress the importance of standardization and integration. These working policies have three advantages: they enable us to avoid the negative effects of uncontrolled de-centralized data processing, they make the statistical production process easier to manage, and they increase efficiency.

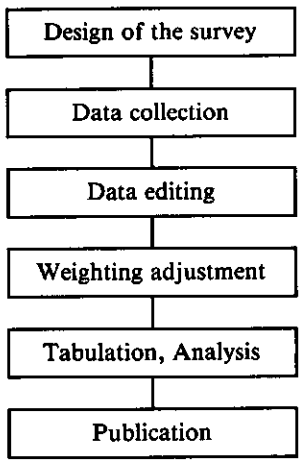
The Blaise System, developed by the CBS, is discussed as the backbone of an integrated survey processing system. On the one hand, the power of this system lies in the consistency it enforces in the various steps of data collection and data processing. On the other hand, it also promotes standardization between different departments. Since all departments use the same software for processing their surveys, everybody speaks the same "language", and so exchange of information between departments is easier and less error prone.

## 2. THE STATISTICAL PRODUCTION PROCESS

National statistical offices collect data on persons, households and establishments and transform this information into useful statistics. Production of statistical information is often a complex, costly and time-consuming process. This section describes the various steps the

[1] Jelke G. Bethlehem and Wouter J. Keller, Netherlands Central Bureau of Statistics, Automation Department, P.O. Box 959, 2270 AZ Voorburg, The Netherlands.

**Table 1**
The statistical production process




statistical office has to go through, the problems that it may encounter, and the decisions it has to make. An overview of the process is given in table 1.

The first step is, of course, the design of the survey, in which the statistician specifies the population to be investigated, the data to be collected, and the characteristics to be estimated. Since statistical offices collect most data by means of (sample) surveys, a questionnaire has to be defined, containing the questions to be asked of the respondents. This questionnaire is the first practical description of the data to be collected. Furthermore, in the case of a sample survey, the statistician also has to specify a sampling design, and he must see to it that the sample is selected properly.

The second step in the process is *data collection*. Traditionally, in many surveys the questionnaires are completed in face-to-face interviews: interviewers visit respondents, ask questions, and record the answers on (paper) forms. The quality of the collected data tends to be good. However, since it typically requires a large number of interviewers, who may all have to do much travelling, it can be expensive and time-consuming. Therefore telephone interviewing is sometimes used as an alternative. The interviewers call the respondents from the statistical office, and thus no more travelling is necessary. However, telephone interviewing is not always feasible: only connected people can be contacted, and the questionnaire should not be too long nor too complicated. A mail survey is cheaper still: no interviewers at all are needed. Questionnaires are mailed to potential respondents with the request to return the completed forms. Although reminders can be sent, the persuasive power of the interviewer is lacking, and therefore response tends to be lower in this type of survey, and so does the quality of collected data.

If the data are collected by means of paper forms, completed questionnaires have to undergo extensive treatment. In order to produce high quality statistics, it is vital to remove any errors. This step is called *data editing*. Three types of errors can be distinguished: A *range error* occurs if a given answer is outside the valid set of answers, *e.g.* an age of 348 years. A *consistency error* indicates an inconsistency in the answers to a set of questions. An age of 8 years may be valid, and a marital status "married" is not uncommon, but if both answers are given by the same person, at least in the Netherlands, there is something definitely wrong. The third

type of error is the *routing error*. This type of error occurs if the interviewer or the respondent fails to follow the specified branch or skip instructions, *i.e.* the route through the questionnaire is incorrect: irrelevant questions are answered, or relevant questions are left unanswered.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterwards, at the office. In many cases, particularly for household surveys, respondents cannot be contacted again, so other ways have to be found to do something about the problem. Sometimes it is possible to determine a reasonable approximation of a correct value by means of an imputation technique, but in other cases an incorrect value is replaced by the special code indicating the value is "unknown".

In addition to data editing, another activity is sometimes carried out during this stage of the production process: *coding of open answers*. A typical example is the question about the occupation of the respondent. Questions are easiest to process if a respondent selects one possibility from a list of pre-coded answers. However, for a question like occupation this set of pre-coded answers would be very long, and thus it would be very hard for the respondent to select the proper answer. This problem is avoided by letting the respondent formulate his own answer, and then literally copying the answer on the form. To enable analysis of this type of information, answers must be classified afterwards. This is a time-consuming and costtly job, which must be carried out by experienced subject-matter specialists.

After data editing, the result is a "clean" file, *i.e.* a file without errors. However, this file is not yet ready for tabulation and analysis. In the first place, the sample is sometimes selected with unequal probabilities, *e.g.* establishments are selected with probabilities proportional to their size. The reason is that a clever choice of selection probabilities makes it possible to produce more accurate estimates of population parameters, but only in combination with an estimation procedure which corrects for this inequality. In the second place, representativity may be affected by nonresponse, *i.e.* for some elements in the sample the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, the results will be biased.

In order to correct for unequal selection probabilities and nonresponse, a *weighting adjustment* procedure is often carried out. Every record is assigned some weight. These weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status and area reflects the known distribution of these characteristics in the population.

In the case of item non-response, *i.e.* answers are missing on some questions, and not all questions, an *imputation procedure* can also be carried out. Using some kind of model, an estimate for a missing value is computed and substituted in the record.

Finally, we have a clean file which is ready for *analysis*. The first step in the analysis phase will nearly always be tabulation of the basic characteristics. Constructing a table is not as simple as it may look at first sight. The composition of rows and columns (often built from a number of variables), the quantities displayed in cells (counts, means, percentages), the way in which percentages are computed, treatment of multiple-response variables, the position of totals and subtotals, and many other things, can make life very difficult.

Many statistical offices also carry out analysis on their data in order to reveal the underlying structures, and thus to gain insight in the data. Information obtained in this way may improve a later survey, and thus improve quality or reduce costs.

The results of the analysis will be *published* in some kind of report. Usually it will contain tables and graphs. It is important to present the statistical information in such a way that the proper "message" is conveyed. Graphs, tables and text should be simple and clear. Particular attention should be paid to graphs, because a visually ambiguous or confusing graph will quite easily lead to wrong interpretation.

## 3. THE NEED FOR INTEGRATION

The computer has always been important in statistical information processing. In the beginning to computer was only used for activities like sorting, counting and tabulation. In the sixties and seventies, with the emergence of mainframes and statistical packages, it become possible to carry out extensive analysis. The computer was also increasingly used for data editing, weighting adjustment and imputation. The use of computers for data collection is more recent. This first occurred for telephone interviewing (CATI). In the last decade, the advent of the small laptop computers has made it possible for interviewers to take the computer with them to the homes of the respondents. This way of computer assisted face-to-face interviewing is denoted by CAPI.

It will be clear that the computer is used for more and more activities. Hardware and software are available for nearly every step in the production process. Also an increasing number of people are making use of the automation tools. At first, only the computer specialists had access to "their" machines, but now statisticians and subject matter experts have become computer-oriented, and therefore make increasing demands for suitable software and hardware to do their jobs. Simple and straightforward electronic data processing can, and is, now carried out by the subject matter departments themselves, leaving design and maintenance of complex information systems to be carried out by the computer specialists of the automation department. As a consequence of these developments the work of the statisticians and subject-matter experts have changed. They used to be specialists in their own (narrow) field, but now they have acquired more general knowledge and experience in a much broader field containing subject matter aspects, statistical methodology and computer processing. So the specialists have vanished, and a new group with general knowledge of all aspects of survey processing has emerged.

Automation of the statistical production process is nice, but one should be aware of the dangers. Although application of computers promises increased efficiency and quality, an uncontrolled and unco-ordinated use of the new technology may easily lead to chaos, and hence to less productivity. Factors affecting the efficiency of the statistical production process are:

- **Different departments are involved.**
  Many people deal with the information: respondents fill in forms, subject-matter specialists check forms and correct errors, data typists enter the data in the computer, and programmers construct editing programs. Transfer of material from one person/department to another can be a source of error, misunderstanding and delay.

- **Different computer systems are involved.**
  Various data processing activities may be carried out on different computer systems. Transfer of files causes delay, and incorrect specification and documentation may produce errors.

- **Repeated specification of the data.**
  In almost every step of the process, the structure of the data must be specified. The particular system or department has to know about the data: What is the meaning of the variables? Which values are permitted? Are there any constraints on the routing? Which relationships between variables have to be checked? Although essentially the same, the form of specification may be completely different for every step. Every system uses its own "language". The first specification is the questionnaire itself. Another specification may be needed for data-entry, and yet another for the checking program, for tabulation and analysis, *etc.* It is clear that this is not the most efficient way to deal with the information.

The CBS solution to these problems is *integration*. In this context, integration has three different aspects: integration of work, hardware, and software. Let us first have a look at integration of the work.

Traditional data processing consist of what we call *macro cycles*. All survey data as a whole goes through cycles: from one department to another, and from one computer system to another. First the paper forms are cleaned manually by the subject matter department, then data on the forms are entered by the data entry department, next the files are transfered to a mainframe computer system. A program checks the data for consistency, detected errors are printed on lists that are send back to the subject matter department for corrections. This process of data entering and data editing has to be repeated a number of times before the data can considered to be "clean".

The idea behind integration of work is that the macro cycles should be replaced by *micro cycles*. Not the whole data file, but instead only one record at a time should cycle around. Micro cycles means that cycling should take place within one computer system, and that this should be controlled by one department. Going from macro cycles to micro cycles comes down to concentrating all data processing activities in one department, and that is the subject-matter department. Since the subject-matter statisticians are the ones with most knowledge about the area covered by a survey, they are best equipt to deal with the data, to solve problems, and to produce high quality statistics. Of course, they need proper instruments to do their job, *i.e.* powerful and user-friendly software and hardware.

The idea that automation of data processing activities should be carried out exclusively by computer specialists is out of date. More and more the subject-matter statisticians become aware of the possibilities and usefulness of the computer for their own work. So the time has come for subject-matter departments to take simple and straightforward survey data processing into their own hands. Of course, the automation department is responsible for providing the proper automation infrastructure. And this department stays in charge of design and maintenance of complex information systems.

The second aspect of integration is integration of hardware. The idea is to concentrate work on one type of computer as much as possible. Taking into account that a large number of inexperienced statisticians will have to use the computer, the obvious choice is the microcomputer. Microcomputers offer user-friendliness at a relative low price, and moreover, there is an abundance of useful software.

Being aware of the fact that statistical offices process huge quantities of data, one may wonder whether microcomputers have the capacity to carry out all work, and indeed can take over from the large workhorses, the mainframes. To be able to answer this question, it is useful to distinguish between two kind of data processing activities. In the first place there are record oriented activities. These are activities for which only one record at a time is needed. Examples of record oriented activities are data entry and data editing. Record oriented activities are generally very well suited for interactive processing. In the second place, there are file oriented activities. These activities can only be carried out properly if the whole file is available. Examples are the computation of weights and tabulation. Because of their size, file oriented activities are often processed in a batch-wise fashion.

The viewpoint of a few years ago was that record oriented activies could be carried out on microcomputers but file oriented activites had to take place on mainframes. With the increasing power of microcomputers, attention is shifting in the direction of the microcomputer. At this moment, the policy of the CBS is that all record oriented activities have to be carried out on microcomputers and file oriented activities can in many cases (say, with data files of less than 50 megabytes) also be carried on microcomputers. However, for data storage and large batch jobs we still need mainframes.

The users of the computer environment should be confronted as little as possible with the mainframe. Therefore, the CBS is moving in the direction of front end/back end systems. The front end consists of microcomputers, and that is what the statisticians use to specify their problems. The back end is a mainframe or mini-computer, and is used bulk work, maybe even without the user knowing it. Particularly for database applications the client/server approach looks very promising. In this approach, the real database activities take place on a dedicated minicomputer, whereas the activities are specified, initiated and controlled by the micro-computers at the desks of the users.

## 4. STANDARDIZATION

The CBS makes an increasing use of microcomputers (running under MS-DOS) in many steps of the statistical production process. On the one hand, this opens new ways towards efficient information processing, but on the other hand, it creates new problems that have to be dealt with. If every department is free to select and purchase its own type of computer and software, the automation infrastructure may easily get out of control, and turn into chaos. Departments will not talk the same "language" anymore, because they use different data formats and different software. It is clear that this calls for a strong policy on standardization. The CBS has adopted such a policy, and in practice it means that there are only one or two software packages available for a particular task.

Another advantage of standardization is that it limits the amount of training that has to be provided for the users. In order to cope with the problem of training a large number of new microcomputer users, the CBS runs an average of 50 one-day courses per month (occupying three fully equipped lecture rooms every working day).

Attention should also be paid to the way in which the microcomputers are used in the organization. Distribution of a lot of stand-alone microcomputers may seem a simple solution, but there are also problems that have to be solved. In the first place, it is very easy to copy (confidential) data files on local hard disks, so we have a data security problem. Furthermore, activities like making back-ups and archiving are often neglected by the users in the subject-matter departments. Also communication between departments (*e.g.* sharing data files) is only possible by exchanging floppy disks. Finally, distribution of new releases of software packages, including their documentation, is often cumbersome in large organizations with a lot of stand-alone microcomputers.

To avoid the above mentioned problems, the CBS has installed approximately 60 local area networks (LANs). Every department has its own LAN. Ten to sixty microcomputers are connected to a high-end 386-based fileserver with a storage capacity of up to 600 Megabytes. In this environment there are in total nearly 2,300 microcomputers, half of them based on the Intel 386SX micro-processor. Security is guaranteed by means of password protection in a login-procedure, by encryption, and by using floppy-less workstations (of the 2,300 microcomputers only 60 have a floppy or hard disk drive). Archiving and backing-up the LANs is carried out in a centralized way by the automation department. A full backup of more than 15 Gigabytes is carried out every night. It is clear that version control and updating software can more easily be realized in such an environment. Distribution and installation of new software releases on a LAN is easy, since, with one command one can upload the new version to all fileservers. All software licenses are based on concurrent usage, which is checked by home-made software.

The role of microcomputers in the statistical production process is growing, but for the time being, there are still applications (like the use of large databases) that need mainframe or minicomputer systems. In this environment, the CBS has adopted Oracle as the standard

database system. Development of a database application is preferably carried out on a microcomputer, whereas actually running it takes place on a mini computer. Recently, the CBS realized a client/server architecture based on a distributed database system. Microcomputers in the network serve as front ends and the minicomputers as back ends.

So, as the use of the data processing instruments is brought closer to the subject-matter specialists at the departments (de-centralization), standardization and coordination of the work environment of the subject-matter users demands strong centralization. More details about the automation infrastructure can be found in Keller, Metz and Bethlehem (1990).

## 5. INTEGRATION OF THE SURVEY PROCESS

The previous section discussed the need for integration in the survey process. Particular attention was paid to concentrating the work in subject-matter departments, and standardization of the hardware and software instruments. But standardization of software is not enough. The efficiency of the production process can be increased even more by integrating the required standard software into one system. This section describes how such an integrated system for survey processing is implemented at the CBS.

An integrated system for survey processing should be based on a powerful language for the specification of questionnaires. This specification is the "knowledge base", containing all knowledge about the questionnaire and the data. The system should be able to exploit this knowledge, *i.e.* it must be able to automatically generate all required data processing applications. On the one hand it means the automatic generation of software for data collection, data entry and data editing, and on the other hand the automatic generation interfaces for other data processing software, *e.g.* for tabulation and analysis. In this way repeated data specification is no longer necessary, and consistency is enforced in all data processing steps.

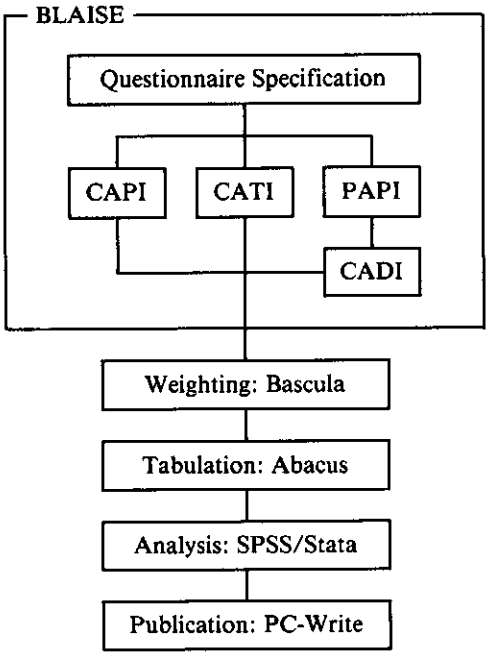
The backbone of the integrated survey processing system developed by the CBS is the Blaise System. In the design phase of the survey, the questionnaire is specified in the Blaise language. And it is this specification that is used throughout the whole survey process to extract the information necessary to carry out the various data processing steps. Table 2 summarizes the integrated system for survey processing.

The Blaise System can produce three kinds of programs: CADI, CAPI and CATI programs. CADI stands for Computer Assisted Data Input. It integrates data entry and data editing by offering an interactive environment for processing paper questionnaire forms. The Blaise System can also produce the software required to carry out CAPI or CATI interviewing. The Blaise System is discussed in more detail in section 6.

Whatever form of data collection is used, the result will be a "clean" data file, *i.e.* a file in which no more errors can be detected. The next step in the process will often be the computation of adjustment weights. The program Bascula will take care of this. It is able to read the Blaise data files directly, and extract the information about the variables, *i.e.* the meta-information, from the Blaise specification. Running Bascula will cause an extra variable to be added to the data file containing the adjustment weight for each case. More about Bascula can be found in section 7.

Now the file is ready for tabulation, and for that, the integrated system offers the program Abacus. This program is also able to read and understand the data files created in the previous step of the process. See section 8 for details. Tabulation may be followed by a more extensive analysis of the data. For that purpose the Blaise System can generate interfaces for the statistical packages SPSS and Stata. More about this in section 9.

**Table 2**
Integrated survey processing




Finally, a publication will be prepared using the standard wordprocessor PC-Write. Since this wordprocessor runs on the same computer system as the other software, it is easy to import generated tables and results of statistical analysis into the text.

## 6. THE BLAISE SYSTEM

The Blaise System was developed by the CBS, and it derives its name from the famous French theologian and mathematician Blaise Pascal (1623-1662). The basis of the Blaise System is the Blaise language, which is used to create a formal specification of the structure and contents of the questionnaire. The Blaise language has its roots, in large part, in the programming language Pascal.

The Blaise System runs on microcomputers (or networks of microcomputers) under MS-DOS. It is the backbone of the integrated survey processing system, and as it is intended to be used by the people of the subject-matter departments, one need not be a computer expert to use the Blaise System. The design goal of the system was to provide subject-matter experts with a powerful but user-friendly tool that enables them to input their knowledge about a survey into the system, and to take care of all subsequent data processing steps.

In the Blaise philosophy, the first step in carrying out a survey is to design a questionnaire in the Blaise language. Such a specification of the questionnaire contains more information than a traditional paper questionnaire. It not only describes questions, possible answers, and conditions on the route through the questionnaire, but also relationships between answers that have to be checked.

The Blaise System can produce programs for CADI, CAPI or CATI. A CADI program is an intelligent and interactive system for data entry and data editing of data collected by means of paper forms. The subject-matter specialist works through a number of forms with a microcomputer, processing them one-by-one. He enters answers to questions at the proper places and, after completion of the form, he activates the check option to test routing and consistency. Detected errors are reported and explained on the screen. Errors can be corrected by consulting the form or calling the supplier of the information. After elimination of all errors, a clean record is written to a file.

The CADI program can also be used for a different way of data processing not mentioned thus far. Sometimes statistical offices do not carry out their own data collection, but they have to create statistics using data files that were generated elsewhere, outside the statistical office. In these cases the data still has to be checked. The Blaise System has a facility to import this kind of data files. With a CADI program, an integral check can be carried out on all records in a batch-wise version. Thus the records are assigned either the status "clean" or "dirty". And the dirty records can be corrected interactively, again with the CADI program.

A CAPI/CATI program can be used for computer assisted interviewing. The paper questionnaire form is replaced by a computer program containing the questions to be asked. This computer program is in control of the interview. It determines the proper next question to be asked, and checks the answers as soon as they have been entered. In the case of CAPI, the interviewing program is loaded into a laptop computer, and the interviewer takes this computer to the homes of the respondents. In the case of CATI, the program is in a desktop computer. The interviewer calls the respondents from a central unit, and carries out the interview by telephone.

The generation of a Blaise CADI/CAPI/CATI proceeds in a number of steps. First, a text editor is used to enter the Blaise specification of the questionnaire, after which it is checked for syntax errors. Detected errors must be corrected, and to do that the system returns to the text editor and places the cursor on the approximate location of the error. After correction, the specification is checked again. If no errors are detected, the specification is transformed into Pascal source code, which in turn is compiled into an executable program.

The Blaise language must serve two somewhat conflicting purposes. On the one hand it must be powerful enough to be able to deal with all kinds of large and complex surveys, and on the other, Blaise questionnaire specifications must be readable enough, for use by subject matter specialists. In fact, a Blaise questionnaire must be self-documenting, i.e. it is the basic description of the survey which can be used by all people involved. Table 3 gives an example of a simple questionnaire in Blaise.

The first part of the questionnaire specification is the QUEST section, containing the definition of all questions that can be asked. A question consists of an identifying name (for internal use in the questionnaire), the text of the question as presented to the respondents, and a specification of valid answers. The next part of this sample Blaise questionnaire is the ROUTE section. It describes under which conditions, and in which order the questions have to be asked. Consistency checks are specified in the CHECK section.

The description above does not exhaust the power of the Blaise language. An overview of the Blaise language can be found in Bethlehem *et al.* (1989b), and more details in Bethlehem *et al.* (1989c).

The Blaise System contains a module for *interactive coding*, thereby providing the possibility of integrating coding either in the data collection phase or in the data entry and data editing phase. The module contains two different tools. The first tool implements a hierarchical approach to coding. Coding of an answer starts by entering the first digit of the code by selecting

**Table 3**

A simple Blaise questionnaire

```
QUESTIONNAIRE Work "The Work Survey";

QUEST
  SeqNum    "Sequence number of the interview?": 1..1000 (KEY);
  Age       "What is your age?": 0..99;
  Sex       "Are you male or female?": (Male, Female);
  MarStat   "What is your marital status?":
              (Married  "Married",
              NotMar    "Not married")
  Job       "Do you have a job?": (Yes, No);
  JobDes    "What kind of job do you have?": STRING[20];
  Income    "What is your yearly income?":
              (Less20   "Less than 20,000",
              Upto40    "Between 20,000 and 40,000",
              More40    "More than 40,000");
  Travel    "How do you usually travel to your work?":
              SET [3] OF
              (Walking  "Walking",
              Bicycle   "By bicycle",
              Car       "By car or motorcycle",
              PubTrans  "By bus, tram, train or metro",
              Other     "Other means of transport");
  OthTrans  "What other means of transport?": STRING[20];

ROUTE
  SeqNum; Age; Sex; MarStat; Job;
  IF Job = Yes THEN
    JobDes; Income; Travel;
    IF Other in Travel THEN OthTrans ENDIF
  ENDIF

CHECK
  IF Age < 15 "Respondent is younger than 15" THEN
    MarStat = NotMar "he/she is too young to be married!"
  ENDIF

ENDQUESTIONNAIRE.
```

the proper category from a menu. After the user enters a digit, the program presents a subsequent menu containing a refinement of the previously selected category. So the description becomes more and more detailed until the final digit is reached. The second tool consists of a dictionary approach to coding. It tries to locate an entered description in an alphabetically ordered list. If the description is not found, the list is displayed, starting at a point as close as possible to the entered description. The list can be made so that almost any description, including permutations, is present. The advantage of this method is that it is simple, fast and controllable. Both coding tools can be used simultaneously.

## 7. BASCULA

The clean file with sample survey data produced by the Blaise System is usually not ready yet for making inference about the population from which the sample has been drawn. The problem is that the data do not constitute a representative sample, and so some adjustment procedure has to be carried out.

In order to account for unequal selection probabilities and nonresponse, one often has to compute adjustment weights. Post-stratification is a well-known technique. Every record is assigned some weight, and these weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population. Two major problems can make application of post-stratification difficult: empty strata and lack of adequate population information. Research has been carried out at the CBS in order to improve weighting techniques. The result was a new general method for weighting, in which weights are obtained from a linear model which relates the target variables of a survey to auxiliary variables. Post-stratification is a special case of this method. Because of the generality of the method, different weighting schemes can be applied that take advantage of the available population information as much as possible, and at the same time avoid the above mentioned problems. See Bethlehem and Keller (1987) for more details.

Bascula is a general weighting program, running on microcomputers under MS-DOS. It combines several weighting techniques. In the first place, traditional post-stratification can be carried out. And if the number of empty strata is small, one can instruct the program to collapse (*i.e.* combine) these strata with neighbouring strata. In the case of many empty strata, or lack of sufficient population information, Bascula can carry out the linear weighting technique described above or apply iterative proportional fitting (also called multiplicative weighting, or raking ratio estimation). The resulting weights can either be added to the data file, or be stored in a separate file.

Bascula is able to read the Blaise data file directly, and also extracts the required information about the variables from the Blaise specification. The information about the population has to be provided by the user. The program is menu-driven, making it user-friendly. It will carry out a complete post-stratification if possible. If not, the user has to decide either to carry out linear or multiplicative weighting.

Presently, Bascula can only be used on a microcomputer. In the future, a back end will be developed that will run on our mainframe environment. Bascula was particularly developed for use in social and demographic surveys, where post-stratification is combined with relatively simple estimation procedures. For use in economic surveys, different software will be developed. This software will concentrate on stratified sampling designs in combination with more complex estimators (ratio and regression estimation).

## 8. ABACUS

Tabulation is one of the basic activities in the statistical production process, and it was one of the first to be automated. Many tabulation packages have already been developed in the world, but many are not very user-friendly. This is partly caused by the fact that proper generation of a complex table needs a lot of parameters to be specified: the variables to be used in the various dimensions (rows, columns, layers), whether to concatenate variables (display all values of a variable, followed by all values of another variable) or to nest variables (display for every value of one variable all possible values of another variable) within a dimension, the

displayed cell quantity (counts, percentages, totals, averages), whether to display totals and subtotals, and many layout features. To be able to cope with all these parameters, traditional packages have control languages to specify tables, and these languages are often not very easy to learn and to use.

Abacus is a tabulation package, running on microcomputers under MS-DOS. While Abacus may be seen as yet another tabulation package, it was developed with very specific design goals. In the first place, no control language is used to specify a table. The program is menu-driven instead. The user designs his table in an interactive, simple, and intuitive way, without having to know about any control language. In the second place, Abacus can directly read the data file created by the Blaise System, as well as Ascii files. The meta-information, *i.e.* the information about the variables in the file, can be generated by the Blaise System, or can (in the case of separate Ascii files) be entered interactively by the user. Thirdly, the program can produce camera-ready tables.

Another striking property of Abacus is its speed. A table produced by SPSS-Tables in 3 minutes was generated by Abacus in about 6 seconds (all timings based on the same 386SX based microcomputer). The reason for this is that the Abacus program is rather small, so it can use a large part of the memory as a working area which allows for a table of up to 90,000 cells.

Tables produced by Abacus can have up to three dimensions (layers, rows and columns). Every dimension can hold up to 10 variables, which may be nested or concatenated. In the example in table 4, the column variables "Employment" and "Sex" are concatenated while in the row dimension the variables "Region" and "Town" are nested. In this example no variable has been placed in the layer dimension.

This table contains simple counts, but Abacus can also calculate totals of quantitative variables, make percentages tables, and averages tables. It is also possible to have more than one (up to 10) items in the cells of the table. In that case, the user has to decide to put each item in a separate row, column or layer. If the data has been collected by means of a sample survey, Abacus can accommodate weighted data, using the weights that are, for example, computed by Bascula. The only thing the user has to do is to specify the variable containing the weights.

**Table 4**

An example of a two-dimensional table

| Number of Records | The population of Samplonia | | | | |
|---|---|---|---|---|---|
| | Total | Employment | | Sex | |
| | | Job | No Job | Male | Female |
| Total | 1,000 | 341 | 659 | 511 | 489 |
| Agria | 293 | 121 | 172 | 145 | 148 |
| Wheaton | 144 | 60 | 84 | 70 | 74 |
| Greenham | 94 | 38 | 56 | 44 | 50 |
| Newbay | 55 | 23 | 32 | 31 | 24 |
| Induston | 707 | 220 | 487 | 366 | 341 |
| Oakdale | 61 | 26 | 35 | 36 | 25 |
| Crowdon | 244 | 73 | 171 | 128 | 116 |
| Smokeley | 147 | 49 | 98 | 80 | 67 |
| Mudwater | 255 | 72 | 183 | 122 | 133 |

Source: Samplonian Statistical Office.

Much attention has been paid to the layout of the table, because the tables produced should be camera-ready. Therefore there are many options in Abacus to control the layout. It is possible to specify up to 10 lines of text for the header and for the footer of the table, and one can select both horizontal and vertical rules (as in the example), only horizontal rules or no rules at all. The layout of the text in column headers and the width of the columns can also be influenced.

A rounding procedure can be carried out to protect confidential data in the table. Cell totals, but also marginal totals are rounded to a multiple of some specified constant, *e.g.* 5. Abacus can provide both normal rounding and random rounding. If the user is not satisfied with the resulting table, he can import the output of Abacus into the spreadsheet program Lotus 123, and carry out further processing there. A final feature to be mentioned here is the possibility of creating new variables by recoding existing variables (*e.g.* from age to age classes). More details about Abacus can be found in Bethlehem *et al.* (1989a).

## 9. ANALYSIS

The CBS has not developed any software for statistical data analysis, the main reason being that there are already enough good statistical packages available. The CBS itself uses the packages SPSS (both on mainframe and micro) and Stata (on micro). To make these packages part of the integrated system for survey processing, tools have to be available to export the data from Blaise to them. This is realized in two steps. First, the data file is converted from the Blaise format to Ascii format, and second, the information about the variables, as available in the Blaise questionnaire, is translated in such a way that it can be understood by the particular package. Thus, a setup file is created. By running this setup from within the statistical package, a system file is created. And by loading the system file, the user can start straight away with his analysis, without having to bother about specifying the variables, labels, *etc.*

The procedure above only works for SPSS and Stata, and not for other packages. Of course, this approach could be implemented for every known statistical package, but that would require a large programming effort. Instead, a different road was taken. The Blaise System has a special setup generator utility. The user "paints" the structure of the setup file in a word processor, and by running the setup generator with this general setup description and the Blaise questionnaire as input, a real setup file is created. So, with the setup generator the user can generate setup files for his own favourite package.

## 10. CONCLUSION

The advent of the microcomputer has had a considerable impact on the work of the national statistical office. The subject matter statistician is making use of it more and more, and for his work, he needs an integrated survey processing system like the one based on the Blaise System. The power of this system lies on the one hand in the consistency it enforces in the various steps of data collection and data processing. This makes the whole process easier to manage and to control. On the other hand it also encourages standardization between different departments. Since all departments use the same software for their surveys, exchange of information between departments is easier and less error prone.

The integrated approach to survey processing was developed with in mind a highly centralized organization, like that of the Netherlands Central Bureau of Statistics. In such an organization, this approach can lead to a substantial increase in efficiency. However, not all statistical offices have a centralized structure. Particularly in larger countries, data processing is often

decentralized. Regional offices take care of data processing in their own regions, and the resulting data files are sent to the central office. In the central office, the regional files are combined into one national file. The integrated approach can also be applied successfully in such an environment: the central office develops the Blaise questionnaire, and copies of the generated data entry program are sent to each regional office. This ensures consistency of data collection and data editing at the regional level. The regional data files will all have the same Blaise format, so combining them into one national file will be a simple job using the tools of the Blaise System. Since all regional files will be "clean", no further editing will be necessary at the national level. The only job for the national office will be to tabulate and analyse, and to publish the results. Furthermore, if either the regional offices or the central office can make regional publications.

The Blaise System has been tested and used since the middle of 1986 for a substantial number of surveys. The system is developed in close cooperation with the users. Every new version contains enhanced features. Abacus has been in use for over a year, and is very popular among its users. The Bascula program is still in development. The first prototype has just been released.

We did not mention the possibility of exporting data and meta-information from the Blaise System to the Paradox database package. In the future, a link will also be established between Blaise and the Oracle database system. In this way, a client/server architecture can be realized for Blaise users.

At the end of the statistical production chain, there are some aspects of publication that still have to be dealt with. In the first place, software will be developed to asses the risk of disclosure of confidential (private) information in statistical information to be published. Tools will also be offered to protect tables or data files against these risks. Finally, statistical offices engage more and more in electronic publication of statistical information, *i.e.* statistical information on floppy disks, CD-ROM, *etc.* To help the users of this type of information in selecting the subset of information they need, user-friendly software must be made available to them. This software is now being developed.

## REFERENCES

BETHLEHEM, J.G., VAN BUITENEN, A.A.A., HUNDEPOOL, A.J., ROESINGH, M.J., and VAN DE WETERING, A. (1989a). Abacus 1.0, A Tabulation Package, Compact Guide. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN, L.F.M. (1989b). Blaise 2.0/An Introduction. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN,L.F.M. (1989c). Blaise 2.0/Language Reference Manual. CBS report, Voorburg: Netherlands Central Bureau of Statistics.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Offical Statistics* 3, 141-154.

KELLER, W.J., BETHLEHEM, J.G., and METZ, K.J. (1990). The impact of micromputers on survey processing at the Netherlands Central Bureau of Statistics. *Proceedings of 1990 Annual Research Conference, U.S. Bureau of the Census*, 637-645.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 57–68
Statistics Canada



# Research and Testing of Telephone Survey Methods at Statistics Canada

J. DOUGLAS DREW[1]

ABSTRACT

Findings from the research and testing of telephone and computer assisted survey methods for household surveys are presented, followed by discussion of how these findings will influence the redesign of household surveys at Statistics Canada during the 1990's. Significant emphasis is given in the presentation to the Canadian Labour Force Survey.

KEY WORDS: Data collection; Household surveys; Sample design.

## 1. INTRODUCTION

The 1980's have seen significant changes in survey taking due to advances in technology and the development of modern telephone survey methods, and the pace of these changes will likely accelerate during the 1990's. In this paper we will describe the research, testing, and development of methods that will form the infrastructure underpinning the data collection activities for Statistics Canada's household surveys during the 1990's. This research has focused in particular on the Canadian Labour Force Survey, and was carried out from 1985 to 1989 with a view to identifying improvements to be implemented during the 1991 post-censal redesign of the survey.

## 2. RESEARCH AND TESTING PROGRAM FOR LFS

The Canadian Labour Force Survey (LFS) is the largest household survey conducted by Statistics Canada, with a sample size of 62,300 households per month. It follows a rotating panel design in which households remain in the sample for six consecutive months, after which they are rotated out. It is based on a multi-stage area sample, with a decentralized interviewing staff of 1,000 local interviewers located across Canada and reporting to one of five Regional Offices.

Until the early 1970's, all interviewing was face to face. In 1972 telephone interviewing was introduced in large urban areas for follow-up interviews with households after they had received a face to face interview during their first month in the sample. In the literature, such telephone follow-up is referred to as "warm telephoning", to distinguish it from "cold telephoning" where the telephone interview is not preceded by a face to face interview (Groves *et al.* 1988).

The warm telephoning was initially restricted to major urban areas due to the frequency of party lines in smaller urban and rural areas and concerns this raised about the confidentiality of the data being collected. However, during the 1981 redesign of the survey, warm telephone interviewing was tested for the small urban and rural areas, and it was found respondents were willing to be interviewed by telephone, and the procedure had no impact on response rates or survey estimates (Choudhry 1984). The extension of telephoning to these areas in 1984 resulted in a 10% reduction in the data collection costs for the survey.

[1] J. Douglas Drew, Assistant Director, Household Surveys Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

In 1985, following introduction of a redesigned sample, a research program was started to investigate what further improvements in data collection could result from: (i) more use of the telephone in collection, (ii) telephone survey methods where the telephone is used both for sampling and for data collection, and (iii) Computer Assisted Interviewing (CAI) methods.

In the study of telephone and CAI methods it was useful to characterize the survey design in terms of a number of design factors as follows:

(i) **Mode of collection**. As already noted, the current mode of collection is warm telephoning, with an initial face to face interview and predominantly telephone interviews in later months. Alternative modes of collection include cold telephone interviewing with face to face follow-up of telephone nonrespondents, and cold telephoning without face to face follow-up.

(ii) **Organization of interviewers**. The LFS currently features a local organization, with the local interviewers doing a mixture of face to face interviews and telephone interviews from their homes. An alternative is a central organization, with interviewers working out of one or more central cites – in the case of Statistics Canada, its five Regional Offices across the country. A third organizational model is a mixed one where interviewing is done by a combination of local and central interviewers.

(iii) **Technology**. The current technology for the survey is traditional paper and pencil. The alternative technology considered most viable for household surveys is Computer Assisted Interviewing. CAI is commonly referred to as CATI (Computer Assisted Telephone Interviewing) when done centrally, and CAPI (Computer Assisted Personal Interviewing) when done locally using portable computers that interviewers use for face to face interviews and for telephone interviewing from their homes.

(iv) **Frame and sample design**. The survey has been based on an area sampling design since its inception in 1945. Alternatives include: telephone frames based on either Random Digit Dialing (RDD) methods (Waksberg 1978) or a combined use of RDD and list frames of published telephone numbers; other list frames which unlike telephone lists, are conceptually complete; and dual frame methods combining two or more of the above frame options.

In the following sections findings from the research and testing program pertaining to each of these design factors are discussed.

## 3. MODE OF COLLECTION

A major test was carried out from 1985 to 1989 to determine the impact of cold telephone interviewing with face to face follow-up as an alternative to the current warm telephoning. This test was referred to as the Telephone First Interview Test. The test was embedded into the ongoing LFS in urban areas of Quebec and Ontario. The methodology is reported fully in Drew, Choudhry and Hunter (1988). In brief, newly sampled LFS dwellings were matched to lists purchased from the telephone companies on the basis of address information. Match rates of 65% were obtained. Test and control samples were selected from the matched dwellings such that each test dwelling was paired with a control dwelling within the same sampling unit (city block). For the test dwellings, telephone numbers were provided to interviewers, who were to attempt a telephone interview, but to use face to face follow-up if necessary. Interviewers were unaware of the existence of a control sample, and followed normal procedures for all dwellings for which telephone numbers were not provided.

**Table 1**

Telephone First Interview Test (October 1985 - March 1989)
Estimate for Test Treatment as Percent of Estimate for Control Treatments

| Characteristic | Ontario | | Quebec | |
|---|---|---|---|---|
| | Percent | t-statistic | Percent | t-statistic |
| Employment | 98.5 | −1.22 | 97.2 | −1.64 |
| Unemployment | 96.3 | −0.94 | 111.8 | 2.27* |
| Not in LF | 101.1 | 0.88 | 98.2 | −0.63 |
| Pop. 15+ | 909.2 | −0.86 | 98.5 | −1.31 |
| Pop. in Hhld = 3+ | 97.8 | −0.76 | 94.1 | −1.26 |
| Pop. in 1 person Hhlds | 100.6 | 0.25 | 104.1 | −0.93 |
| Pop. in 2 person Hhlds | 101.1 | 0.26 | 101.0 | 0.43 |
| Pop. in 3 person Hhlds | 98.6 | −0.56 | 94.7 | −1.28 |
| Emp. Male 15-24 | 95.2 | −0.75 | 88.7 | −2.00* |
| Emp. Male 25+ | 98.5 | −1.11 | 96.7 | −1.54 |
| Emp. Female 15-24 | 99.0 | −0.11 | 111.5 | 1.44 |
| Emp. Female 25+ | 99.2 | −0.65 | 97.1 | −1.11 |
| Unemp. Male 15-24 | 105.6 | 0.53 | 119.0 | 1.53 |
| Unemp. Male 25+ | 94.5 | −0.80 | 96.7 | −0.31 |
| Unemp. Female 15-24 | 99.6 | −0.15 | 119.4 | 1.30 |
| Unemp. Female 25+ | 90.9 | −1.22 | 123.9 | 2.71** |
| Not in LF Male 15-24 | 99.8 | −0.07 | 99.9 | 0.47 |
| Not in LF Male 25+ | 105.6 | 1.57 | 101.3 | 0.07 |
| Not in LF Female 15-24 | 101.9 | 0.56 | 95.3 | −0.56 |
| Not in LF Female 25+ | 99.2 | −0.14 | 97.3 | −0.92 |
| Pop. Male 15-24 | 97.2 | −0.49 | 95.2 | −0.73 |
| Pop. Male 25+ | 99.9 | −0.09 | 97.7 | −1.49 |
| Pop. Female 15-24 | 99.9 | 0.16 | 105.4 | 0.95 |
| Pop. Female 25+ | 98.9 | −1.28 | 98.4 | −1.24 |

* t-statistic significant at 5% level
** t-statistic significant at 1% level

No significant differences in response rates were found between the test and control samples. For Quebec, response rates were 96.1% for both samples, while for Ontario, the rate of 96.3% for the test sample was marginally lower than that of 96.5% for the control sample.

When comparing the labour force estimates obtained in the test and control samples, certain estimates from Quebec early in the test for the period October 1985 to February 1987 were significantly different. In particular the employed and unemployed males in households of three or more persons were underestimated in the test sample (see Drew, Choudhry and Hunter 1988). Table 1 presents data over the full life of the test from October 1985 to March 1989. For Quebec, a few statistically significant differences existed - which stemmed from the influence of the earlier time period. When the data are analyzed from March 1987 onwards, these differences are not significant. In Ontario, there were no significant differences. Speculating on the differences in Quebec, their co-incidence with a program of inspection of welfare recipients carried out by the provincial government suggests that measures external to the survey led to a climate in which there was decreased trust of cold telephone interviews. We were fortunate

**Table 2**

Nonresponse Rates: Warm Telephoning versus Cold Telephoning with/without Face to Face Follow-up

| Method | Test 1 | Test 2 |
|---|---|---|
| Warm telephoning with letter (ongoing LFS) | 4.1 | 5.6 |
| Cold telephoning with letter | 6.9 | 9.8 |
| Cold telephoning without letter | 8.5 | - |

Test 1: October 1985 - September 1986; Ontario and Quebec
Test 2: July 1988 - March 1989; Nova Scotia and Alberta

to have been conducting the test during this period, because the finding that survey results obtained under cold telephoning are more subject to external influences than are those obtained under warm telephone interviewing will be an important consideration in any decisions on extension of telephone interviewing.

Cold telephoning without face to face follow-up was also studied. Two Telephone Sampling Tests were carried out in which the LFS was conducted as a central telephone survey, with interviewing from the Regional Offices. Nonresponse rates for the two tests and comparable rates for the ongoing LFS are presented in Table 2.

The first test studied two sampling methods - RDD, and a combination of list sampling for published numbers and RDD for nonpublished numbers. The list sampling featured introductory letters, but the RDD sampling did not. Differences in response rates would seem to point to the positive effects on response rates of an advance letter. For both tests, the comparison of warm versus cold telephone interviewing revealed nonresponse rates for cold telephoning which were higher at a 5% significance level. The second test was based solely on a list sample of published numbers.

An important issue is the nonresponse bias, if any, resulting from the higher nonresponse under cold telephoning without face to face follow-up. As a proxy to these extra nonrespondents, Laflamme (1990) looked at non-first-month-in-sample households from the ongoing LFS who had a telephone, but who were interviewed face to face. He found that size of the proxy group, at 3.5% of respondents, was close to the size of the extra nonresponse under cold telephoning without face to face follow-up. Further, he found the unemployment rate for the proxy group was 12.8%, versus 7.4% for persons in households interviewed by telephone. Exclusion of the proxy group from the sample would have lowered the national unemployment rate from 8.1% to 7.9%. This is clearly a serious bias, given the accuracy required for the LFS national estimates. As the proxy assumption seems a reasonable one, these findings raise serious concerns about cold telephone interviewing without face to face follow-up for the LFS.

Table 3 compares unemployment and participation rates for the first telephone sampling test with corresponding estimates from telephone households in the LFS. The only estimate found to be significantly different at a five percent level from estimates produced for the telephone population from the ongoing LFS was the unemployment rate for Quebec for the RDD treatment. It is worth noting that the test was carried out at the same time that problems emerged with estimates for Quebec in the Telephone First Interview Test. Another point worth noting is that while other differences in unemployment rates were not statistically significant, the rates for cold telephoning were higher. Other researchers have observed differences in the same direction, also without being able to attribute statistical significance to them. These data might benefit from a meta analysis.

**Table 3**

Telephone Sampling Test (October 1985 - September 1986)
Unemployment and Participation Rates

| Province | Design | Unemployment Rate (S.D.) | | Participation Rate (S.D.) | |
|---|---|---|---|---|---|
| Quebec | LIST | 12.3 | (0.78) | 64.1 | (1.08) |
| | RDD | 13.0* | (0.88) | 62.8 | (1.28) |
| | LFS | 10.9 | (0.27) | 63.4 | (0.29) |
| Ontario | LIST | 7.3 | (0.59) | 69.0 | (1.11) |
| | RDD | 7.9 | (0.63) | 69.0 | (1.18) |
| | LFS | 6.9 | (0.16) | 69.0 | (0.20) |

* Significant difference between RDD and LFS Unemployment rates for Quebec

In summary, the test results showed cold telephoning without face to face follow-up yielded higher nonresponse rates than the current warm telephoning method, and while inconclusive, there was some evidence that it yielded higher unemployment rates. On the other hand, cold telephoning with face to face follow-up, apart from the one period of time in Quebec, was found to yield data comparable to that under warm telephoning.

On the basis of these findings, it was decided to implement cold telephoning with face to face follow-up for the LFS apartment frame sample, which constitutes roughly 4% of the overall sample. The availability of the telephone number for apartment frame units, it was reasoned, would help overcome problems in gaining access to highrise apartment buildings, and allow for more attempts to find persons at home than is feasible with face to face interviewing. These expectations seem to have been borne out. As reported by Dufour (1990), while first month nonresponse rates for the apartment sample continue to be higher than corresponding first month rates for the non-apartment sample, the gap has narrowed from a difference of 8.7 percentage points in the year before the change to a difference of 5.7 percentage points during the first five months under the new procedure.

Another change to the mode of collection for the ongoing LFS was to introduce telephone follow-up of the first month in sample households which could not be contacted during an initial visit to the dwelling. This procedure was introduced in 1986, and led to a $100,000 per year savings in data collection costs.

The combined effect of the telephone first interview for the apartments, and the telephone follow-up for first month nonrespondents has been an increase in the overall telephoning rate for the survey from 80% in 1985 to 83% in 1990.

## 4. ORGANIZATION OF INTERVIEWING STAFF

During the testing program, two alternatives to the current local organization of the interviewing staff were studied. The telephone sampling tests already described considered a "central" organization where all of the interviewing was done out of the Regional Offices. Another test examined a mixed organization, in which the current warm telephoning mode of collection was retained. The test of the mixed organization was carried out from January 1988 to March 1989 in two Census Metropolitan Areas in which Regional Offices are located - Montreal and Halifax. Its primary objective was to measure the cost implications of such a mixed organization.

The test methodology consisted of face to face collection by local interviewers for first month in sample cases, and telephoning by central interviewing staff working out of the Regional Offices for most non-first month cases. Whenever nonresponse follow-up was required for households initially assigned to the central interviewers, this was carried out by the local interviewers. This methodology was initially tested for the Labour Force Survey by Muirhead *et al.* (1975) and has been extensively studied by the United States Bureau of the Census (1987), where the centralized interviewing is being done using Computer Assisted Telephone Interviewing (CATI).

One of the complexities of the method was the practice followed for the first half of the test of transferring cases requiring nonresponse follow-up from the central to the local interviewers at the mid-point of the interviewing week. For the second half of the test, this so-called re-cycling was restricted to cases where the telephone number was determined to be no longer valid. During the first half of the test, nonresponse rates were 8.0% for the test treatment versus 6.1% for the control procedures corresponding to the decentralized interviewing used for the ongoing LFS. The gap narrowed to 7.3% versus 6.7% during the second half.

From the first telephone sampling test, interviewing costs per household were estimated to be $2.72 for central data collection with telephone list sampling, versus $3.53 for RDD sampling. The extra costs for RDD methods is due to the time spent in screening for residential telephone numbers. These costs include $0.46 per household for long distance charges. This amount was estimated based on long distance rates and data on length of calls, since record keeping practices in the regional offices did not permit the extraction of actual costs incurred. Comparable costs for the ongoing LFS were $4.76 per household for interviewer fees and expenses. The test of the mixed organization yielded savings relative to the ongoing LFS of $0.78 per household in interviewer fees and expenses. The above cost comparisons do not factor cost of office space and equipment into the costs under the centralized and mixed organizations. Nor do they consider the costs of transferring documents to and from local interviewers under the mixed and local organizations, which under the current paper and pencil technology is accomplished by express mailing of documents, but under CAI scenarios would be transmitted electronically.

The mixed organization was considered only for Regional Office cities, as extension beyond Regional Office cities would imply greater long distance telephoning. More importantly, the sample design for smaller urban and rural areas is clustered so that primary sampling units yield sample sizes corresponding to an interviewer assignment. Centralization of the telephone portion of the sample would necessitate more clustering of the sample in order to retain a sufficient workload for the local interviewers. Also in medium sized urban centres where there are currently four to five interviewers, the number of local interviewers under the mixed organization would be reduced to one or two, significantly reducing the flexibility to have interviewers fill in for one and other during vacations and illness.

Of the three organizational models considered, all had advantages and disadvantages. The local organization yielded the lowest nonresponse rates, albeit at the highest per unit data collection cost. The mixed organization had marginally higher nonresponse and marginally lower costs, and was limited in where it could be applied. In the final analysis, the mixed organization was seen as introducing a lot of complexity for at best marginal gain. The central organization, which offered substantial savings in data collection costs, resulted in a 68-75% increase in nonresponse relative to the local organization. As discussed in section (3), there is evidence that this extra nonresponse would introduce a serious nonresponse bias into the LFS estimates. Moreover there are concerns that the gap in nonresponse rates attainable under local versus central organizations might widen in the future, as increasing exposure to telephone

solicitation and increasing availability of telephone screening technology renders the population less receptive to telephone interviewing. Such developments favour survey design strategies which, although they may allow for flexibility to do telephoning, also allow for face to face follow-up wherever needed. The local organization best offers this flexibility. On the basis of the above considerations it has been decided to retain the current local organization.

## 5. TECHNOLOGY

Catlin, Ingram, and Hunter (1988) carried out a controlled study comparing CATI and paper and pencil interviewing. In the study the LFS questionnaire was administered to RDD samples of 1,000 households per month per treatment over a period of nine months. All interviewing took place from Statistics Canada headquarters in Ottawa.

The study was part of a collaborative research effort with the United States Bureau of the Census (USBC), and the CATI software used was developed by the USBC. The wording of the questionnaires was purposively the same for both treatments. Features unique to the CATI treatment were automatic branching, some basic on-line edits, and automated call scheduling.

Three quality improvements were discernable for CATI relative to paper and pencil methods. First, the overall rate of edit failures during post-collection data processing was 50% lower for CATI. Second, there was a virtual elimination of branching errors under CATI. Importantly, this occurred for certain portions of the questionnaire, which, although infrequently encountered, have a bearing on determination of labour force status, and which under paper and pencil interviewing are subject to high levels of branching errors. Third, the average household size reported under CATI was 3% higher, which represents roughly a 50% reduction in the underenumeration in the LFS relative to the Census. This improvement seems to stem from the enforced probing built into the CATI instrument for additional household members and for persons temporarily away.

Based on these findings, it has been decided that the introduction of Computer Assisted Interviewing should be one of the major thrusts of the 1991 post-censal redesign of the LFS. Due to the preference for maintaining a local organization of interviewing staff, a CAPI implementation is being planned for.

## 6. FRAME AND SAMPLE DESIGN

### Telephone Frames

Telephone coverage and the extent to which characteristics of those without telephones differ from characteristics of those with telephones are important factors in the design of telephone survey methods – particularly as regards frame strategies.

In an international review of telephone coverage, Trewin and Lee (1988) found telephone coverage in Canada to be one of the highest in the world at 97-98%. As is typical of the situation in most countries they surveyed, persons in non-telephone households in Canada tend to have lower incomes and higher rates of unemployment.

Table 4 gives the percentage of non-telephone households in Canada from 1976 to the present. Telephone coverage, while already high in 1976 has been steadily edging upwards, although it appears to have levelled off over the last few years at around 98.5%.

**Table 4**

Non-telephone Households by Province (%)

| | 1976 | 1981 | 1985 | 1987 | 1990 |
|---|---|---|---|---|---|
| Canada | 3.5 | 2.4 | 1.8 | 1.5 | 1.5 |
| Newfoundland | 10.0 | 6.0 | 5.1 | 3.6 | 1.9 |
| Prince Edward Island | – | – | – | – | 2.8 |
| Nova Scotia | 7.5 | 4.6 | 3.5 | 3.2 | 1.5 |
| New Brunswick | 5.8 | 5.3 | 5.3 | 3.3 | 2.2 |
| Quebec | 3.3 | 2.1 | 1.6 | 1.5 | 1.5 |
| Ontario | 2.5 | 1.9 | 1.0 | 1.0 | 1.2 |
| Manitoba | 4.1 | 2.3 | 2.7 | 2.4 | 1.7 |
| Saskatchewan | 3.6 | 2.5 | 2.3 | 2.4 | 2.3 |
| Alberta | 3.0 | 2.4 | 2.0 | 1.8 | 2.0 |
| British Columbia | 4.2 | 2.8 | 2.4 | 1.3 | 1.5 |

Source: Statistics Canada, Estimates from Household Facilities & Equipment Survey

**Table 5**

Labour Force Characteristics by Telephone Status

| Province | Telephone Status | Unemployment Rate | Participation Rate |
|---|---|---|---|
| Nova Scotia | published | 9.0 | 71.9 |
| | non-published | 9.8 | 70.2 |
| | non-telephone | 17.2 | 62.3 |
| Alberta | published | 6.3 | 80.7 |
| | non-published | 8.2 | 81.5 |
| | non-telephone | 11.1 | 67.0 |

Laflamme (1990) undertook a study comparing characteristics of the non-telephone and telephone universes. The study included a breakdown of those with published versus non-published numbers, obtained by linking telephone numbers supplied by LFS respondents to lists of published telephone numbers. Two provinces were included in the study, Nova Scotia and Alberta. For Nova Scotia 9.7% of numbers, and for Alberta 11.2% of numbers were found to be non-published. Unemployment and participation rates reported by Laflamme are given in Table 5.

This study replicates findings from earlier studies that the labour force characteristics of persons without telephones are very different from those with telephones. The labour force characteristics differ but to a lesser extent between persons with published versus non-published numbers.

While the non-telephone population accounts for only 1.0%-1.5% of the population, the differences in labour force characteristics are sufficiently large that simply excluding the non-telephone population is not a viable option for the LFS, given the accuracy required for the national employment and unemployment estimates (coefficients of variation of 0.5% and 2% respectively).

Another difficulty with telephone frames, particulary for panel surveys is their rapid deterioration. Drew, Dick and Switzer (1989) found a 0.5 - 1.0 % rate of additions and deletions to the stock of published residential numbers per month. Hence telephone samples cannot remain representative of the telephone universe for very long unless they are updated. The authors proposed a strategy of updating samples for a panel survey using files of published numbers acquired on an ongoing basis from telephone companies. An operational test of the procedure over a nine month period was a success. Their procedure applied only to published numbers sampled from a list frame, and did not provide a solution to the problem of keeping a sample selected using Random Digit Dialing methods up to date over the life of a panel.

Because of the coverage and updating problems with telephone frames, approaches where dwellings as opposed to telephone numbers are the samping units are seen as having more promise for large scale panel surveys. It is worth noting that the situation can be quite different for other surveys. Catlin *et al.* (1984) showed that coverage biases for general population characteristics were less than for labour force characteristics. Further for smaller surveys (*e.g.*, those with sample sizes of 10,000 or less) small biases are less important given the larger relative sampling errors for these surveys. These findings led to establishment of an RDD household survey capacity in 1986. It has been used for numerous one-time surveys and for the General Social Survey, which is an annual survey of 10,000 households.

**Area Frame**

As has already been described, it is possible in urban areas to match selected addresses from an area frame to telephone lists in order to permit cold telephone interviewing. The experiences with the LFS have been that telephone numbers can be obtained in this fashion for approximately 60% of households. These match rates are based on exact matching after standardization of the address information, and they could be improved through use of record linkage methods. With telephoning for a substantial portion of the first month cases, the clustering of the sample could be reduced somewhat, but a clustered sample remains a constraint imposed by an area frame design.

It is planned to investigate the feasibility of extending these procedures to rural areas, which would entail changing the type of information collected when dwelling lists are created for the survey. The information currently collected tends to be descriptive of the physical characteristics of the dwelling, whereas to successfully match with lists of telephone subscribers, information such as name (often readily available on mail boxes), street name and number, or in their absence the rural route number and postal code, would be required.

**Address Register**

Statistics Canada is constructing an Address Register in urban areas of Canada. It will be used in the 1991 Census to improve coverage by providing an independent check on the dwelling lists created by the Census enumerators (Drew, Royce and van Baaren 1989). The Address Register will be a machine readable list of addresses constructed by linkage of various administrative data sources, including lists of customers with published numbers purchased from telephone companies. During the use of the Address Register in the 1991 Census, its coverage will be updated to correspond to that of the 1991 Census.

It is planned during the 1991 post-censal redesign of the Labour Force Survey to conduct studies into use of the Address Register as a frame for household surveys in urban areas. If the conclusion is that it should be adopted as a frame, the Address Register will be updated on an ongoing basis following the 1991 Census.

An advantage of the Address Register as a frame over the area frame is that telephone and non-telephone households are known ahead of time. Hence the two can be sampled as separate strata - with a reduced amount of clustering for the telephone stratum, for which a significant portion of the first month interviews could be done by telephone. The non-telephone stratum would include those households with non-published numbers and those households without telephones. Evidence from earlier studies showed the refusal rate when cold telephoning households with non-published numbers to be 12% compared to 4% for all households. Under warm telephoning there is no corresponding increase in the refusal rates for households with non-published numbers, and there is a good success rate in converting these households to respond by telephone in later months. This finding and the desire to be sensitive to privacy concerns of individuals support the face to face interview of such households, which account for an estimated 10-15% of numbers.

**Dual Frame**

In urban areas, if the coverage of the Address Register as the sole frame is not adequate, a dual frame design in which the Address Register is supplemented by a small area sample will be considered. There are different forms the supplementary sample could take. A promising option, not involving the expense of building and maintaining both a conventional area frame and an Address Register, would be to use an interval approach in which a sample of consecutive dwellings on the Address Register would be selected and checked in the field. Any dwellings found between the Address Register dwellings constitute a sample of dwellings missing from the Address Register.

Mian (1990) has studied dual frame methods considering a cost and variance optimization for the general case where neither of the frames needs to cover the entire universe. This was felt to be a practical model since the area frame, while conceptually complete, in practice suffers from 3-4% undercoverage relative to the Census, in addition to the 5% of the population which is not represented because of nonresponse. Extension of Mian's model to include a non-sampling error component will permit factoring into the optimization what we know or may wish to assume about the coverage and nonresponse biases under alternative frame approaches. It can be used in the context of dual frames combining the Address Register and area frames in urban areas, and combining area and telephone frames in rural areas.

## 7. 1991 POST-CENSAL REDESIGN OF LFS

The Labour Force Survey is redesigned following each decennial population census. Redesigns have normally focused on redesign of the sample, but in the 1970's a major revision was carried out encompassing a sample redesign, changes to the questionnaire content, wording of questions, and survey outputs, and a major overhaul of the survey processing systems including introduction of a network of mini-computers in the regional offices to support survey operations and regional data capture. In contrast, redesign efforts during the 1980's were restricted to a sample redesign.

While decisions on the scope of the 1991 post-censal redesign have yet to be taken, an effort falling somewhere between the major revision in the 70's and minimal redesign in the 80's appears needed. The work on the redesign is at an early planning stage, and is proceeding through four sub-projects focusing on: (i) content and questionnaire issues, (ii) modernization of the survey processing systems and review of survey outputs, (iii) development, testing, and implementation of Computer Assisted Interviewing, and (iv) sample redesign. (Drew *et al.* 1991).

Sub-projects (iii) and (iv) are those which will be concerned with telephone and CAI methods. Current plans for these sub-projects as they relate to the survey design factors described in earlier sections of this paper are briefly summarized below.

**Technology and Organization**

Based on the positive findings from testing of Computer Assisted Interviewing for the LFS reported by Catlin *et al.* (1988), it has been decided to make the adoption of CAI one of the major thrusts of the redesign. Moreover, for reasons already discussed, a decision has been taken that the current local organization of the interviewing staff should be retained, so that the implementation of CAI methods will take the form of Computer Assisted Personal Interviewing (CAPI). Specifically, local interviewers will be equipped with lightweight portable computers that they will carry with them for face to face interviewing, and that they will use to conduct telephone interviews from their homes. The mix of face to face and telephone interviewing may remain much the same as it is currently - 83% telephone and 17% face to face - or it may shift to more telephoning if it is decided to adopt more cold telephone interviewing of households during their first month in the sample.

The work plan for the Computer Assisted Interviewing sub-project includes a field test during 1991 of a touch screen portable computer, and in later years development or acquisition of CAI software, a combined test of CAI and questionnaire alternatives, development of on-line editing, and an automated version of the interviewer manual embedded into a help screen accessible during interviewing.

**Frame and Mode of Collection**

A key research finding was that cold telephone interviewing with face to face follow-up yields response rates and labour force estimates comparable to those under the current warm telephone collection procedure for the LFS, which features face to face interviewing for the first month households in the sample. One exception was observed in Quebec, where, as discussed, for a period of time data differences were seen. The importance of face to face follow-up in maintaining response rates favours the retention of the dwelling as the sampling unit, and supplying the telephone number to interviewers to give them greater flexibility to use both telephone and face to face interviewing in obtaining first month interviews.

In urban areas, both the current area frame and the Address Register as a frame are consistent with this approach. In rural areas, as described earlier in the paper, research into the feasibility of matching area frame addresses to telephone lists to provide interviewers with telephone numbers will be studied. It is also planned to continue to investigate dual frame methods which in urban areas might consist of the Address Register and an area frame, and in rural areas an area frame and a telephone frame.

## 8. SUMMARY

The current data collection methodology for the Labour Force Survey consists of a face to face interview for households during their first month in the sample and predominantly telephone interviewing in later months. The introduction of telephone interviewing for the later months took place during the 1970's for major urban areas and during the 1980's for remaining areas. In both instances the introduction of telephone interviewing resulted in significant cost savings without any impact on the response rates or survey estimates. Prior to the telephone and CAI research and testing program begun in 1985, 80% of LFS interviews were done by

telephone. This has moderately increased to 83% through the introduction of telephone follow-up for households which could not be contacted during an initial face to face visit, and by supplying interviewers with telephone numbers for the apartment sample.

The primary benefit of the research and testing program has been to identify the frame and data collection options to pursue during the 1991 post-censal redesign of the survey, including the retention of the current local organization of interviewers,the adoption of Computer Assisted Personal Interviewing, the retention of frame and sample design approaches in which the dwelling is the unit of selection, and the provision of interviewers with telephone numbers to permit the flexibility to use a combination of telephone and face to face interviewing to obtain first month interviews.

## REFERENCES

CATLIN, G., CHOUDHRY, H., and HOFMANN, H. (1984). Telephone ownership in Canada. Internal report, Statistics Canada.

CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and data quality. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*), 437-450. New York: Wiley.

CHOUDHRY, G.H. (1984). Results from telephone interviewing experiment in the non self representing areas of the Labour Force Survey. Internal report, Statistics Canada.

DREW, J.D., and GAMBINO, J. (1991). Plans for the 1991 post censal redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

DREW, J.D., CHOUDHRY, H., and HUNTER, L. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Metholodogy*, (Eds. R. Groves *et al.*), 233-246. New York: Wiley.

DREW, J.D., DICK, P., and SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DREW, J.D., and JAWORSKI, R. (1986). Telephone survey development on the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DREW, J.D., ROYCE, D., and van BAAREN, A. (1989). Address register research at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DUFOUR, J. (1990). Implantation de la première entrevue par téléphone pour la base d'appartements de l'enquête sur la population active. Internal report, Social Survey Methods Division, Statistics Canada.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II, and WAKSBERG, J. (Eds.) (1988). *Telephone Survey Methodology*. New York: Wiley.

LAFLAMME, F. (1990). Étude comparative entre trois différentes populations visées par l'EPA selon leur type de service téléphonique. Internal report, Social Survey Methods Division, Statistics Canada.

MIAN, I.U.H. (1990). Dual frame estimation of proportions in sample surveys. Internal report, Social Survey Methods Division, Statistics Canada.

MUIRHEAD, R.C., GOWER, A.R., and NEWTON, F.T. (1975). The telephone experiment in the Canadian Labour Force Survey. *Survey Methodology*, 1, 158-180.

TREWIN, D., and LEE, H. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*). New York: Wiley, 9-24.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 576-579.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 69–78
Statistics Canada



# Marginal and Approximate Conditional Likelihoods for Sampling on Successive Occasions

D.R. BELLHOUSE[1]

ABSTRACT

Marginal and approximate conditional likelihoods are given for the correlation parameters in a normal linear regression model with correlated errors. This general likelihood approach is applied to obtain marginal and approximate conditional likelihoods for the correlation parameters in sampling on successive occasions under both simple random sampling on each occasion and more complex surveys.

KEY WORDS: Likelihood inference; Sampling in time; ARMA models; State space models.

## 1. INTRODUCTION

Consider a finite population of $N$ units which may be sampled on $k$ occasions. Let $y_{tj}$ denote the measurement on the $j^{th}$ population unit taken on the $t^{th}$ occasion; $j = 1, \ldots, N$ and $t = 1, \ldots, k$. It is assumed that any two units, say $j$ and $j'$, are independent, but that measurements of the same unit across time are correlated. In particular, assume that for any $j$,

$$(y_{1j}, y_{2j}, \ldots, y_{kj})^T \sim N(\mu, \sigma^2\Omega), \tag{1}$$

where $\Omega$ is a $k \times k$ correlation matrix and where $\mu$ is the $1 \times k$ vector of fixed means $(\mu_1, \mu_2, \ldots, \mu_k)^T$. In view of the explicit model assumption in (1), a model-based approach to survey estimation is used in this paper. Based on samples taken over the $k$ occasions, it is of interest to estimate $(\mu_1, \mu_2, \ldots, \mu_k)^T$. The form of the model-based estimates $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)^T$, if obtained by maximum likelihood or generalized least squares, for example, will depend on $\sigma^2$ and the parameters in $\Omega$. It is therefore necessary to obtain good estimates of $\sigma^2$ and the parameters in $\Omega$.

The notation of Bellhouse (1989) is used to describe the sampling scheme considered here, namely one-level rotation sampling. On any occasion, $c$ rotation groups are sampled. Rotation group $r (r = 1, 2, \ldots, k + c - 1)$, denoted by $G_r$, consists of $m_r$ sample units. On occasion $t (t = 1, \ldots, k)$, the sample consists of the units in $G_t, G_{t+1}, \ldots, G_{t+c-1}$, so that the total sample size on occasion $t$, $n_t = m_t + m_{t+1} + \ldots + m_{t+c-1}$. On occasion $t + 1$, $G_t$ is dropped from the sample and $G_{t+c}$ is added. Each rotation group is chosen without replacement from previously unchosen units in the population. The total sample size over all $k$ occasions is $m = n_1 + n_2 + \ldots + n_k$. The maximum number of occasions that a unit remains in the sample is $c$.

If $c$ is small, then estimates of the correlation parameters in $\Omega$ can be unstable, leading to instability in the estimates of interest $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)^T$. Viewed another way, the total number of parameters is at least $k + 2$ and increases with time, *i.e.* with the addition of new occasions. Since the dimension of the parameter space increases with time, maximum likelihood estimates of parameters may be biased and inconsistent. The problem of the stability of estimates has been addressed in sampling on successive occasions, for example, by Blight

[1] D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B9.

and Scott (1973), who assume that the elements of $(\mu_1, \mu_2, \ldots, \mu_k)^T$ follow a time series process. On using this assumption the dimension of the parameter space is fixed at a relatively small number so that the problems of instability, bias and inconsistency are resolved. In this paper, a different approach is taken. Here the fixed means assumption is retained and marginal and approximate conditional likelihoods are derived for the parameters in $\Omega$, treating the fixed means as nuisance parameters.

Marginal likelihood estimation was introduced as a general method for eliminating nuisance parameters from the likelihood function (Fraser 1967; Kalbfleisch and Sprott 1970). Cox and Reid (1987) introduced approximate conditional likelihoods which also address this problem. They argued that the approximate conditional likelihood was preferable to the profile likelihood obtained by replacing the nuisance parameters in the likelihood by their maximum likelihood estimates when the parameters of interest are given. Bellhouse (1990) established the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model.

Following on the work of Cox and Reid, Cruddas *et al.* (1989) obtained an approximate conditional likelihood for the correlation parameter in several short series of autoregressive processes of order one with common variance and autocorrelation parameters. Based on a simulation study, Cruddas *et al.* (1989) showed that the estimate based on the approximate conditional likelihood had a much smaller bias and better coverage properties of the confidence interval than the maximum likelihood estimate from the profile likelihood. The situation described by Cruddas *et al.* (1989) applies directly to sampling on successive occasions in sample surveys. In order to reduce the response burden, individuals in a survey are retained in the sample for relatively short periods of time. It is expected that the use of marginal and approximate conditional likelihoods will improve the estimates of correlation parameters and consequently improve the estimates of the mean for each occasion.

Within a rotation group, the sample measurements on an individual are usually modelled by an autoregressive moving average process (ARMA), *i.e.* the parameters in $\Omega$ are comprised of the correlation parameters in the ARMA process. See Binder and Hidiroglou (1988) for a review of the application of time series models to sampling on successive occasions. Consequently, it is of interest to obtain marginal and approximate conditional likelihoods under ARMA models with application to rotation sampling. The marginal and approximate conditional likelihoods for the correlation parameters in a normal model are obtained in Section 2. The general results of Section 2 are illustrated in Section 3 by applying the results to sampling on successive occasions assuming simple random sampling of units in rotation groups. In Section 4, some methods are given to apply these likelihood methods to complex surveys.

## 2. MARGINAL AND APPROXIMATE CONDITIONAL LIKELIHOODS FOR CORRELATION PARAMETERS UNDER A NORMAL MODEL

Let $y$ be a vector of sampled observations of dimension $m \times 1$ which follows the linear model

$$y = X\beta + \epsilon \tag{2}$$

with error vector $\epsilon \sim N(\mathbf{0},\sigma^2\Phi)$, where $\Phi$ is the $m \times m$ correlation matrix and where $\beta$ is the $p \times 1$ vector of regression coefficients so that $X$ is $m \times p$. The log-likelihood for $\beta$, $\sigma^2$ and $\Phi$ is given by

$$L(\beta,\sigma^2,\Phi) = -\left\{m \ln\sigma + (\ln \mid \Phi \mid)/2 + (y - X\beta)^T\Phi^{-1}(y - X\beta)/(2\sigma^2)\right\}. \tag{3}$$

For a given value of $\Phi$,

$$\hat{\beta} = (X^T\Phi^{-1}X)X^T\Phi^{-1}y$$

and

$$s^2 = y^T\Phi^{-1}y - y^T\Phi^{-1}X\,(X^T\Phi^{-1}X)^{-1}\,X^T\Phi^{-1}y \tag{4}$$

are jointly sufficient for $\beta$ and $\sigma^2$.

A marginal likelihood for $\Phi$ is obtained by making a transformation of the data $y$ to the sufficient statistics $\hat{\beta}$ and $s^2$ and the ancillary statistic

$$a = \Phi^{-\frac{1}{2}}(y - X\,(X^T\Phi^{-1}X)^{-1}\,X^T\Phi^{-1}y)/s,$$

where $\Phi^{-\frac{1}{2}}$ is the $m \times m$ dimensional matrix such that $\Phi^{-1} = \Phi^{-\frac{1}{2}}\Phi^{-\frac{1}{2}}$. The marginal likelihood for $\Phi$ is the marginal distribution of the ancillary $a$ times the product of the differentials $da_i$, $i = 1, \ldots, m$. See Kalbfleisch and Sprott (1970, eqs. 6 and 10) for a general discussion and a general expression for $\Pi da_i$. Bellhouse (1978) and, later independently Tunnicliffe Wilson (1989), showed that the marginal likelihood for $\Phi$ under the normal model is given by

$$L_M(\Phi) = \left\{ |\,\Phi\,|^{\frac{1}{2}}\,|\,X^T\Phi^{-1}X\,|^{\frac{1}{2}}\,s^{m-p}\right\}^{-1}. \tag{5}$$

Note that (4) is proportional to the maximum likelihood estimate of $\sigma^2$ given $\Phi$ and that $s^2(X^T\Phi^{-1}X)^{-1}$ is proportional to the estimated variance-covariance matrix of the maximum likelihood estimate of $\beta$ given $\Phi$. Then (5) can be written as

$$L_M(\Phi) = \frac{|\,\text{est var}(\hat{\beta})\,|^{\frac{1}{2}}}{s^m\,|\,\Phi\,|^{\frac{1}{2}}}. \tag{6}$$

To obtain an approximate conditional likelihood, it is first necessary to transform the parameters to achieve parameter orthogonality between the parameters of interest and the nuisance parameters, which now may depend on the parameters of interest. Sets of parameters are orthogonal if the associated information matrix is block diagonal, with each block as the information matrix for each parameter set. The conditional likelihood is related to the distribution of the data $y$ conditional on the maximum likelihood estimate of the nuisance parameters for fixed values of the parameters of interest. The approximate conditional likelihood is obtained by applying two approximations to this conditional distribution. See Cox and Reid (1987, Section 4.1) for a discussion of the derivation. For example, let $\Theta$ be the vector of parameters of interest and let $\Lambda$, possibly depending on $\Theta$, be the vector of nuisance parameters orthogonal to $\Theta$. The full likelihood of the data for parameters $\Theta$ and $\Lambda$ is denoted by $L(\Theta,\Lambda)$ and the profile likelihood for $\Theta, L(\Theta,\hat{\Lambda})$ is the likelihood with $\Lambda$ replaced by its maximum likelihood estimate. The approximate conditional likelihood for $\Theta$ is

$$L(\Theta,\hat{\Lambda})\;|\;I(\Theta,\hat{\Lambda})\;|^{\frac{1}{2}},$$

where $I(\Theta,\hat{\Lambda})$ is the observed information matrix for $\Lambda$ at a fixed value of $\Theta$. See Cox and Reid (1987, eq. 10).

Following Cruddas *et al.* (1989), Bellhouse (1990) suggested, for model (2), the parameter transformation $\lambda = \ln\sigma + (\ln|\Omega|)/(2m)$ leaving $\beta$ the same. The log-likelihood under the new parameterization is denoted by $L(\beta, \lambda, \Phi)$ and can be obtained from (3). If the entries of $\Phi$ are functions of a parameter $\phi$, then the nuisance parameters $\lambda$ and $\beta$ are each orthogonal to $\Phi$, *i.e.*

$$-\frac{1}{m}E\left[\frac{\partial^2 L(\beta,\lambda,\Phi)}{\partial\phi\partial\lambda}\right] = 0$$

and

$$-\frac{1}{m}E\left[\frac{\partial^2 L(\beta,\lambda,\Phi)}{\partial\phi\partial\beta}\right] = 0,$$

when each entry of $\Phi$ is a continuous and differentiable function of $\phi$. Moreover, in this case the approximate conditional likelihood for $\Phi$, $L_C(\Phi)$ is the same as the marginal likelihood $L_M(\Phi)$, given by (5) or (6). See Bellhouse (1990) for details.

The marginal and approximate conditional likelihood in (5) or (6) can be evaluated at any $\Phi$ using state space models in the approach of Harvey and Phillips (1979). For any given $\Phi$, once the recursions to estimate $\beta$ and $\sigma^2$ are complete, the value of $s^2$ and $|\Phi|^{1/2}$ can be calculated from Harvey and Phillips (1979, eqs. 5.6 and 6.6, and 4.3 respectively). It is then necessary only to obtain $X^T\Phi^{-1}X$ and its determinant. The value of $X^T\Phi^{-1}X$ may be obtained from the final step in the recursive equations of Harvey and Phillips (1979, eq. 3.4).

## 3. SIMPLE RANDOM SAMPLING ON SUCCESSIVE OCCASIONS

### 3.1 Some General Results for Rotation Sampling

Suppose rotation group $G_r$ first appears in the sample on occasion $u$ and last appears on occasion $v$. Then $u$ is either 1 or $r$ and $v$ is either $r + c - 1$ or $k$. The total number of occasions on which a unit in $G_r$ is present in the sample is $b = v + 1 - u$. Let $\bar{y}_{u,r}, \ldots, \bar{y}_{v,r}$ be the sample means or elementary estimates for $G_r$ on occasions $u, u+1, \ldots, v-1, v$ respectively. Then under model (1), the contribution of $G_r$ to the log likelihood in (3) is

$$-\left\{bn_r \ln\sigma + (n_r/2)\ln(|\Omega_r|) + \left[n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1)\,\mathrm{tr}(\Omega_r^{-1}S_r)\right]/(2\sigma^2)\right\}, \quad (7)$$

where $x_r^T$ is the $1 \times b$ vector $(\bar{y}_{u,r} - \mu_u, \bar{y}_{u+1,r} - \mu_{u+1}, \ldots, \bar{y}_{v-1,r} - \mu_{v-1}, \bar{y}_{v,r} - \mu_v)$, where $S_r$ is the $b \times b$ matrix of sample variances and covariances for observations within the rotation group, and where $\Omega_r$ is the $b \times b$ correlation matrix on the observations on a single unit within the rotation group. Note that the parameters in $\Omega$ as given in expression (1) will also be the parameters in $\Omega_r$. The correlation matrix $\Omega$ is based on measurements from all occasions 1 through $k$; the correlation matrix $\Omega_r$ is from the subset of the data observed from occasions $u$ through $v$. By the independence assumption, the full log likelihood is obtained by summing (7) over all rotation groups.

Given the parameters in $\Omega$, or equivalently the parameters in $\Omega_1, \ldots, \Omega_{k+c-1}$, expressions for the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$, for $\mu$ and $\sigma^2$ respectively, may be found. Likewise, $V(\hat{\mu})$, the estimated variance-covariance matrix of $\hat{\mu}$ may be obtained. This is illustrated for a first-order autoregressive process in Section 3.2. Then the marginal likelihood for the parameters in $\Omega_1, \ldots, \Omega_{k+c-1}$ is given by (5) with the expressions in (5) given by

$$| \Phi |^{1/2} = \prod_{r=1}^{k+c-1} \Omega_r,$$

$$| X^T \Phi^{-1} X |^{1/2} = V(\hat{\mu}) / s^k, \tag{8}$$

$$s^2 = \sum_{r=1}^{k+c-1} \{ n_r \hat{x}_r^T \Omega_r^{-1} \hat{x}_r + (n_r - 1) \operatorname{tr}(\Omega_r^{-1} S_r) \},$$

and $p = k$, where $\hat{x}_r$ is $x_r$ with the $\mu$'s in $x_r$ replaced by their maximum likelihood estimates.

### 3.2 First-Order Autoregressive Processes

When specific forms of the correlation matrices $\Omega_1, \ldots, \Omega_{k+c-1}$ are used, some simplifications to the general form of the marginal likelihood for correlation parameters, given by (5) and (6), may be obtained. For example, assume the first-order autoregressive model

$$y_{tj} = \mu_t + \phi (y_{t-1,j} - \mu_{t-1}) + \epsilon_{tj}, \tag{9}$$

where $\epsilon_{tj} \sim N(0,\sigma^2)$ for $t = 1, \ldots, k$ and $j = 1, \ldots, N$, and where the $\epsilon$'s are mutually independent. Model (9), essentially Patterson's (1950) model, is a special case of (1). As in Section 3.1, the vector of regression parameters $\beta = (\mu_1, \ldots, \mu_k)^T$. When the data vector $y$ contains the measurements on each unit grouped by all the occasions on which it was sampled, as in the rotation sampling description of Section 3.1, the correlation matrix $\Phi$, now a function of the autoregressive parameter $\phi$, can be written as a direct sum of matrices, each of which are the correlation matrices of a first-order autoregressive process.

The following notation, similar to Patterson (1950), is used to denote various sample sizes, means and sums of squares and cross products (corrected for the appropriate mean) for occasion $t$:

$\pi_t$ = the proportion of units on occasion $t$ that are matched with units from the previous occasion $(t - 1)$;

$n_t$ = the number of units sampled on occasion $t$;

$\bar{y}_t'$ = the mean of the units on occasion $t$ that are matched with units from the previous occasion $(t - 1)$;

$\bar{y}_t''$ = the mean of the units on occasion $t$ that are unmatched with units from the previous occasion $(t - 1)$;

$\bar{y}_t$ = the mean of all the units on occasion $t$;

$\bar{x}_t'$ = the mean of the units on occasion $t$ that are matched with units from the following occasion $(t + 1)$;

$syy_t'$ = the sum of squares among units on occasion $t$ which are matched with units from the previous occasion $(t - 1)$;

$syy_t''$ = the sum of squares among units on occasion $t$ which are unmatched with units from the previous occasion $(t - 1)$;

$sxx_t'$ = the sum of squares among units on occasion $t$ which are matched with units from the following occasion $(t + 1)$;

$syy_t$ = the sum of squares among all the units on occasion $t$;

$sxy_t'$ = the sum of cross products for measurements on sample units from occasion $t$ matched with sample units from $(t - 1)$.

Under the special case of model (9), and after much algebra, it may be shown that (7) summed over all rotation groups $r$, the log-likelihood for the data reduces to

$$L(\mu_1, \ldots, \mu_k, \sigma^2, \phi) = -m \ln\sigma + (d/2)\ln(1 - \phi^2) - \{A(\mu,\phi) + B(\phi)\}/(2\sigma^2), \tag{10}$$

where $d$ is the distinct number of units sampled (irrespective of the number of occasions on which a unit is sampled) and $m$ is the total sample size $(n_1 + \ldots + n_k)$. Further in (10),

$$A(\mu,\phi) = (1 - \phi^2)n_1(\bar{y}_1 - \mu_1)^2 + \sum_{t=2}^{k} \left[\pi_t n_t \{\bar{y}_t' - \mu_t - \phi(\bar{x}_{t-1}' - \mu_{t-1})\}^2 + (1 - \pi_\tau)n_t(1 - \phi^2)(\bar{y}_t'' - \mu_t)^2\right] \tag{11}$$

and

$$B(\phi) = (1 - \phi^2)\, syy_1 + \sum_{t=2}^{k} \left\{\phi^2\, sxx_{t-1}' - 2\phi\, sxy_t' + syy_t' + (1 - \phi^2)\, syy_t''\right\}. \tag{12}$$

For any given value of $\phi$ the maximum likelihood estimator is $\hat{\mu} = G^{-1}z$ and $\hat{\sigma}^2 = \{A(\hat{\mu},\phi) + B(\phi)\}/m$, where $A(\hat{\mu},\phi)$ is (11) with $\mu$ replaced with its maximum likelihood estimate and where $G$ is a symmetric $k \times k$ band matrix of band width 3 and $z$ is a $k \times 1$ vector. The nonzero entries of $G$ are

$$g_{tt} = \pi_t n_t + (1 - \pi_t)n_t(1 - \phi^2) + \pi_{t+1}n_{t+1}\phi^2, \quad \text{for} \quad t = 1, \ldots, k$$

and

$$g_{t,t+1} = -\pi_{t+1}n_{t+1}\phi, \quad \text{for} \quad t = 1, \ldots, k - 1,$$

where $\pi_1 = \pi_{k+1} = 0$. The entries of $z$ are

$$z_t = \pi_t n_t(\bar{y}_t' - \phi\bar{x}_{t-1}') + (1 - \pi_t)n_t\bar{y}_t''(1 - \phi^2) - \pi_{t+1}n_{t+1}(\bar{y}_{t+1}' - \phi\bar{x}_t'),$$

for $t = 1, \ldots, k$, where $\pi_1 = \pi_{k+1} = 0$ and $\bar{y}_1'' = \bar{y}_1$. The vector of estimated means $\hat{\mu}$ is unbiased for $\mu$ under model (9) and its variance-covariance matrix is $\sigma^2 G^{-1}$. It follows from (5) or (6) that the marginal and approximate conditional likelihood for $\phi$ is

$$L_M(\phi) = \frac{(1 - \phi^2)^{d/2}}{\{A(\hat{\mu},\phi) + B(\phi)\}^{(m-k)/2} \mid G \mid^{1/2}}. \tag{13}$$

### 3.3 Example

The data for this example are forestry data taken from Cunia and Chevrou (1969, p. 220). The data are the merchantable volume of timber per plot measured on three occasions with partial replacement of the sample units. In rotation sampling it is assumed that once a unit

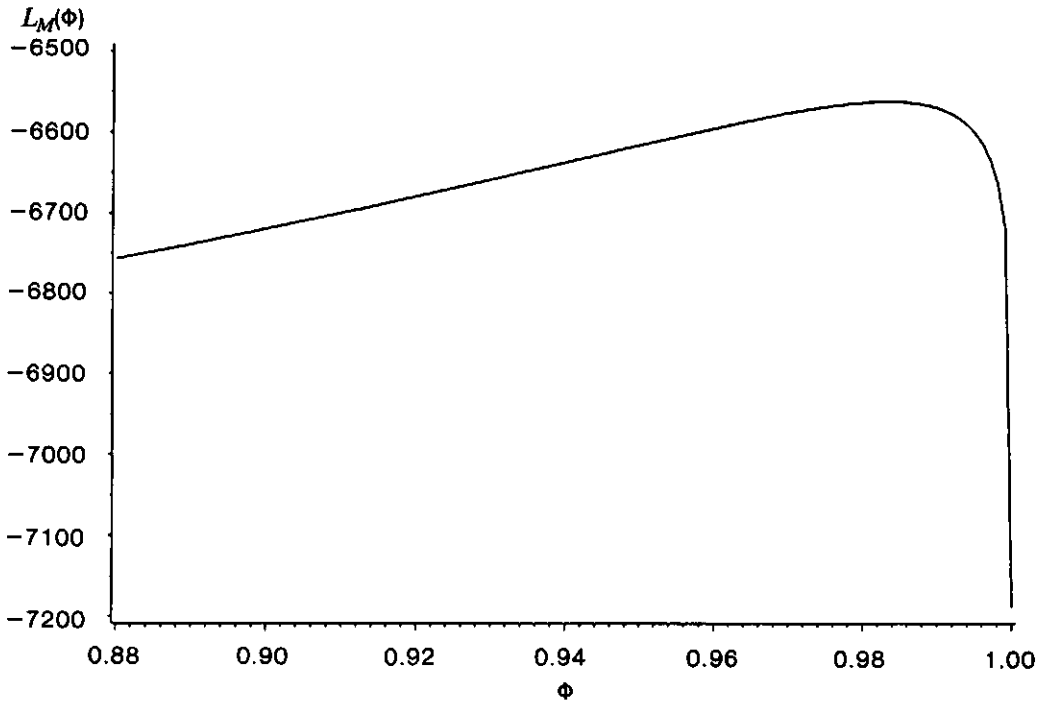



**Figure 1.** Marginal Likehood for the AR(1) Parameter

is dropped from the sample it is not selected again. In view of this assumption an adjustment to the data in Cunia and Chevrou was made. In particular, the measurements from sample units matched on the first and third occasions without matching units on the second occasion were dropped from the current example. From the remaining data the following calculations may be made:

$\pi_2 = 86/139$, $\pi_3 = 76/100$, $n_1 = 104$, $n_2 = 139$, $n_3 = 100$, $\bar{y}_2' = 161.5581$, $\bar{y}_3' = 179.9211$, $\bar{y}_1'' = 154.0673$, $\bar{y}_2'' = 167.2075$, $\bar{y}_3'' = 181.125$, $\bar{x}_1' = 147.6512$, $\bar{x}_2' = 163.4342$, $syy_2' = 864129.2$, $syy_3' = 555369.5$, $syy_1'' = 943948.5$, $syy_2'' = 266820.7$, $syy_3'' = 271762.6$, $sxx_1' = 800753.5$, $sxx_2' = 559850.7$, $sxy_2' = 812435.7$, $sxy_3' = 550943.6$, $d = 181$, and $m - k = 340$.

On substituting these data into (13) the marginal and approximate conditional likelihood of the data for the autoregressive order one parameter $\phi$ may be obtained. This is shown in Figure 1.

## 4. COMPLEX SURVEYS

There are several ways in which one may proceed to analyze time series data from complex surveys. Each method that can be put forward will depend upon the sample information that is available.

If data are available at the micro level, then variance-covariance matrices based on the complex design can be computed for the elementary estimates for each rotation group. A pseudo marginal likelihood is obtained by replacing $\bar{x}_r$ and $S_r$ in (5) and (8) by their complex survey counterparts. A similar approach is taken, for example, by Roberts, Rao and Kumar (1987) in logistic regression analysis for complex surveys: obtain a likelihood or a set of likelihood equations and replace the usual statistics by their complex survey counterparts.

Under simple random sampling, $S_r$ estimates the finite population variance-covariance matrix for measurements on the occasions covered by rotation group $r$. Consequently, in a complex design, $S_r$ is replaced by a design-consistent estimate of the corresponding finite population variance-covariance matrix. For example, Kilpatrick (1981) looked at a stratified sampling design on two occasions for evaluation of the standing volume of state forests in Northern Ireland; the strata were based on the times, beginning in the 1920's, at which the forests were planted. In order to calculate the stratified sampling equivalent to $S_r$, it is necessary to have the estimates of the means on each occasion, strata means, strata variances, and strata covariances for the unmatched and matched samples from the two occasions. For a stratified population, the finite population variance (or covariance) may be decomposed into terms comprising the variation (or covariation) between strata and the variation (or covariation) within strata; see, for example, Cochran (1977, eq. 5.32). Estimates of the means and strata means would be used to obtain a consistent estimates of the between strata variation or covariation component and estimates of the strata variances and covariances would be used to obtain estimates of the within strata variation and covariation. Unfortunately, only certain strata variance and covariance estimates were relevant to Kilpatrick's study, so that there is insufficient published data in the article to calculate a maximum marginal likelihood estimate for the correlation between timber volumes on the two occasions.

In many cases the data at the micro level will not be available. The estimation procedure then depends upon the data that are available. One scenario is considered here; others could be formulated. Suppose that only the elementary estimates and their design effects are available. Let $\bar{y}_{t,r}$ be the estimate from rotation group $G_r$ on occasion $t$ based on a sample of size $m_r$. Let $\text{deff}_{t,r}$ be the design effect associated with $\bar{y}_{t,r}$. If $\sigma^2/m_r$ is the variance of $\bar{y}_{t,r}$ under simple random sampling, then on appealing to the Central Limit Theorem,

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} \sim N(0,\sigma^2/m_r) \tag{14}$$

approximately. The modelling may proceed by assuming, within $G_r$, an ARMA-type process such as

$$(\bar{y}_{t,r} - \mu_t)/(\text{deff}_{t,r})^{1/2} = \phi\,(\bar{y}_{t-1,r} - \mu_{t-1})/(\text{deff}_{t-1,r})^{1/2} + \epsilon_t, \tag{15}$$

where $\epsilon_t$ has constant variance. This may be easily cast into the framework of model (2), where the data vector $y$ contains data of the form $\bar{y}_{t,r}/(\text{deff}_{t,r})^{1/2}$, where $\beta$ is $(\mu_1, \mu_2, \ldots, \mu_k)^T$, and where $X$ contains entries of the form $1/(\text{deff}_{t-1,r})^{1/2}$. The marginal likelihood, obtained as a special case of (5) or (6), may be evaluated using the state space models of Harvey and Phillips (1979) as noted in Section 2. Marginal and approximate conditional likelihood estimation is especially desirable under the model given by (14) and (15). The estimate of $\phi$ in this case is based on the variation between elementary estimates within each rotation group; the variation within elementary estimates is not available. The length of time a rotation group remains in the sample is short so that the problems of bias and inconsistency in the maximum likelihood estimates will be applicable here.

## 5. DISCUSSION

Binder and Dick (1990) have also suggested the use of marginal likelihood estimation techniques for sampling on successive occasions. In their framework, suppose that the survey estimates of the means, say $\bar{y}_t$, are available for each occasion $t = 1, \ldots, k$. Also, the matrix, say $S$, of variances and covariances of the surveys estimates is available. As in Binder and Dick (1989, 1990), among several others, the $\bar{y}_t$'s may be modelled by

$$\bar{y}_t = \mu_t + e_t, \tag{16}$$

where $e_t$ is the survey error at time $t$ with variance-covariance matrix estimated by $S$. The means on each occasion, $\mu_t$ for occasion $t$, follow an ARMA process. Model (16) is a special case of the random coefficients regression model, so that the appropriate marginal likelihood is different from (5).

A marginal or approximate conditional likelihood for correlation parameters in a random coefficients regression model is obtained as follows. Suppose in model (2) that $\beta$ is a random vector modelled by $\beta = W\delta + u$, where $W$ is a $p \times q$ matrix of known values, $\delta$ is a $q \times 1$ vector of parameters, and $u \sim N(0, \gamma^2\Gamma)$, independent of $\epsilon$. Under the composite model $y = XW\delta + Xu + \epsilon$, the log-likelihood for $\delta$, $\Omega$, $\Gamma$, $\gamma^2$, and $\kappa = \sigma^2/\gamma^2$, denoted by $L(\delta,\kappa,\gamma^2,\Gamma,\Omega)$, is given by (3), with $\Omega$ replaced by $\kappa\Omega + X\Gamma X^T$ and $X\beta$ replaced by $XW\delta$. Likewise, the marginal likelihood, denoted by $L_M(\kappa,\Gamma,\Omega)$, is given by (5), with $X$ replaced by $XW$ and $\Omega$ replaced by $\kappa\Omega + X\Gamma X^T$. This yields

$$L_M(\kappa,\Gamma,\Omega) = \left\{ | \kappa\Omega + X\Gamma X^T |^{1/2} \, | (XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW |^{1/2} \, g^{m-q} \right\}^{-1}, \tag{17}$$

where

$$g = y^T(\kappa\Omega + X\Gamma X^T)^{-1}y$$

$$- y^T(\kappa\Omega + X\Gamma X^T)^{-1}XW((XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW)^{-1}(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}y.$$

Now the dimension of $\Omega$ may be large in comparison to $\Gamma$; this can be the case in sampling on successive occasions. As an alternate approach, one could take the likelihood implied by (3), multiply it by the distribution for $\beta$, and integrate over $\beta$ to obtain the likelihood for the parameters under the random coefficient model. This will yield matrices of the same dimension as $\Gamma$.

Since $S$ is available, an estimate of $\Omega$, the correlation matrix of the survey error, may be easily obtained. An estimate of $\kappa = \sigma^2/\gamma^2$, may also be obtained. From assumptions which lead to the marginal likelihood in (17), it is necessary to assume that $e_t$ in (16) is a stationary random variable. Then an estimate of $\sigma^2$ is the average of the diagonal elements in $S$. If $\gamma^2$ is the variance of the $\mu$'s then the variation between $\bar{y}_t$, $t = 1, \ldots, k$ provides an estimate of $\sigma^2 + \gamma^2$. From these two estimates, an estimate of $\kappa$ may be obtained. Under model (16), $X$ in (17) is the $k \times k$ identity matrix, while $W$ is a $k \times 1$ column vector of 1's. The resulting marginal likelihood is a pseudo likelihood since some of the parameters have been replaced by estimates. In this case, the pseudo marginal likelihood for the parameters in $\Gamma$ (pseudo since $\kappa$ and $\Omega$ have been replaced by their estimates) and is given by (17) with the appropriate substitutions. The parameters in $\Gamma$ are the correlation parameters in the ARMA process on $\mu_t$. If $k$, the number of occasions, is relatively large in comparison to the number of parameters in $\Gamma$, then the marginal and approximate conditional likelihood estimates should be similar to the maximum likelihood estimator. For ease of computation, it seems that the full likelihood approach using the state space models as outlined by Binder and Dick (1989a, Section 3) appears to be the simplest approach to use in this situation.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author would like to thank the referee for his comments on an earlier draft of the paper.

## REFERENCES

BELLHOUSE, D.R. (1978). Marginal Likelihoods for distributed lag models. *Statistische Hefte*, 19, 2-14.

BELLHOUSE, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *Journal of Statistical Planning and Inference*, 21, 69-74.

BELLHOUSE, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, 77, 743-746.

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics, Volume 6 (Sampling)* (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: North-Holland, 187-211.

BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

BINDER, D.A., and DICK, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 239-253.

BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York: Wiley.

COX, D.R., and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society*, Series B, 49, 1-39.

CRUDDAS, A.M., REID, N., and COX, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika*, 76, 231-237.

CUNIA, T., and CHEVROU, R.B. (1969). Sampling with partial replacement on three or more occasions. *Forest Science*, 15, 204-224.

FRASER, D.A.S. (1967). Data transformations and the linear model. *The Annals of Mathematical Statistics*, 38, 1456-1465.

HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

KALBFLEISCH, J.D., and SPROTT, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*, Series B, 32, 175-208.

KILPATRICK, D.J. (1981). Optimum allocation in stratified sampling of forest populations on successive occasions. *Forest Science*, 27, 730-738.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

TUNNICLIFFE WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Series B, 51, 15-27.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 79–90
Statistics Canada



# Sample Rotation and Estimation in the Survey of Employment, Payroll and Hours

IOANA SCHIOPU-KRATINA and K.P. SRINATH[1]

ABSTRACT

The current Survey of Employment, Payroll and Hours, conducted by the Labour Division of Statistics Canada is a major monthly survey collecting data from a large sample of business establishments. This paper describes the methodology of the survey. The description of the stratification, sample size determination and allocation procedures is brief, whereas the description of the rotation procedure is more detailed because of its complexity. Some of the possible simplifications of the design are also highlighted.

KEY WORDS: Establishment; Response burden; Sampling frame.

## 1. INTRODUCTION

### 1.0 Objectives of the Survey

The Survey of Employment, Payroll and Hours (SEPH) is a monthly establishment based survey conducted by Statistics Canada.

The main objectives of SEPH are:

(i) to provide monthly estimates of the total number of paid employees, average weekly earnings, average hourly earnings, average weekly hours and other related variables at the industry division-province level.

(ii) to provide the above estimates for Canada at the three digit Standard Industrial Classification (SIC) level.

(iii) to provide standard errors of all the estimates produced.

It is also intended to produce estimates at the three digit SIC-province level annually.

The survey covers all industries with the exception of agriculture, fishing and trapping, private household services, religious organizations and military services. For a detailed description of the objectives and uses of SEPH, see Cottrel-Boyd *et al.* (1980).

This article describes the sample selection and rotation as well as the estimation procedure adopted for the survey. Chapter 2 presents the sample selection and rotation procedure in detail. Chapter 3 is devoted to the estimation procedure. Some of the details relating to Chapter 2 are given in the Appendix. The Appendix also presents a simplified estimator of the number of live units.

For a complete description of the SEPH methodology, see Schiopu-Kratina and Srinath (1986).

### 1.1 Preliminary Definitions

Some of the terms used in this article are defined here for convenience.

(i) Establishment – An establishment is the smallest unit that is a separate operating entity capable of reporting all elements of basic industrial statistics. The establishment is the statistical unit for SEPH. We will use the term unit for establishment.

[1] Ioana Schiopu-Kratina and K.P. Srinath, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

(ii) Employment reporting unit (ERU) – For purposes of detailed geographical statistics, the establishment is often sub-divided into reporting units based mainly on location and sometimes on other considerations like payroll, *etc.*

(iii) Standard Industrial Classification (SIC) – 1970. Each establishment is assigned a Standard Industrial Classification (SIC) code according to the nature of its activity. These SIC codes are defined in the SIC Manual (Statistics Canada 1970). For the purpose of the survey, 16 industry divisions (groups) which are groupings of specific three digits SIC's have been created.

In SEPH, the total number of paid employees associated with a unit is the characteristic chosen as a measure of size of that unit.

There are four size groups in SEPH and their boundaries are defined as follows: 0-19 for size group 1, 20-49 for size group 2, 50-199 for size group 3 and 200 or more for size group 4.

(iv) A super-stratum is defined by an industry division, province and size group. With 16 industry divisions, 12 provinces and territories and 4 size groups, there are 768 super-strata.

(v) A stratum is defined as a three digit SIC, province and size group. It is the finest level of detail for which estimates are obtained.

(vi) The take-all portion of the population consists of units which are all included in the sample with certainty. It contains units in size group 4 and pre-specified units of the population. The take-some portion of the population consists of the remaining units which are subject to sampling as described in the following sections.

## 2. SAMPLE SELECTION AND ROTATION

### 2.0 Sample Size Determination and Allocation Procedure

In SEPH, the take-some sample size is determined at the industry group – province level based on a designed coefficient of variation of the estimate of the total number of employees for that industry group-province. The required sample size and the sampling fractions are calculated at the super-stratum level, using $X$-proportional allocation, where $X$ represents the total number of employees. The sampling fractions are held constant from one month to the next. Details about the allocation procedure can be found in Schiopu-Kratina and Srinath (1986).

The actual selection is made at the stratum level. Due to the minimum sample requirements at that level, the number of sampled units is larger than the required sample size at the industry group – province level (see (2.2)).

### 2.1 Sample Selection

Let us now consider a specific stratum. Let $N$ be the size of the take-some portion of the population and $n$ the size of the take-some sample in this stratum.

Whereas the allocation of the sample to the super-strata is $X$-proportional, the allocation at the stratum level within each super-stratum is essentially proportional to the total number of units in the take-some portion of that stratum.

The sampling fraction in each stratum is given by the formula:

$$f = \max\left(f', \frac{1}{100}\right), \tag{2.1}$$

where $f'$ is calculated at the corresponding super-stratum level. In order to reduce the instability of the estimates caused by small values of the sampling fraction, it was decided to set 1/100 as a minimum sampling fraction for all strata.

The detailed calculations of the sample size at stratum level are given in section 2.3 (see the derivation of the formula (2.8)). A systematic sample is drawn from each stratum.

### 2.2 The Rotation Scheme

The sample rotation (partial periodic replacement of the sample) in SEPH is designed primarily to reduce the response burden. From previous surveys, it appeared that the average response rate in strata in which there was no rotation was significantly lower than the average response rate in strata in which there was rotation. Also, the existence of a large portion of units common to the sample for two consecutive months improves the reliability of the estimates of month-to-month change relative to the estimate of change based on two independent monthly samples. Rotation of the sample in each stratum has to be done under certain constraints such as keeping the units out of the sample for a certain period of time after they rotate out of sample.

The monthly sample consists of 14 groups numbered from 0 to 13. Group 0 contains the take-all units of the stratum. Groups 1 to 13 are called rotation groups. The labels 1 to 12 on rotation groups indicate the month in which the units other than births rotated into the sample. For example, rotation group 1 contains mostly units which entered the sample in January and births, rotation group 2 contains mostly units which entered the sample in February and births, *etc.* Rotation group 13 contains units which have completed 12 months in the sample. These units are the oldest in terms of the time spent in the sample and are eligible to be rotated out. Each month, births are selected and allocated at random to the rotation groups.

At the time of the monthly selection and rotation, all units in the reference month are transferred to rotation group 13. In February, for example, all units in the rotation group 2 are transferred to rotation group 13. A replacement group is selected from "eligible for selection" units, and newly recorded units (births). The units of the replacement group (with the exception of 11/12 of all births) are then placed in rotation group 2, and they are not eligible to rotate out for at least 12 months. If sufficient units are available for a replacement group, the contents of group 13 are removed from the sample and are not eligible for reselection for 12 months. Otherwise, some units in group 13 are retained in the sample until such time that there are enough available units outside the sample to form a replacement group. This is done in order to maintain the minimum sample size or attain a sample size large enough to provide estimates with prespecified reliability. This way, in general at least 11/12 of the units stay in the take-some portion of the sample for two consecutive months.

The units that have left the sample are assigned to a waiting group which is divided into subgroups. A subgroup consists of units which were all removed from the sample in the same month. The waiting group contains 12 subgroups in every stratum. The time each unit has spent outside the sample is thus recorded to ensure that the units will not be reselected for at least 12 months. The units that have spent the required amount of time in the "not eligible for selection" group are transferred to the "eligible for selection" group and are thus assigned a positive probability of reselection.

To summarize, the entire take-some population at any given time consists of four groups of units. These are:

(i) units that are in the sample for that month;

(ii) units that are eligible for selection (E.F.S.);

(iii) units in the waiting group which have rotated out of the sample less than 12 months ago and which are not eligible for selection (N.E.F.S.);

(iv) births, *i.e.* units that have not been previously recorded on the frame.

The process of monthly selection and rotation involves an exchange of units among these groups. Some units leave group (i) for group (iii) and new units enter group (i) from group (ii), after some selected births from group (iv) have been transferred to group (ii). The remainder of the births are allocated to groups (ii) and (iii) after selection. This is done in order to insure that the sample is representative of the population in any given month.

## 2.3 Determination of the Sample Size and Weights

### 2.3.1 Monthly Updates

The sampling frame contains a large number of units which are inactive, out of business, out of scope *etc.* Apart from the burden of retaining an increasing number of inactive units on the frame, the estimators based on samples drawn from such a population are likely to have a high variance, due to the fact that the sample contains a high proportion of zero observations. Ideally, all such units should be eliminated from the sampling frame before the monthly sample is drawn. The frame is updated each month, after a monthly selection and rotation and prior to the next. For this reason, the indices we use to denote births and deaths on the frame are one unit higher than those used for the sample size in the sample selection preceeding the update. For example, after the initial sample selection, say $n(0)$ units are in the sample, of which $d(1)$ units are subsequently found to be dead units. Then $D(1)$ denotes the number of dead units in the out-of-sample portion of the population and $B(1)$ the number of units registered as births that month. In calculating the required sample size for the following month $n(1)$, one must take into account these updates (see (2.3)) as well as the size of the population at the time of the first sample selection $N(0)$.

### 2.3.2 Determination of Sample Size

The population of a given take-some stratum is a function of time and it will be denoted by $N(t)$ say, where $t$ is a positive integer which increases by one unit from one month to the next. The required sample size is, for each month:

$$n'(t) = [fN(t) + 0.5]. \tag{2.2}$$

Here $[a]$ is the largest integer number which is not greater than $a$. The constant 0.5 is used for a better approximation in the rounding off procedure.

Suppose $d(t)$ units are eliminated from the sample (in-sample deaths) and $D(t)$ from the rest of the population of the stratum (out-of-sample deaths). Also let $B(t)$ new units be recorded during the same time interval (births).

As a result, the size of the population of the cell at the time of the $t^{\text{th}}$ selection is:

$$N(t) = N(t-1) - d(t) - D(t) + B(t). \tag{2.3}$$

Since the updates are not exhaustive, undetected inactive (dead) units are expected to exist in the population.

Let $n_t(t)$ be the number of live units left in the previous month sample (after the updates) *i.e.*:

$$n_t(t) = n(t-1) - d(t). \tag{2.4}$$

We assume that there are no undetected dead units in the sample at this point. We can think of the population of a stratum as consisting of two domains: the domain of live units and the domain of dead units. The size of the dead domain is not known, but an estimate $\hat{U}_d(t)$ can be calculated based on the information given by the sample and the updates (see Appendix). Let $\hat{U}_d(t)$ be an estimate of the number of undetected dead units in the population at the time of the $t^{\text{th}}$ monthly selection. Then:

$$N(t) = \hat{U}_\ell(t) + \hat{U}_d(t), \tag{2.5}$$

where $\hat{U}_\ell(t)$ is the estimate of the number of live units.

The probability of choosing a dead unit when selecting a unit at random from the out-of-sample units is:

$$\hat{P}_d(t) = \min\left\{\frac{\hat{U}_d(t)}{N(t) - n_\ell(t)},\ 1\right\}. \tag{2.6}$$

The required number of live units in the sample is:

$$n'_\ell(t) = f\hat{U}_\ell(t). \tag{2.7}$$

The replacement sample size is calculated in such a way as to ensure that the expected number of live units in the sample after selection is $n'_\ell(t)$.

Now assume that at the time of the $t^{\text{th}}$ sample selection and rotation, of the $n_\ell(t)$ live units in the sample, $n_o(t)$ units are eligible to rotate out.

Since there are $n_\ell(t) - n_o(t)$ live units left in the sample, $n'_\ell(t) - n_\ell(t) + n_o(t)$ more live units are required in the sample for the $t^{\text{th}}$ month.

In order to represent the births in the sample adequately, $b(t) = fB(t)$ births should be selected at random and included in the sample.

Therefore:

$$\ell(t) = \max(n'_\ell(t) - n_\ell(t) + n_o(t) - b(t),\ 0),$$

live units should be selected from the eligible for selection group and added to the sample along with the selected births. Taking into account the existence of an unknown number of inactive units in the population and integerizing, it is required that:

$$n_i(t) = \min\left(\left|\frac{\ell(t)}{1 - \hat{P}_d(t)} + 0.5\right|,\ n''(t)\right),$$

more units rotate into the same sample, with $\hat{P}_d(t)$ given by (2.6) and $n''(t) = N(t) - n_o(t)$.

In calculating $n_i(t)$, we made the assumption that there are no inactive units among the births, so the expansion factor $[1 - \hat{P}_d(t)]^{-1}$ is applied only to the "older" units in the E.F.S. group.

The sample size $n(t)$ for the $t^{\text{th}}$ month is:

$$n(t) = \max\{n_\ell(t) - n_o(t) + n_i(t) + b(t),\ m\}. \tag{2.8}$$

In (2.8), $m$ represents the minimum required sample size for a stratum, which is presently set at 3. This additional requirement increases the sample size by 3,000 units in all strata, of which 1,800 are expected to be in the sample for a considerable length of time.

Of the $n_i(t)$ units which rotate in, $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$ are expected to be found inactive and $\hat{n}_t(t) = n_i(t) - \hat{n}_d(t)$ active. Thus, of the $n(t)$ units in the sample after the $t^{\text{th}}$ monthly selection and rotation, $\hat{n}_t(t) = n_t(t) - n_o(t) + \hat{n}_t(t) + b(t)$ are expected to be alive and they represent the $\hat{U}_t(t)$ units of the live domain at the proper rate $f$ (see (2.7) - (2.8)), when $n(t) > m$ in (2.8).

### 2.3.3 Determination of Weights

The weight $w(t)$ used for estimation for the $t^{\text{th}}$ month is expressed in terms of the size of the population and sample at time $t$. However, the use of $N(t)/n(t)$ as weight for estimation could lead to an overestimation of the live units in the population. Indeed, $n(t)$ in formula (2.8) was chosen so that the expected number of live units in the sample equals the required sample size. The number of dead units in the sample drawn as described above may not represent the size of the dead domain at the proper rate. In (2.8), $n_i$ is drawn from the general population and is thus expected to preserve the proportion between the dead and the live domain. No deaths are expected to be found among births and thus $b(t)$ properly represents the birth subgroup of the population. There are, however, $n_t(t) - n_o(t)$ units left in the sample from a previous selection, after rotation and the updates on the sample. The proportion of deaths among them is likely to be much smaller than the corresponding proportion in the general population, in spite of the fact that the updates are based on information from sources other than the survey. Then the value of $N(t)/n(t)$ should be adjusted for the underrepresentation of dead units in the sample. This gives, when $n(t) > m$,

$$\frac{N(t)}{n(t) + \hat{u}(t)} = \frac{1}{f}, \tag{2.9}$$

where $\hat{u}(t)$ will be determined subsequently (see (2.10)). The value of $\hat{u}(t)$ represents the "deaths" that have to be added to the sample to correctly represent the dead units in the population. Notice that when the first sample is drawn or if a redraw takes place, such an adjustment is not needed, *i.e.* $\hat{u}(0) = 0$.

In order to find a formula for $\hat{u}(t)$, we use (2.5) in the numerator of (2.9) and (2.8) in the denominator. By (2.7) - (2.8), we must also have:

$$\frac{\hat{U}'_d(t)}{\hat{n}_d(t) + \hat{u}(t)} = \frac{1}{f}.$$

With $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$, we obtain from above

$$\hat{U}'_d(t) = \frac{1}{f}[\hat{n}_d(t) + \hat{u}(t)] \quad \text{or} \quad \hat{u}(t) = f\hat{U}'_d(t) - \hat{n}_d(t). \tag{2.10}$$

The death adjustment is given by:

$$v(t) = \begin{cases} \hat{u}(t) & \text{if } \hat{u}(t) \geq -\hat{n}_d(t) \\ 0 & \text{if } \hat{u}(t) < -\hat{n}_d(t) \end{cases} \tag{2.11}$$

and the weight used in estimation is:

$$w(t) = \frac{N(t)}{n(t) + \hat{v}(t)}. \tag{2.12}$$

Note that the weight in (2.12) is defined using an estimate and so it is a random variable.

The use of the weight defined by (2.12) implies that the estimate of the number of live units in the population, defined by $\hat{U}_t(t) = w(t)\hat{n}_t(t)$ does not exceed $N(t)$, the size of the population at the time of the $t^{\text{th}}$ sample selection.

Let us define:

$$\hat{U}_d(t) = w(t)\,[\hat{v}(t) + \hat{n}_d(t)].$$

By (2.10) - (2.11), it follows that $\hat{U}_d(t) \geq 0$ and its minimum value is 0 when $\hat{v}(t) - \hat{n}_d(t) = 0$. By (2.5), the maximum value of $\hat{U}_t(t)$ is then $N(t)$.

The restriction that the estimator of live units be truncated at $N(t)$ has implications on estimation which will be discussed in section 3.1. The estimator $\hat{U}_d(t)$ is calculated recursively (see the Appendix) using 2.10 - 2.11 and the fact that $\hat{v}(0) = \hat{u}(0) = 0$.

It has to be noted that the formula (2.11) is slightly different from the formula giving the death adjustment in SEPH. Firstly, for the sake of simplicity, we did not consider here the cases when the minimum sample size $m$ has to be used. In such cases, the use of the sampling fraction $f$ in (2.10) is not appropriate. In SEPH, the previous month weight is used in (2.10) in lieu of $f$ in all instances. Secondly, the death adjustment in SEPH is always taken to be positive. Formula (2.11) shows that it could be negative, as long as it is larger than $-\hat{n}_d(t)$. The actual instances in which this happens, or, more generally, when $\hat{u} \leq 0$ are very rare.

The Appendix presents a formula for the estimator of the live units which does not require the use of the death adjustment.

### 2.4 Sampling of Births

As mentioned previously, every month new units are added to the frame. Since it is believed that these new units (births) may differ from the "old" units, a special birth strategy was designed, aimed at adequately representing the births in the sample.

Ideally, if $B$ births are added during the current month and if $f$ is the stratum sampling fraction, then $b = fB$ births should be selected in the sample during that month. The selected births are randomly assigned to the rotation groups described in section 2.2. This ensures the same probability of rotating out of the sample for births as for the "old" units, so that the age distribution of the in-sample units is the same as the out-of-sample units.

Using the notation of the previous section, let $n_i$ be the number of units required to rotate in at the time of the monthly sample selection excluding births and $N'$ the number of units in the E.F.S. group (group (ii) of section 2.2). The birth strategy consists of a two-phase selection procedure. This procedure involves the formation of a common pool of births and "older" units from which a sample is then drawn. This was thought necessary because usually, the birth group is too small for sampling births separately each time. There are two ways of forming the common pool depending on the sizes of the birth group and the E.F.S. group. If:

$$\frac{n_i}{N'} > f, \tag{2.13}$$

then $b'$ births are preselected from the birth group and a common pool of size $(N' + b')$ is formed where:

$$b' = \frac{b\,N'}{n_i}. \tag{2.14}$$

Inequality (2.13) ensures that $b' \leq B$ which means that the birth group is large enough and the preselection can take place. From the common pool of $N' + b'$ units, $n_i + b$ units are selected next and added to the sample.

The choice of $b'$ as given by (2.14) ensures that the expected number of births in the sample is the desired one, since the probability of selecting one birth from the pool is $b'/(b' + N')$ and therefore the expected number of births when $n_i + b$ units are selected without replacement from this pool is $(n_i + b)\,b'/(b' + N')$ which by (2.14) equals $b$. Similarly, it is easy to see that the expected number of "older" units is $n_i$.

In the complementary situation, when (2.13) does not hold, a common pool of size $(n' + B)$ units is formed where:

$$n' = n_i/f \leq N'. \tag{2.15}$$

Then $n'$ "older" units are selected from the E.F.S. group. Let us note that in this situation $b'$ as given by (2.14) is larger than $B$ and so the first procedure cannot be applied.

We now calculate the expected number of "old" units in the sample. Since the probability of selecting one "old" unit is now $n'/(n' + B)$ and $n_i + b$ units are drawn from the pool of $n' + B$ units and placed in the sample, the expected number of "old" units is $(n_i + b)\,n'/(n' + B)$ which equals $n_i$. The expected number of births is $b$.

It has to be noticed that an underrepresentation of births may occur in some situations. For example, if in some stratum no units are required to rotate in, then for the month in question the births will not be represented in the sample taken from that stratum. However this situation usually arises when the population size is small and in the long run, the representation of births can be expected to average out correctly.

The $b$ births actually selected are randomly assigned to the rotation groups 1-12 in the sample, resulting, on the average, in $b/12$ births assigned to each of these rotation groups. This ensures that the probability of a birth rotating out is the same as that of any other unit.

In order to keep the age distribution of the units in the groups (i) - (iii) (see section 2.1) constant, the non-selected births will be allocated to the E.F.S. and the N.E.F.S. group at random. That is, if $N'$ is the number of units in the E.F.S. group and $N''$ is the number of units in the N.E.F.S. group, then $N'(B - b)/(N' + N'')$ unselected births will be assigned to the E.F.S. group and $N''(B - b)/(N' + N'')$ to the N.E.F.S. group.

## 3. ESTIMATION

### 3.0 Introduction

In this chapter we describe the procedure for estimating the characteristic "the total number of paid employees".

As indicated in section 2.0, only the reliability of the estimates of the total employment at the industry group-province level of aggregation is prespecified. The estimates of characteristics other than the total employment have varying degrees of reliability. For example, estimates of average weekly earnings are expected to have higher reliability than the total employment.

The finest level of aggregation at which the estimates are published is the SIC-province level, but the basic "building blocks" for producing the estimates are the strata.

An outlier in SEPH is an observation in the take-some portion of the sample which is larger than a prespecified value.

The weight of each stratum is first calculated as in section 2.3, then it is adjusted for outliers in that stratum. Estimates of the total employment are computed for each stratum using the adjusted weights. There is no adjustment for nonresponse, as values for the nonrespondents are imputed. The estimate of the total employment for each stratum is obtained by adding the following totals:

(i) the total employment for take-all units

(ii) the total employment for outlier units

(iii) the sum of the weighted values of employment for take-some units, excluding outliers.

Since the weight assigned to each outlier is one, outliers are treated as take-all units for the purpose of estimation and are therefore not used in the variance estimation.

### 3.1 Estimation of the Total of a Characteristic

Let us consider a specific stratum for a given month of the survey. Let $N$ be the size of the take-some portion of the population of the stratum and $n$ the number of take-some units in the sample for that month. If $\hat{v}$ is the death adjustment (see (2.11)), then the original weight assigned to each unit in the sample for the purpose of estimation is (see (2.12)):

$$w = \frac{N}{n + \hat{v}}. \tag{3.1}$$

However, if $t$ outliers are present in the sample with $t \geq 1$ then this weight is modified by giving each outlier a weight of 1 and assigning to the remaining units in the take-some portion of the sample the weight:

$$w' = \frac{N - t}{n + \hat{v} - t}. \tag{3.2}$$

Let $S$ represent the set of in-sample units. If $Y(u)$ represents the value of employment corresponding to the unit $u$ in the sample, then the estimate of the total employment in the stratum is:

$$\hat{Y} = \sum_{u \in S} w(u) Y(u). \tag{3.3}$$

Where $w(u) = 1$ if $u$ is an outlier or a take-all unit and $w(u) = w'$ for all other units in the sample (see (3.2)).

Estimates of totals at any level of aggregation higher than the stratum are obtained adding the estimates of the stratum totals, for all strata in the level of aggregation considered.

Since for the purpose of estimation the outliers may be considered take-all units, we may replace, for the sake of simplicity $w'$ by $w$, with $w$ given by (3.1). Then $N$ and $n$ should be modified accordingly (see (3.2)).

Let $N_t$ be the size of the live domain in the population of the stratum and $n_t$ the number of live units in the sample. Let $\bar{Y}_t$ represent the average employment of the live units in the sample.

Since only the live units contribute to the total employment, an estimate of the cell total is:

$$\hat{Y}_\ell = \hat{U}'_\ell \bar{Y}_\ell. \tag{3.4}$$

We consider the case $fN > m$.

In (3.4), $\hat{U}'_\ell$ is an estimate of $N_\ell$, the number of live units in the population and is given by:

$$\hat{U}'_\ell = \frac{1}{f} n_\ell. \tag{3.5}$$

In (3.5), $f$ is the stratum sampling fraction which is held fixed. The estimator based on the values of $\hat{U}'_\ell$ is unbiased, that is:

$$E(\hat{U}'_\ell) = N_\ell. \tag{3.6}$$

As a consequence of the unbiasedness of the estimator based on (3.5), $\hat{U}'_\ell$ may exceed $N$ in some instances, as low possible outcomes of $\hat{U}'_\ell$ compensate for it.

In SEPH, the estimate of live units $\hat{U}_\ell$ is defined by:

$$\hat{U}_\ell = w n_\ell \tag{3.7}$$

with the weight given by (3.1).

The definition of the weight in SEPH (see the end of section 2.3) implies that:

$$\hat{U}_\ell = \begin{cases} \hat{U}'_\ell & \text{if} \quad n_\ell \leq fN \\ N & \text{if} \quad n_\ell > fN. \end{cases} \tag{3.8}$$

The estimate of the total employment in SEPH is defined by:

$$\hat{Y}_\ell = \hat{U}_\ell \bar{Y}_\ell. \tag{3.9}$$

An estimate of the total employment based on (3.5) is:

$$\hat{Y}'_\ell = \hat{U}'_\ell \bar{Y}_\ell. \tag{3.10}$$

The estimator based on (3.8) is biased and consequently the estimator of the total employment in SEPH is also biased.

However, the mean square error of the estimator based on (3.9) conditioned on $n_\ell$ is smaller than the mean square error of the estimator based on (3.10), conditioned on $n_\ell$. We now sketch a proof of this claim.

It is not difficult to see that, for each particular outcome, the bias $B_\ell$ of the estimator based on (3.9) and conditioned on $n_\ell$ is given by:

$$B_\ell = (\hat{U}_\ell - N_\ell) \bar{Y}_\ell. \tag{3.11}$$

Similarly, we obtain for the estimator based on (3.10):

$$B'_\ell = (\hat{U}'_\ell - N_\ell) \bar{Y}_\ell. \tag{3.12}$$

We show that the conditional mean square error of the estimator (3.9) is smaller than the conditional mean square error of the estimator (3.10). The same result then holds for the unconditional mean square errors. We condition on the realized sample size of live units $n_\ell$. From (3.8) - (3.10), $\text{Var}[\hat{U}'_\ell \bar{Y}_\ell \mid n_\ell] - \text{Var}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] = [n_\ell^2 f^{-2} - N^2]\, 1\{n_\ell > fN\}\text{Var}[\bar{Y}_\ell \mid n_\ell]$. Notice that $n_\ell^2 f^{-2} - N^2 > 0$ on the set $\{n_\ell > fN\}$. We now compare $[B'_\ell]^2$ and $B_\ell^2$:

$$[B'_\ell]^2 - B_\ell^2 = 1\{n_\ell f^{-1} > N\}\{[\hat{U}'_\ell - \mathrm{N}_\ell]^2 \bar{Y}_\ell^2 - (N - N_\ell)^2 \bar{Y}_\ell^2\}.$$

But $\hat{U}'_\ell - N_\ell = f^{-1} n_\ell - N_\ell > N - N_\ell$ if $n_\ell f^{-1} > N$.

Therefore $[B'_\ell]^2 - B_\ell^2 \geq 0$. Since $\text{MSE}[\hat{U}'_\ell \bar{Y}_\ell \mid n_\ell] = \text{Var}[\hat{U}'_\ell \bar{Y}_\ell \mid n_\ell] + [B'_\ell]^2$ and $\text{MSE}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] = \text{Var}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell] + [B_\ell]^2$, the term-by-term comparison leads to the conclusion that:

$$\text{MSE}[\hat{U}'_\ell \bar{Y}_\ell \mid n_\ell] \geq \text{MSE}[\hat{U}_\ell \bar{Y}_\ell \mid n_\ell].$$

This important property motivates the choice of the estimator (3.9) over (3.10) for the total employment in SEPH.

## APPENDIX

In this Appendix, we use the notation of section 2.3.2.

We first derive the formula for $\hat{U}_d(t)$, the size of the dead domain used in SEPH for the $t^{\text{th}}$ selection.

Recall $\hat{U}_d(t) = w(t)[\hat{v}(t) + \hat{n}_d(t)]$. At the time of the $t^{\text{th}}$ update, $d(t+1)$ dead units are found in the sample. We can replace therefore $\hat{n}_d(t)$, the estimated number of dead units in the sample by $d(t+1)$ in order to obtain an update of $\hat{U}_d(t)$, namely $\hat{N}_d(t+1)$:

$$\hat{N}_d(t+1) = w(t)[\hat{v}(t) + d(t+1)]. \tag{1.1}$$

Formula (1.1) uses the death adjustment from the previous month. The initial value of $\hat{v}$ is $\hat{v}(0) = 0$ (see the remark after (2.9) and the definition of $\hat{u}$) and $w(0) = f^{-1}$. Now the estimate of the size of the dead domain for the $(t+1)^{\text{th}}$ monthly selection is:

$$\hat{U}_d(t+1) = \max(\hat{N}_d(t+1) - D(t+1) - d(t+1), 0). \tag{1.2}$$

Notice that $\hat{U}_\ell(t+1)$ can be calculated from (2.5) when $\hat{U}_d(t+1)$ is known and vice versa. An alternative form for $\hat{U}_\ell(t+1)$ is obtained recursively as follows. Let us assume that $\hat{U}_\ell(t)$ is known before the $t+1^{\text{th}}$ selection of th sample (recall that $t = 0$ is used for the first, or original selection). Then $\hat{U}_d(t)$ is also known and can be used to calculate $\hat{P}_d(t)$, the probability of selecting a dead unit from the out-of-sample units (see formula 2.6). This probability is then used to calculate the required number of units which should rotate in as described in 2.3.2, as well as the expected number of live units in the sample at the time of the $(t+1)^{\text{th}}$ selection, $\tilde{n}_\ell(t)$.

Then the weight used in estimation for the next selection is $\hat{U}_\ell(t)/\tilde{n}_\ell$. After selection, the $(t+1)^{\text{th}}$ update takes place and the actual number of live units in the sample is found to be $n_\ell(t+1)$. The estimate of the size of the live domain for the following selection can be calculated

$$\hat{U}_\ell(t+1) = \min\left\{\frac{\hat{U}_\ell(t)}{\tilde{n}_\ell(t)} n_\ell(t+1) + B(t+1), N(t+1)\right\}$$

and so forth. To initiate the process, note that the weight used in the first estimation is $w(0) = f^{-1}$ and after the first update,

$$\hat{U}_\ell(1) = \min\{w(0) \times n_\ell(1) + B(1), N(1)\}.$$

## REFERENCES

COTTREL-BOYD, T.M., DUNN, M.R., HUNTER, G.E., and SRINATH, K.P. (1980). Development of the redesign of the Canadian establishment based employment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 8-15.

SCHIOPU-KRATINA, I., and SRINATH, K.P. (1986). The methodology of the Survey of Employment, Payroll and Hours. Working Paper No. BSMD-86-010E, Statistics Canada.

STATISTICS CANADA (1970). *Standard Industrial Classification Manual.* Catalogue 12-501, Ottawa: Statistics Canada.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 91–98
Statistics Canada



# Estimating a System of Linear Equations with Survey Data

PHILLIP S. KOTT[1]

ABSTRACT

This paper develops a framework for estimating a system of linear equations with survey data. Pure design-based sample survey theory makes little sense in this context, but some of the techniques developed under this theory can be incorporated into robust model-based estimation strategies. Variance estimators with the form of the single equation "linearization" estimator are nearly unbiased under many complex error structures. Moreover, the inclusion of sampling weights in regression estimation can protect against the possibility of missing regressors. In some situations, however, the existence of missing regressors can make the estimation of a system of equations ambiguous.

KEY WORDS: Sampling weights; Putative missing regressor; Robust; Nearly unbiased.

## 1. INTRODUCTION

Kott (1991) showed that design-based techniques developed for estimating a single linear regression equation could be exploited in a more conventional model-based framework. In particular the use of sample weighted regression was shown to help protect against the possible existence of missing regressors, while the so-called linearization variance estimator was shown to produce nearly unbiased estimators of mean squared error for many complex variance structures.

This paper extends those results to the estimation of a system or "grouping" of linear equations, a topic of considerable interest to econometricians (see, for example, Johnston 1972, pp. 238-241). Two simple examples may shed some light onto the subject for those not already schooled in econometric methods or their equivalent.

Suppose we have a sample of farmers and want to estimate the relationship between the amount of planted soybean acres and the size of the farm. Zellner (1962) showed in effect that even if a simple quadratic equation with independent and identically distributed errors correctly described the universe, a better estimator than the one produced by ordinary least squares (OLS) might exist. This estimator could be found by taking into consideration other linear relationships, say between planted *corn* acres and farm size, that had errors terms correlated with those in the original relationship. Zellner called the system-wide estimation of a group of such equations "seemingly unrelated regression." Oddly, in order for Zellner's generalized least squares (GLS) estimator to produce different results from OLS, it is necessary for some equations to contain regressors not found in other equations. Alternatively, one can think of each equation as containing the same regressors but with certain coefficients constrained to zero.

A second example concerns a sample of firms each producing one output, $y$, from two inputs, $x_1$ and $x_2$, with unit prices, $p_1$ and $p_2$. Economists often assume that each firm possesses the same technology (plus or minus an error term). Given $p_1$, $p_2$, and $y$, each firm would choose $x_1$ and $x_2$ so as to minimize total cost, $c = p_1x_1 + p_2x_2$. Suppose that the

[1] Phillip S. Kott, Special Assistant for Economic Survey Methods, U.S. Bureau of the Census, Room 3061-3, Washington, DC, 20233, USA.

relation between $p_1$, $p_2$, $y$ and the cost minimizing $c$ can be expressed by the following equation (on average):

$$\log(c) = b_0 + b_1\log(p_1) + b_2\log(p_2) + b_3\log(y). \quad (1)$$

Economic theory tells us that a rational firm faced with implicit cost equation (1) would choose its level of $x_1$ so that

$$x_1p_1/c = b_1. \quad (2)$$

Naturally, in order to estimate equations (1) and (2), we need to add a stochastic structure. For simplicity, assume that both equations (1) and (2) fit the behavior of all firms subject to respective independent (across firms) and identically distributed random errors. Observe that in addition to the strong possibility that the error terms in the two equation will be correlated for a particular firm, there is also a coefficient ($b_1$) shared by both equations.

When faced with a system of linear equations in which the coefficients are known to be constrained, the design-based approaches to linear regression reviewed in Kott (1990a) make little sense. For that reason, although design-based *practice* inspires many of the procedures discussed here, only the extended model-based approach introduced in Kott (1991) will be used to justify them.

Section 2 lays out the theoretical model for the estimation of a system of linear equations based on data from the full population. Section 3 introduces the sample weighted analogues of full population OLS and GLS estimators for a system of linear equations. Section 4 addresses robust mean squared error estimation of both the sample weighted OLS and GLS estimators employing a straigtforward generalization of the linearization variance estimator (see, for example, Shah, Holt, and Folsom 1977). Section 5 discusses a general method for developing test statistics that can be used to evaluate, among other things, whether sample weighted OLS and GLS are actually estimating the same thing. Section 6 explores a simple example. Section 7 sketches an extension of the methodology developed here to what econometricians call "simultaneous equations." In the stochastic version of equation (1), for example, many economists believe that $\log(y)$ should be treated as a random variable and that $\log(c)$ can be assumed to be fixed. This causes a simultaneity bias if not specifically addressed by techniques like two and three stage least squares (see Johnston 1972, pp. 341-420). Finally, section 8 contains a brief discussion.

## 2. FULL POPULATION ESTIMATION

### 2.1 The Unconstrained System:

Suppose we have a population containing $M$ data points. Each data point $i$ is associated with $G + \tilde{K}$ observed variables satisfying the following model:

$$Y = \tilde{X}\tilde{\beta} + U + V, \quad (3)$$

where $Y$ is an $M \times G$ matrix of observed dependent variables (the $i$th row of $Y$ contains the dependent variables associated with the $i$th data point),

$\tilde{X}$ is an $M \times \tilde{K}$ matrix of observed independent or regressor variables (the $i$th row of $\tilde{X}$ contains the independent variables associated with the $i$th data point),

$\tilde{\beta}$ is an $\tilde{K} \times G$ matrix of parameters,

$U$ is an $M \times G$ matrix satisfying the relationship $\lim_{M-\infty} \tilde{X}'U/M = 0_{\tilde{K} \times G}$ (a matrix of zeroes) - this assumes that there is an underlying process generating the data points which could in principle generate points *ad infinitum* (see Kott 1991), and

$V$ is a $M \times G$ matrix of random variables such that $E(V) = 0_{M \times G}$ and $E(v_{is}v_{it}) = \sigma_{st(i)}$.

It is well known that if $U \equiv 0_{M \times G}$, $E(v_{is}v_{jt}) = 0$ for $i \neq j$, and $\sigma_{st(i)} = \sigma_{st}$ for all $i$, then

$$\tilde{B}_{\text{OLS}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y \tag{4}$$

is the best linear unbiased estimator for $\tilde{\beta}$ (see, for example, Johnston 1972, p. 240). This means that the $g$th column of $\tilde{B}_{\text{OLS}}$, call it $B_{.g}$, is the best linear unbiased estimator of $\beta_{.g}$, where

$$y_{.g} = \tilde{X}\beta_{.g} + u_{.g} + v_{.g}, \tag{5}$$

and $y_{.g}$, $u_{.g}$, and $v_{.g}$ are the $g$th columns of $Y$, $U$, and $V$, respectively. Equation (5) can be viewed as the $g$th equation in the system of equations represented by equation (3).

Let us call the matrix $U$ in equation (3) the *putative missing regressor* matrix. Usually, in conventional (*i.e.*, model-based) regression analysis that part of the dependent variable (or variables) not capturable by a linear combination of the independent variables is (are) assumed to be purely random. Here, however, we follow Kott (1991) and allow for the possiblility of non-random missing regressors. Note that even when $U \neq 0_{M \times G}$, $\tilde{B}_{\text{OLS}}$ remains nearly (*i.e.*, asymptotically) unbiased.

### 2.2 The (Possibly) Constrained System:

Efficient estimation is a more complicated matter when there are constraints on some elements of $\tilde{\beta}$; for example, when $\tilde{\beta}_{kg}$ is known to be zero or when $\tilde{\beta}_{hf}$ is known to equal $\tilde{\beta}_{hj}$.

In this paper, we are interested in a (possibly) constrained systems of equations that can be modelled directly with the following equation:

$$y = X\beta + u + v, \tag{6}$$

where $y = (y_{.1}', y_{.2}', \ldots, y_{.G}')'$, $u$ and $v$ are defined in an analogous manner, $X$ is an $MG \times K$ matrix, $\beta$ is a $K \times 1$ vector, and $K \leq G\tilde{K}$. By definition, $\lim_{M-\infty} X'u/M = 0_K$.

When the original $\tilde{\beta}$ in equation (3) is unconstrained, $K = G\tilde{K}$, and

$$X = \begin{bmatrix} \tilde{X} & & & \\ & \tilde{X} & & \\ & & \ddots & \\ & & & \tilde{X} \end{bmatrix}.$$

When the original $\tilde{\beta}$ is constrained, however, $K < G\tilde{K}$. For example, when an element of $\tilde{\beta}$ is known to be zero, it can be removed from the $\beta$ vector in equation (6) along with the column of the $X$ matrix that corresponds to it. When two elements in the same row of $\tilde{\beta}$ are known to be equal, the second can be removed from $\beta$, and $X$ can be adjusted accordingly (it will no longer be block diagonal).

When $u \equiv 0_{MG}$ and $\text{Var}(v) = \Sigma \otimes I_M$ (where $\Sigma = \{\sigma_{st}\}$), then $b_{\text{OLS}} = (X'X)^{-1}X'y$ is an unbiased estimator for $\beta$, but $b_{\text{GLS}} = (X'[\Sigma^{-1} \otimes I_M]X)^{-1}X'[\Sigma^{-1} \otimes I_M]y$, where $I_M$ is the $M \times M$ identity matrix, is the best linear unbiased estimator. In practice, the elements of $\Sigma$ have to be estimated from the sample, say by $\hat{\sigma}_{gf} = r_{\cdot g}'r_{\cdot f}/M$, where $r_{\cdot g} = y_{\cdot g} - X_{\cdot g}b_{\text{OLS}}$.

It is well known that $b_{\text{OLS}}$ and $b_{\text{GLS}}$ are equal when the parameter matrix in (3) is unconstrained (again, see Johnston 1972, p. 240). Turning to the constrained case, if $u \neq 0_{\text{MG}}$, then both $b_{\text{OLS}}$ and $b_{\text{GLS}}$ are nearly unbiased estimators of $\beta$ when $\lim_{M\to\infty}\bar{X}'U/M = 0_{\bar{K}\times G}$ holds as we originally assumed. Unfortunately, $b_{\text{GLS}}$ may not be nearly unbiased under the weaker assumption that $\lim_{M\to\infty}X'u/M = 0_K$, which is more in line with the extended model in Kott (1991) when (6) is viewed as a single equation.

To see why this is, let $X_{\cdot g}$ denote the $M \times K$ matrix formed from the $\{(g-1)M+1\}$th through the $\{gM\}$th row of $X$ and $\Sigma^{-1} = \{\sigma^{fg}\}$, then

$$E(b_{\text{GLS}} - \beta) \propto X'[\Sigma^{-1} \otimes I_M]u/M =$$

$$\sum_g X_{\cdot g}'\left(\sum_f \sigma^{fg}u_{\cdot f}\right)\Big/M = \sum_g \sum_f \sigma^{fg}X_{\cdot g}'u_{\cdot f}\Big/M,$$

which approaches zero as $M$ grows large under the stronger assumption but not necessarly the weaker one.

## 3. ESTIMATION WITH SURVEY DATA

Suppose now that we observe variables values for only a random sample of the population. Let $P = \text{diag}\{p_i\}$, where $p_i$ is the probability of selection for data point $i$. Let $S = \text{diag}\{s_i\}$, where $s_i = 1$ if data point $i$ is in the sample and 0 otherwise. Finally, let $W = (m/M)P^{-1}S = \text{diag}\{w_i\}$ be the matrix of sampling weights, where $m$ is the sample size. When all the $p_i = m/M$, note that $W = S$.

It is not difficult to show that for many sample designs and populations (see Kott 1990b and 1991), the *sample weighted OLS estimator*:

$$\hat{\beta}_{W\cdot\text{OLS}} = (X'[I_G \otimes W]X)^{-1}X'[I_G \otimes W]y \tag{7}$$

is a design consistent estimator for $b_{\text{OLS}}$, which means that $\text{plim}_{m\to\infty}(\hat{\beta}_{W\cdot\text{OLS}} - b_{\text{OLS}}) = 0_K$. Under similar conditions, *sample weighted GLS estimator*:

$$\begin{aligned}\hat{\beta}_{W\cdot\text{GLS}} &= (X'[I_G \otimes W][\hat{\Sigma}^{-1} \otimes I_M]X)^{-1}X'[I_G \otimes W][\hat{\Sigma}^{-1} \otimes I_M]y \\ &= (X'[\hat{\Sigma}^{-1} \otimes W]X)^{-1}X'[\hat{\Sigma}^{-1} \otimes W]y,\end{aligned} \tag{8}$$

where

$$\hat{\sigma}_{gf} = r_{\cdot g}'Wr_{\cdot f}\Big/\sum_{i=1}^{M} w_i, \quad \text{and} \quad r = y - X\hat{\beta}_{W\cdot\text{OLS}},$$

is a design consistent estimator for $b_{\text{GLS}}$. Like $b_{\text{OLS}}$ and $b_{\text{GLS}}$ (and for the same reasons), $\hat{\beta}_{W\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{GLS}}$ are equal for an unconstrained system of equations.

If $\hat{\beta}_{W\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{GLS}}$ are design consistent, both are also nearly unbiased estimators of $\beta$ when $\lim_{M\to\infty}\tilde{X}'U/M = 0_{K\times G}$, because $b_{\text{OLS}}$ and $b_{\text{GLS}}$ are; however, $\hat{\beta}_{W\cdot\text{GLS}}$ - like $b_{\text{GLS}}$ - may not be nearly unbiased under the weaker assumption that $\lim_{M\to\infty}X'u/M = 0_K$. (Unbiasedness here is always defined with respect to the model in equation (6)).

## 4. MEAN SQUARED ERROR ESTIMATION

Suppose the sample design is such that there are $H$ strata, $n_h$ distinctly sampled PSU's in stratum $h$, and $m_{hj}$ sampled data points in PSU $hj$. Both $\hat{\beta}_{W\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{GLS}}$ have the form $\hat{\beta} = Cy$. Without loss of generality, they can be rewritten as $\hat{\beta} = C^*y^*$, where $y^* = (y_{11}', \ldots, y_{Hn_H}')$ contains only elements corresponding to sampled data points, and $y_{hj}$ is the vector of $G \times m_{hj}$ $y$-values associated with data points in PSU $hj$. Define $r^*$ and $r_{hj}$ in an analogous manner to $y^*$ and $y_{hj}$.

Let $D_{hj}$ be a diagonal matrix of 0's and 1's such that $D_{hj}y^* = (0', \ldots, y_{hj}', \ldots, 0')$, and let $g_{hj} = C^*D_{hj}r^*$. Extending the design-based linearization variance estimator in a straight forward manner, the estimator for the mean squared error of $\hat{\beta} = C^*y^*$ has the form:

$$\text{mse} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h} g_{hj}g_{hj}' - \frac{1}{n_h}\left(\sum_{j=1}^{n_h} g_{hj}\right)\left(\sum_{j=1}^{n_h} g_{hj}'\right)\right]. \tag{9}$$

Under mild restrictions on the sampling design, mse is nearly unbiased when $U$ (from (3)) $\equiv 0_{M\times G}$ and $V$ obeys the following property:

$$|E(v_{sg}v_{tf})| \begin{cases} = 0 & \text{when } s \text{ and } t \text{ are from different PSU's} \\ < Q & \text{otherwise.} \end{cases}$$

See Kott (1991) for the proof in the $G = 1$ case; the extension to the $G > 1$ case is trivial. The estimator mse remains reasonable when $U \not\equiv 0_{M\times G}$ (see Kott 1990a).

## 5. TEST STATISTICS

Let $\hat{\beta}_{I\cdot\text{OLS}}$ and $\hat{\beta}_{I\cdot\text{GLS}}$ be the unweighted counterparts of $\hat{\beta}_{W\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{GLS}}$ derived by replacing the $W$ in (7) and (8) by $S$. One is often interested in determining whether using the sampling weights really matters. This comes down to testing whether $\hat{\beta}_{I\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{OLS}}$ are significantly different; that is, whether they are estimating the same thing.

When weights *are* determined to matter, another question of some interest is whether $\hat{\beta}_{W\cdot\text{OLS}}$ and $\hat{\beta}_{W\cdot\text{GLS}}$ are significantly different; that is, does $\lim_{M\to\infty}\tilde{X}'U/M = 0_{K\times G}$ hold so that these two estimators are estimating the same thing?

A general statistic for testing whether:

$$\hat{\beta}_{(1)} = \sum_h \sum_j \{C_{(1)}{}^*D_{hj}y^*\} \quad \text{and} \quad \hat{\beta}_{(2)} = \sum_h \sum_j \{C_{(2)}{}^*D_{hj}y^*\}$$

are equal is

$$T^2 = [\hat{\beta}_{(1)} - \hat{\beta}_{(2)}]'A^{-1}[\hat{\beta}_{(1)} - \hat{\beta}_{(2)}], \tag{10}$$

where

$$A = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\sum_{j=1}^{n_h} d_{hj} d_{hj}' - \frac{1}{n_h}\left(\sum_{j=1}^{n_h} d_{hj}\right)\left(\sum_{j=1}^{n_h} d_{hj}'\right)\right],$$

$d_{hj} = C_{(1)}{}^{*}D_{hj}r_{hj(1)} - C_{(2)}{}^{*}D_{hj}r_{hj(2)}$, and $r_{hj(f)} = y_{hj} - X_{hj}\hat{\beta}_{(f)}$.

Under the null hypothesis, the test statistic, $T^2$, is asymptotically a $\chi^2$ random variate with $K$ degrees of freedom. Given our concern for robustness, it seems prudent to question the null hypothesis when prob$(\chi^2_{(K)} > T^2)$ is at considerably less than the standard 0.1 or 0.05 level, but not when $T^2$ is less than its expected value, $K$.

## 6. AN EXAMPLE

Consider the following example synthesized from data from the National Agricultural Statistics Service's June 1989 Agricultural Survey. The data set, previously analyzed in Kott (1990a), is briefly described below.

A sample of 17 primary sampling units was selected from among 4 strata. These PSU's were then subsampled yielding a total sample of 252 farms. Although the sample was random, not all farms had the same probability of selection.

Suppose we are interested in estimating the parameters, $\beta_1$ and $\beta_2$, of the following equation:

$$y_{1i} = x_{1i}\beta_1 + x_{2i}\beta_2 + u_{1i} + v_{1i}, \tag{11}$$

where $i$ denotes a farm,

$y_{1i}$ is farm $i$'s planted soybeans to cropland ratio when $i$'s cropland is positive, zero otherwise;

$x_{1i}$ is 1 if farm $i$ has positive cropland, zero otherwise; and

$x_{2i}$ is farm $i$'s cropland divided by 10,000.

(Note: dropping all sampled farms with zero cropland from the regression equation will have no effect on the parameter estimation, but it can affect mean squared error estimation.)

Letting $\hat{\beta}_{(1)}$ in equation (10) be the pure OLS estimator for the vector $(\beta_1, \beta_2)'$, and $\hat{\beta}_{(2)}$ be the sample weighted estimator, one computes a $T^2$ of 4.58. Under the null hypothesis that OLS and sample weighted least squares are estimating the same thing (for which $u_{1i} \equiv 0$ is sufficient but not necessary), $T^2$ is asymptotically $\chi^2_{(2)}$. We cannot reject this null hypothesis at the 0.1 level. Nevertheless, since $T^2$ is considerably greater than 2, it seems that the existence of a putative missing regressor is more than likely. Thus, the sample weighted regression estimator should be employed rather than the OLS estimator.

Table 1 displays both the pure OLS and the sample weighted coefficient estimates. Although the sample weighted estimator for $\beta_2$ is not significantly different from zero at the 0.1 level, we retain it in the model because it exceeds its estimated root mean squared error. This parallels the reasoning for preferring sample weighted regression over OLS.

Notice the loss of efficiency that results from using the sample weighted estimator in place of pure OLS. The estimated root mean squared error for the $\beta_2$ estimator more than doubles (note: both root mean squared errors were estimated using equation (9)).

**Table 1**
Alternative Estimates for Equation (11)

| | |
|---|---|
| OLS | $y_{1i} = \underset{(.044)}{0.268x_{1i}} - \underset{(3.95)}{0.92x_{2i}} + u_{1i} + v_{1i}$ |
| sample weighted | $y_{1i} = \underset{(.075)}{0.191x_{1i}} + \underset{(9.95)}{12.15x_{2i}} + u_{1i} + v_{1i}$ |
| sample weighted GLS | $y_{1i} = \underset{(0.71)}{0.197x_{1i}} + \underset{(6.97)}{10.26x_{2i}} + u_{1i} + v_{1i}$ |

Numbers in parentheses are root mean squared errors.

We can increase the efficiency of the sample weighted estimator by adding a second farm equation and estimating it and (11) as a system. Let

$$y_{2i} = x_{1i}\beta_3 + u_{2i} + v_{2i}, \tag{12}$$

where $y_{2i}$ is farm $i$'s planted corn to cropland ratio when $i$'s cropland is positive, zero otherwise.

The sample weighted estimators in Table 1 and their estimated root mean squared errors are unchanged under system-wide sample weighted OLS. The system approach, however, allows us to calculate sample weighted GLS estimator for $\beta_1$ and $\beta_2$ which are also displayed in the table. Observe that the estimated root mean squared error for $\beta_2$ is reduced by approximately 30% without a loss of robustness, assuming that sample weighted OLS and GLS are estimating the same thing.

The $T^2$ value for a test comparing the sample weighted OLS and GLS estimators for the vector $(\beta_1, \beta_2)'$ is 0.97. This number is considerably less than 2. Thus, the two estimators do appear to be estimating the same thing. That is to say, there is no additional regressor in one equation related to the putative missing regressor in the other (which is not surprising since when an $x_{2i}$ term was added to the right hand side of equation (12), its estimated coefficient was less than its estimated root mean squared error).

## 7. SIMULTANEOUS EQUATIONS

In a simultaneous equation framework, some of the columns of the dependent variable matrix, $Y$, (see (3)) are actually contained on the right hand side of the $g$th equation (see (5)). Formally, we can write

$$y = Y_{(\cdot)}\alpha + X\beta + u + v \quad \text{or} \quad y = Z\delta + u + v,$$

where

$$Y_{(\cdot)} = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \\ \cdot \\ \cdot \\ Y_{(G)} \end{bmatrix},$$

$$Z = (Y_{(\cdot)}X), \quad \text{and} \quad \delta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Most of the columns of $Y_{(g)}$ are 0-vectors. The rest (no more than G-1 columns) are columns of $Y$ from equation (3).

Define $\hat{Y}_{(g)}$ as $\tilde{X}(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'WY_{(g)}$. Now replace $X$ in (5) by $\hat{Z} = (\hat{Y}_{(.)}\ X)$ and proceed as before. Equation (7) produces $\hat{\delta}_{W\cdot\mathrm{OLS}}$, akin to two stage least squares, while (8) produces $\hat{\delta}_{W\cdot\mathrm{GLS}}$, akin to three stage least squares. Mean squared error estimation follows along the same line of reasoning that produced equation (9).

## 8. DISCUSSION

The purpose of this paper was to show how procedures developed in the design-based survey sampling literature – in particular, sample weighted regression and the linearization mean squared error estimator – could be adopted to the estimation of a system of linear equations.

One somewhat unexpected discovery was when estimating the parameters of a constrained linear system, the sample weighted analogues of OLS and GLS might be estimating different things. On further reflection this is not so suprising. If there are missing regressors in our working model, perhaps we don't always know enough about the true model to put constraints on the parameters in the first place.

It is important to realize that mse in equation (9) can be used to estimate the mean squared error of parameter estimators even when there are no missing regressors. The advantages of mse to conventional practice is that it allows for the possibility of heteroscedasticity and complex correlations across data points (but within PSU's).

If there are no missing regressors, however, the following estimator has all the advantages of mse and is generally more efficient:

$$\mathrm{mse}' = \sum_{h=1}^{H} \sum_{j=1}^{n} \frac{n}{n-1} \{g_{hj} g_{hj}' - gg'/n\}, \tag{13}$$

where $n = \sum n_h$ and $g = \sum \sum g_{hj}$ (note: if $Xq = (1, \ldots, 1)'$ for some $K$-vector $q$, then $g = 0$).

When there *are* missing regressors the diagonal elements of mse′ may tend to be biased upward. The reasoning here follows that in Wolter (1985) for collapsed strata variance estimators in design-based sampling theory.

## REFERENCES

JOHNSTON, J. (1972). *Econometric Methods*, (second edition). New York: McGraw Hill.

KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, forthcoming.

KOTT, P.S. (1990a). What does performing a linear regression on survey data mean? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.

KOTT, P.S. (1990b). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

SHAH, B.V., HOLT, M.M., and FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.

Survey Methodology, June 1991
Vol. 17, No. 1, pp. 99–119
Statistics Canada



# The Evaluation of Errors in National Accounts Data: Provisional and Revised Estimates

LUIGI BIGGERI and UGO TRIVELLATO[1]

ABSTRACT

This article is a critical review of recent developments in the evaluation of the reliability of provisional national accounts estimates. First, we will sketch a theoretical outline of the process used to produce successive estimates of an aggregate, and will reflect upon its implications regarding the design of the analyses of errors in provisional data. Second, particular attention will be focused upon the choice of elementary measurements of errors and suitable integrative accuracy indices, and the impact of revisions on constant price aggregates and implicit deflationary factors. Finally, the results of some empirical analysis on discordances between provisional and revised estimates will be summarily discussed on the basis of a comparison of national accounts data in Canada, Italy, and the United States.

KEY WORDS: Revision of data; National accounts; Accuracy indices; Constant price aggregates.

## 1. INTRODUCTION

The evaluation of the reliability of national accounts estimates is important because of the effects of these estimates on economic policy analysis and decisions. At the same time, these estimates are difficult to evaluate, because they are affected by a very large number of complex sources of error.

Thus, it is difficult to arrive at a single reliability criterion that will be both convincing at the conceptual level and practicable. On the other hand, it seems reasonable and feasible to establish a certain number of partial criteria that could be used to measure the principal features of reliability. For a review of these various criteria, see Novak (1975) and Trivellato (1987).

The purpose of this paper is to describe and discuss critically certain recent developments in the analysis of errors in provisional national accounts estimates.

National accounts aggregates as well as many other economic aggregates and indicators are initially published in the form of provisional estimates that are subsequently subjected to various revisions.

The process of revision is determined above all by the need to ensure that information can be made quickly available and, at the same time, by the time required to collect and process all the data currently used to estimate the aggregates ("timeliness vs. accuracy," according to the effective polarization of Wilton and Smith 1974). Typically, this gives rise to a revision process characterized by a preliminary estimate and by a subsequent series of routine revisions that follow closely upon one another. From time to time, it may happen that, in order to improve the accuracy or relevance of the data, modifications are introduced into the classification and accounting plans, and/or in the basic statistical surveys and estimation methods. In turn, this leads to new evaluations, in which case, we speak of occasional or extraordinary revisions.

The characteristic feature of the errors present in provisional estimates resides in the fact that they are present in these estimates, and, by definition eliminated in the revised estimates. Thus, they can be measured by comparing the two estimates. Strictly speaking, it is true that

[1] Luigi Biggeri, Dipartimento Statistico, Università di Firenze, Italia; Ugo Trivellato, Dipartimento di Scienze Statistiche, Università di Padova, Italia.

the difference between a provisional estimate and the corresponding revised estimate is merely a measurement of the error difference between the two successive estimates. Nevertheless, when the process of revision is based on information that is progressively more complete, and on evaluation methods that are increasingly more refined, as is precisely the case in the developed countries, it is reasonable to assume that the final estimate is the closest to the true unknown value, and for purposes of comparison, to treat it as the reference value.

An understanding of the characteristics and behaviour of discordances between the provisional and revised estimates can be truly useful. In fact, significant and frequent differences are a discouraging sign respecting the quality of the basic data and/or the estimation methods used in the preliminary evaluations, and suggest at least that the results should be used carefully. (Naturally, this does not mean that, if the differences are few, this will necessarily represent a guarantee of the quality of the data. The fact that a preliminary estimate is not revised may simply mean that we do not have enough elements to improve it, regardless of its reliability). Moreover, if the differences also seem to be systematic, this may, at the same time, represent a helpful warning both to the user, who will eventually introduce correction factors; and to the producer, who may derive from this, suggestions to improve his provisional estimates.

A review of many statistical analyses of the discordance between provisional and subsequently revised estimates can be found in Trivellato (1986a). Other recent contributions have been made by Zarnowitz (1982), Mankiw, Runkle and Shapiro (1984), Mankiw and Shapiro (1986), McNees (1986), Mork (1987) and Lefrançois (1988).

In this area, there is a considerable amount of literature. Nevertheless, some questions (both methodological and substantive) concerning the way in which these analyses should be carried out have been at least partly neglected or resolved in a way that is not always satisfactory. In particular, this is true for:

(a) The identification of a design for the analysis of the accuracy of provisional estimates that is consistent with the characteristics of the process of revision;

(b) The definition of suitable elementary measurements of the discordances between provisional and revised estimates; the choice of suitable integrative indices that could provide adequate information on the overall accuracy of a series of provisional estimates; and, if the process of revision is carried out in several stages, of indices that could be decomposed in order to be able to verify the convergence of the provisional estimates with the final value.

(c) The examination of the effects of errors in provisional estimates upon the derived series, especially upon constant price aggregates and implicit deflationary factors.

The three following sections will deal with these problems in order. Finally, section 5 will provide a very brief summary of the results of some empirical analyses of the discordances between provisional and revised estimates, on the basis of comparisons between national accounts data from Canada, the United States, and Italy.

## 2. A THEORETICAL OUTLINE OF THE REVISION PROCESS

A sufficiently general illustration of the process of revision of economic aggregates is shown in Table 1, which provides an immediate visualization of the relationships that exist between the reference and publication periods of the various estimates.

Three types of estimates and revisions will clearly emerge from an examination of this illustration:

**Table 1**

Publication Plan for Successive Aggregate Estimates[a]

| Publication Periods | Reference Periods | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $t-2h+1$ | . | . | $t-h$ | $t-h+1$ | . | $t-m$ | . | . | $t-1$ | $t$ | $t+1$ | . | . | $T-m$ | . | . | $T-1$ | $T$ | $T+1$ |
| $t-2h+2$ | ${}_1p_{t-2h+1}$ | | | | | | | | | | | | | | | | | | | |
| . | ${}_2p_{t-2h+1}$ | | | | | | | | | | | | | | | | | | | |
| . | . | . | . | | | | | | | | | | | | | | | | | |
| . | . | . | . | ${}_1p_{t-h}$ | | | | | | | | | | | | | | | | |
| $t-2h+m+2$ | ${}_mp_{t-2h+1}$ | . | . | ${}_2p_{t-h}$ | ${}_1p_{t-h+1}$ | | | | | | | | | | | | | | | |
| $t-2h+m+3$ | $r_{t-2h+1}$ | . | . | . | ${}_2p_{t-h+1}$ | . | | | | | | | | | | | | | | |
| . | | . | . | . | . | . | ${}_1p_{t-m}$ | | | | | | | | | | | | | |
| . | | | . | ${}_mp_{t-h}$ | . | . | ${}_2p_{t-m}$ | ${}_1p_.$ | | | | | | | | | | | | |
| . | | | | $r_{t-h}$ | ${}_mp_{t-h+1}$ | . | . | ${}_2p_.$ | ${}_1p_.$ | | | | | | | | | | | |
| . | | | | | $r_{t-h+1}$ | . | . | . | ${}_2p_.$ | ${}_1p_{t-1}$ | | | | | | | | | | |
| . | ${}_br_{t-2h+1}$ | . | . | ${}_br_{t-h}$ | | . | ${}_mp_{t-m}$ | . | . | ${}_2p_{t-1}$ | ${}_1p_t$ | | | | | | | | | |
| $t+2$ | | | | | | | $r_{t-m}$ | ${}_mp_.$ | . | . | ${}_2p_t$ | ${}_1p_{t+1}$ | | | | | | | | |
| . | | | | | | | | $r_.$ | ${}_mp_.$ | . | . | ${}_2p_{t+1}$ | . | | | | | | | |
| . | | | | | | | | | $r_.$ | ${}_mp_{t-1}$ | . | . | . | . | | | | | | |
| . | | | | | | | | | | $r_{t-1}$ | ${}_mp_t$ | . | . | . | ${}_1p_{T-m}$ | | | | | |
| $t+m+2$ | | | | | | | | | | | $r_t$ | ${}_mp_{t+1}$ | . | . | ${}_2p_{T-m}$ | ${}_1p_.$ | | | | |
| $t+m+3$ | | | | | | | | | | | | $r_{t+1}$ | . | . | . | ${}_2p_.$ | ${}_1p_.$ | | | |
| . | | | | | ${}_br_{t-h+1}$ | | ${}_br_{t-m}$ | . | . | ${}_br_{t-1}$ | ${}_br_t$ | | . | . | . | . | ${}_2p_.$ | ${}_1p_{T-1}$ | | |
| . | | | | | | | | | | | | | . | . | ${}_mp_{T-m}$ | . | . | ${}_2p_{T-1}$ | ${}_1p_T$ | |
| $T+2$ | | | | | | | | | | | | | | | $r_{T-m}$ | ${}_mp_.$ | . | . | ${}_2p_T$ | ${}_1p_{T+1}$ |
| $T+3$ | ${}_sr_{t-2h+1}$ | . | . | ${}_sr_{t-h}$ | ${}_sr_{t-h+1}$ | . | ${}_sr_{t-m}$ | . | . | ${}_sr_{t-1}$ | ${}_sr_t$ | ${}_sr_{t+1}$ | . | . | ${}_sr_{T-m}$ | $r^*$ | ${}_mp^*$ | . | . | ${}_2p^*_{T+1}$ |

[a] The process of revision provides $m$ successive current revisions $({}_2p, \ldots, {}_mp, r)$, one benchmark revision each $h$ period $({}_br)$, and one extraordinary revision when the retrospective series is reconstructed on the basis of time $T+2$. (Source: Biggeri 1984, p.24).

(a) those lying on the main diagonal and the lower diagonals describe a current process of revision that takes place in $m$ steps; from the preliminary estimate ${}_1p_t$ to the final estimate $r_t$ (final in relation to the current process of revision);

(b) those of the ${}_br_t$ type, which lie horizontally, incorporate the benchmark controls where the series is reconstructed with reference to a period that goes from one benchmark to the next;

(c) those of the ${}_sr_t$ type are the result of an occasional revision and also lie horizontally; they concern the retrospective reconstruction of the series for a generally rather long period of time.

The figure also shows how the presence of benchmark and occasional revisions that are superimposed upon the current revision process results in mixed revisions. Thus, the revision that is carried out is not homogeneous with either the previous or the subsequent revisions carried out in the current estimation sequence.

If we have available a chronological series of provisional estimates of an aggregate for times 1 to $n$ and that of its corresponding revised estimates, the problem of evaluating the accuracy of the first in relation to the second can be formally reduced to that of evaluating the validity of these forecasts, a problem that has been amply covered in the literature. Nevertheless, the mechanism of revising national accounts aggregates, an illustration of which is shown in Table 1, has certain characteristic features that we must keep in mind. First, it is important to remember that analyses of the process of revision explicitly take into account the existence of the three types of revisions listed above. Moreover, in order to understand correctly the characteristics of the revisions under consideration, the analyses must also exclude all comparisons that involve mixed revisions.

A second consideration of particular importance concerns the methods and criteria used to analyze the accuracy of provisional estimates. Most studies in the statistical and economic literature on the revision of national accounts aim to establish the accuracy of provisional estimates on the basis of statistical measurements that are generally descriptive. Even though very different, and interesting, approaches have been recently proposed (for example, by Mankiw and Shapiro, 1986; and by Lefrançois, 1988), they cannot be entirely satisfactory for an analysis of the properties of provisional data. In fact, they represent convenient integrations, but cannot offer a real alternative to the analysis of the properties of provisional data, for which it is important to adopt criteria and accuracy measurements that are essentially descriptive.

Finally, when carrying out an analysis of the accuracy of provisional estimates, it seems preferable to treat the provisional and final estimates in a non-symmetrical fashion, and to consider the revised series $r_t$ as the reference series. This choice is motivated precisely by the fact that we propose to evaluate the accuracy of the provisional estimates on the basis of the final estimates (or, in any case, those that are similar to the latter in the comparison under consideration).

## 3. ELEMENTARY MEASUREMENTS AND INTEGRATIVE ERROR INDICATORS IN PROVISIONAL ESTIMATES

### 3.1 The Choice Between Errors and Relative Errors

The error in a provisional estimate ($p_t$) of the level of an economic aggregate can be obtained as follows:

$$v_t = p_t - r_t. \tag{1}$$

However, this could be inadequate to compare the accuracy of several aggregates, since the results depend upon the measuring unit and the order of magnitude of the aggregate under consideration. In this case, it may be preferable to use the relative error, which is defined as follows:

$$e_t = (p_t - r_t)/r_t = p_t/r_t - 1. \tag{2}$$

The choice between analyzing the errors or the relative errors can be a crucial one in many ways, and deserves a more in-depth discussion. As we will see in section 3.2, in fact, the pertinent integrative accuracy indices for provisional estimates are simple averages of $v_t$ and $e_t$ respectively (or of their suitable transformations). Thus, the use of simple means is reasonable if the series have been derived from a purely random process, particularly if they do not have a trend component. In the opposite case, we lose information, and the analysis may lead to obscure or even misleading conclusions. For this reason, it is important to carry out preliminary verifications, which can be done using various tests (see for example Malinvaud 1969, p. 473-481; Kendall 1973, p. 22-28; Box and Jenkins 1970, p. 34-36 and 287-298).

In relation to the specific problem of choosing between errors and relative errors, a particularly useful criterion is offered by the analysis of the parameters of a suitable model of the provisional and final data. In its simplest form, this can be specified as follows:

$$p_t = \alpha + \beta r_t + \epsilon_t, \tag{3}$$

where $\epsilon_t$ is the random error. From model (3) we can obtain:

$$v_t = \alpha + (\beta - 1)r_t + \epsilon_t, \tag{4}$$

$$e_t = (\beta - 1) + \alpha \frac{1}{r_t} + \frac{\epsilon_t}{r_t}. \tag{5}$$

Formulae (4) and (5) show that, generally, both the error and the relative error of the provisional estimate depend upon the level of the corresponding final estimate. From these we can then derive the conditions that must be satisfied in order to justify the use of integrative indices based on the errors or relative errors respectively:

(a) $v_t$ does not depend upon $r_t$ if $\beta$ equals 1, and the $\epsilon_t$s are homoscedastic, and not temporally correlated;

(b) $e_t$ does not depend upon $r_t$ if $\alpha$ equals zero, and the variance of the $\epsilon_t$s is proportional to the square of the level of the series (the $\epsilon_t$s are not temporally correlated).

It will be immediately clear that estimating (4) with ordinary least squares is equivalent to estimating (5) with generalized least squares ( with $E(\epsilon\epsilon') = \sigma_\epsilon^2\Omega$, where $\Omega$ is a diagonal matrix with $w_{tt} = r_t^2$). It is also immediately evident that (3) and (4) differ only by 1 in the angular coefficient. Thus, in order to choose between analyzing the errors or the relative errors, it becomes crucial to verify the homoscedasticity of $\epsilon_t$ in (3). In order to do this, Trivellato, Di Fonzo, and Rettore (1986) developed a simple non-parametric test based on the order of the estimated residuals, which does not depend upon specifying a particular stochastic structure for equation (3); this conforms with the little we know *a priori* about the relationship that exists between provisional and final estimates.

This test can also be used to examine the hypothesis of the stability of the parameters in equation (3). In fact, it is evident that conditions (a) and (b) above are not exhaustive, and that a possible reason for their lack of validity is the presence of instabilities in the revision process. The importance of carrying out this type of verification is based, among other things, on the fact that we study the superposition of occasional revisions onto the current revision process. If the results of the test favour the hypothesis of stability, it is appropriate to analyze the current revision process as a whole. In the opposite case, the analysis must be carried out separately for the two periods preceding and following the occasional revision.

Finally, it is important to point out that specification (3) is intentionally stylized and can be essentially used to explore the process of revising an annual series, when the residual $\epsilon_t$s are not self-correlated. Nevertheless, it is easy to generalize it in various ways: (i) by introducing a vector of non-random explanatory variables in order to take into account any eventual factors that may affect the process of revision (benchmark revisions, seasonal factors, *etc.*); (ii) by assuming that the residuals are self- correlated; (iii) by modelling jointly the process of revision of several series using a seemingly unrelated regression equations system. For these developments, see Trivellato and Rettore (1986), and Bordignon and Trivellato (1989).

### 3.2 Integrative Indices of the Accuracy of Provisional Estimates and "Low Coherence"

In order to characterize suitable integrative indices of the overall accuracy of a provisional estimates vector, it is important to refer to the property of "low coherence," as defined by Trivellato (1986b). In substance, in reference to two series of provisional estimates corresponding to the same series of final estimates, this requires that if the first series shows errors that are smaller than or equal to those of the second in absolute value, and if significant inequality is present in at least one case, the accuracy index of the first series must be smaller than the index of the second series, and thus indicates that the first series of provisional estimates is closer to the final data than the second.

As we will see, the central reference to integrative low coherence indices does not lead to any significant innovations in relation to the normally used measurements (we will essentially use the mean absolute error and the mean quadratic error). This concept, together with the choice of analyzing either the errors or the relative errors, and the eventual identification of homogeneous sub-periods in the revision process can nevertheless lead to an adequate implementation of empirical analyses, while avoiding the lack of accuracy that can often be found in the literature.

It is also important to consider the decompositions of the low-coherence integrative indices. From a slightly different point of view, the presence and weight of specific components can eventually be revealed by looking at the estimated parameters in equation (3) (Mincer and Zarnowitz 1969; Hatanaka 1974; Hempenius 1980; Trivellato 1986a).

The indices that we should emphasize are the following:

(I) **Errors**

If the criteria used to choose between the errors and the relative errors (see section 3.1) lead to the implementation of an analysis of errors, two low-coherence integrative indices would be the mean absolute error and the square root of the mean quadratic error:

$$\bar{v}' = \Sigma \mid v_t \mid /n, \tag{6}$$

$$d_v = \sqrt{(\Sigma v_t^2/n)}, \tag{7}$$

where the sum is extended to $n$ terms in the series.

The following decomposition, similar to Theil's asymmetric decomposition (1966), but developed by treating $r_t$ as a reference series, makes it possible to show the deficiencies in the performance of provisional estimates:

$$d_v^2 = (\bar{p} - \bar{r})^2 + (s_r - \hat{\rho} s_p)^2 + (1 - \hat{\rho}^2)\, s_p^2, \tag{8}$$

where $s_r$ and $s_p$ are the standard deviations of $r_t$ and $p_t$ respectively, and $\hat{\rho}$ is the coefficient of correlation between $p_t$ and $r_t$. From formula (8) we can derive the following relative decomposition:

$$1 = U_v^M + U_v^R + U_v^D. \tag{9}$$

The three terms in equation (9) represent fractions of the mean quadratic error and can be interpreted as follows: a bias component between the means of the two series, $U_v^M$; a regression component attributable to a deviation of 1 from the regression coefficient of $p_t$ over $r_t$, $U_v^R$; and a random error component attributable to the variance of the regression errors, $U_v^D$.

When assessing the quality of estimates, it is also useful to take into account the mean and mean quadratic deviation of the errors; that is, $\bar{v} = \sum v_t/n = \bar{p} - \bar{r}$ and $s_v = \sqrt{(\sum (v_t - \bar{v})^2/n)}$ respectively. Clearly, $\bar{v}$ is not an index of the accuracy of provisional estimates. It provides information only about the direction and dimension of the mean level error. Nevertheless, it is a statistical measurement of remarkable significance for at least two reasons: (i) if $\bar{v} \simeq 0$, then there is no bias component ($U_v^M = 0$); (ii) if $\bar{v} \neq 0$, it will be instructive to examine $\bar{v}$ and $\bar{v}'$ jointly, since $|\bar{v}| \simeq \bar{v}'$ if and only if the errors are always (or almost always) in the same direction. Thus, the comparison between the two indices can show the presence of a possible systematic component in the level errors of the provisional estimates.

**(II) Relative Errors**

When the criterion used in section 3.1 leads to the analysis of the relative errors, two suitable integrative indices are the absolute mean relative error, and the square root of the mean quadratic relative error respectively:

$$\bar{e}' = \sum |e_t| / n, \tag{10}$$

$$d_e = \sqrt{(\sum e_t^2/n)}. \tag{11}$$

The two indices are defined if $r_t \neq 0$ for each $t$, a condition that remains non-restrictive for most economic aggregates.

A reasonable decomposition of $d_e$ is as follows:

$$d_e^2 = [(\overline{p/r}) - 1]^2 + \frac{1}{n} \sum [(p_t/r_t) - (\overline{p/r})]^2, \tag{12}$$

from which can be derived the relative decomposition

$$1 = U_e^b + U_e^c, \tag{13}$$

$U_e^b$ being the fraction of the mean quadratic relative error due bias, and $U_e^c$ the fraction due to the random component.

Two other specific components of the lack of accuracy should also be mentioned; these are, the mean and mean quadratic deviation of relative errors; that is, $\bar{e} = \sum e_t/n = (\overline{p/r}) - 1$ and $s_e = \sqrt{(\sum (e_t - \bar{e})^2/n)}$ respectively. *Mutatis mutandis*, what we have said for $\bar{v}$ is also valid for $\bar{e}$, both in terms of its meaning and for the comparison with $\bar{e}'$.

### 3.3 Decomposition of the Revision Process from the Preliminary to the Final Estimate

The accuracy indices discussed until now involve a comparison between two vectors $p$ and $r$. However, as we have already observed, the process of revision of economic aggregates cannot usually be completed in a single stage. Thus, how can we summarily estimate the convergence of the succession of provisional estimates with the final estimate?

In this section, we will discuss the decomposition of a process of revision that takes place in two stages (a procedure for the most common situation of $m - 1$ steps, can be found in Wilton and Smith 1974). This decomposition will be evaluated with reference to (i) errors; and (ii) relative errors respectively.

When we analyze the errors, the decomposition of the mean quadratic error of ${}_1p$ in relation to $r$ in the two phases from ${}_1p$ to ${}_2p$ and from ${}_2p$ to $r$ is easy to obtain. Let ${}_1v = {}_1p - {}_2p$, ${}_2v = {}_2p - r$, ${}_Tv = {}_1p - r$ and let $d^2_{v1}$, $d^2_{v2}$, $d^2_{vT}$ show the mean quadratic error associated with vectors ${}_1v$, ${}_2v$ and ${}_Tv$ respectively. The result is:

$$d^2_{vT} = d^2_{v1} + d^2_{v2} + \frac{2}{n} \sum {}_1v_t \, {}_2v_t, \tag{14}$$

from which we derive the relative decomposition:

$$1 = D_I + D_{II} + D_{I,II}, \tag{15}$$

where the first two components represent the fraction of $d^2_{vT}$ that can be attributed to the mean quadratic error of the first and second revision respectively, while $D_{I,II}$ represents the interaction between ${}_1v$ and ${}_2v$.

On the other hand, if we consider the relative errors, an integrative decomposition of the process of revision in the two stages may be problematic. Useful indications can be derived from the equation $({}_1p_t - r_t)/r_t = ({}_1p_t - {}_2p_t)/r_t + ({}_2p_t - r_t)/r_t$; thus, by dividing the two members by $({}_1p_t - r_t)/r_t$ and taking into account the mean values, we obtain:

$$1 = \frac{1}{n} \sum ({}_1p_t - {}_2p_t)/({}_1p_t - r_t) + \frac{1}{n} \sum ({}_2p_t - r_t)/({}_1p_t - r_t). \tag{16}$$

If we assume a stable order of estimates of the ${}_1p_t$ type $< {}_2p_t < r_t$ for each $t$, the two terms of equation (16) can be interpreted as mean fractions of the discordance between ${}_1p$ and $r$ eliminated by the first and second revision respectively. Nevertheless, this interpretation can be questioned in terms of order violations, especially when this occurs together with small ${}_1p_t - r_t$ differences. In general, it seems more reasonable to discard the hypothesis that the order between estimates must be respected for the whole period under consideration. In this situation, a qualitative evaluation of the degree of stability of the order relationships between the estimates, and the importance of the two stages in the revision process can be obtained from an examination of values of $\bar{e}$ and $\bar{e}'$, together with comparisons between ${}_1p$ and ${}_2p$, between ${}_2p$ and $r$, and between ${}_1p$ and $r$. Clearly, the relationship $| \bar{e} | = \bar{e}'$ is valid if and only if the order between the two series of estimates involved in the comparison is always respected. However, we can say that the order is "most often" respected when $| \bar{e} | \simeq \bar{e}'$ meaning that:

$$(\bar{e}' - |\bar{e}|)/\bar{e}' < 2f, \tag{17}$$

where $f$ is a fraction of violations of a predetermined order, that is considered to be acceptable.

Equation (17) can be justified as follows: let $e_t$ $(t = 1, 2, \ldots, n)$; the relative errors, $n_1$, are strictly positive; $n_2$s are strictly negative; and $n_3$s equal zero $(n_1 + n_2 + n_3 = n)$. It is easy to verify the following relation:

$$\bar{e}' - |\bar{e}| = \begin{cases} 2n^{-1} \sum_{i=1}^{n1} e_i & \text{if } \bar{e} < 0 \\ -2n^{-1} \sum_{j=1}^{n2} e_j & \text{if } \bar{e} > 0, \end{cases}$$

where quantities $\sum_{i=1}^{n1} e_i$ and $-\sum_{j=1}^{n2} e_j$ correspond to the absolute value of the sum of the sign violations of $\bar{e}$.

The quantity

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2\sum_{i=1}^{n1} e_i \Big/ \left(\sum_{i=1}^{n1} e_i - \sum_{j=1}^{n2} e_j\right) & \text{if } \bar{e} < 0 \\ -2\sum_{j=1}^{n2} e_j \Big/ \left(\sum_{i=1}^{n1} e_i - \sum_{j=1}^{n2} e_j\right) & \text{if } \bar{e} < 0 \end{cases}$$

is thus an index of the importance of the sign violations of $\bar{e}$. This index is bounded as follows: $0 \leq (\bar{e}' - |\bar{e}|)/\bar{e}' \leq 1$; the lower limit is reached when $|\bar{e}| = \bar{e}'$; that is, when $n_1 = 0$, or when $n_2 = 0$ (if $n_1 = n_2 = 0$, there are no order violation problems); and the upper limit is reached when $\bar{e} = 0$; that is, when $\sum_{i=1}^{n1} e_i = \sum_{j=1}^{n2} e_j$.

The relationship $|\bar{e}| \simeq \bar{e}'$ is defined by comparing the value assumed by the index with a critical value calculated in accordance with the hypothesis of the equality, in absolute value, of all the non-zero relative errors. This produces the following result:

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2n_1/(n_1 + n_2) & \text{if } \bar{e} < 0 \\ 2n_2/(n_1 + n_2) & \text{if } \bar{e} > 0. \end{cases}$$

Under this particular hypothesis, the index is thus equal to twice the fraction of strict violations of the sign of $\bar{e}$.

Thus, it is possible: (i) to rank the estimates ${}_1p$ and ${}_2p$ on the basis of a mean criterion; that is, the values of $\bar{e}$ associated with the various comparisons; (ii) to examine the degree of stability of the order on the basis of equation (17); (iii) to use decomposition (16), or to estimate qualitatively in any other way, the magnitude of the two stages, by observing the values of $\bar{e}$ associated with the various comparisons, when there is a high degree of stability.

## 4. THE IMPACT OF ERRORS IN PROVISIONAL ESTIMATES UPON CONSTANT PRICE AGGREGATES

Evidently, the revision process has certain effects upon the derived series, which often contain the information that holds the greatest interest for analysts and policy-makers. This is particularly true in the case of estimates of the rate of variation, seasonally-adjusted series, constant price aggregates, and implicit deflationary factors. The impact of revisions on measurements

of variation has been analyzed descriptively by Trivellato, Di Fonzo and Rettore (1986); and by Rao, Srinath and Quenneville (1989), who attempted to determine the best estimator of variation. The effects of revisions upon the seasonal adjustment procedures have been described in papers by Pierce (1980), Wallis (1982), and Maravall and Pierce (1983). The effects of the revision process upon constant price aggregates and implicit deflationary factors will be examined in this section.

The criteria and indices discussed in section 3.2 could evidently be used to analyze the accuracy of provisional constant price estimates. It would nevertheless be interesting to illustrate the formal characteristics of errors in these estimates, and to show the relationships that exist between estimation errors in current price and constant price aggregates, as well as the implicit deflationary factors. The emphasis will be solely upon aggregates of national economic accounts made up of the flow of goods, and the aggregates obtained using accounting balances for these goods.

Let us consider the simple case of a series of provisional estimates and the corresponding series of final estimates. By observing any aggregate at time $t$, consisting of $k$ elementary goods and services, the final estimates can be obtained as follows (when we have a sum of $k$ elementary goods and services): $A_t = \sum p_t q_t$ is the current price aggregate; $C_t = \sum p_o q_t$ is the constant price aggregate (0 base); $P_t$ is the Paasche price index (that is, $P_t = \sum p_t q_t / \sum p_o q_t$); $Q_t$ is the Laspeyres quantity index (that is, $Q_t = \sum p_o q_t / \sum p_o q_o$); $D_t = A_t / C_t$ is the implicit deflationary factor with a Paasche structure. The corresponding provisional estimates are identified with the subscript $p$ (thus, for example, ${}_pA_t$ and $A_t$ are respectively the provisional and final estimates of the current price aggregate).

In order to explain the characteristics of the errors in provisional estimates of constant price aggregates, it would be useful to remember that the evaluation of the constant price aggregate at time $t$ may be obtained using three methods: (i) if we have a series of quantities for all the goods and services that make up the aggregate and the corresponding prices for the base year, we use the direct relationship: $C_t = \sum p_o q_t$; (ii) if we have a Laspeyres-type quantity index, we multiply the value of the aggregate in the base year by this index: $C_t = (\sum p_o q_o)\, Q_t$; (iii) if we have a Paasche-type price index, we divide the current price aggregate by this index: $C_t = (\sum p_t q_t)/P_t$.

In relation with the various useable evaluation criteria, the error in the provisional level estimate (as well as the relative error) of a constant price aggregate can be obtained as follows:

$$ {}_pC_t - C_t \equiv \sum p_o\, {}_pq_t - \sum p_o q_t = \sum p_o q_o \left( \frac{\sum p_o\, {}_pq_t}{\sum p_o q_o} - \frac{\sum p_o q_t}{\sum p_o q_o} \right) = A_o({}_pQ_t - Q_t), \quad (18.1) $$

$$ \equiv (\sum p_o q_o)\, {}_pQ_t - (\sum p_o q_o)\, Q_t = A_o({}_pQ_t - Q_t), \quad (18.2) $$

$$ \equiv {}_pA_t/{}_pP_t - A_t/P_t = (\sum {}_pp_t\, {}_pq_t)/{}_pP_t - (\sum p_t q_t)/P_t. \quad (18.3) $$

Equations (18) show that, if we adopt the direct evaluation criterion (18.1), or extrapolation by means of an index of quantity (18.2), the estimation error coincides with the estimation error in the index of quantity, multiplied by constant $A_o$. On the other hand, this result cannot be obtained using the other indirect deflation criterion a price index, because, in general, (18.3) cannot be simplified in a way that is useful for our purposes. In other words, in this third case, the error in the estimate of the level of the constant price aggregate generally depends both upon errors in the provisional estimates of quantities and upon errors in the provisional estimates of prices.

The implication of this for the interpretation of errors in the estimates of constant price aggregates can be quickly made clear. Even though, theoretically, the three evaluation criteria are identical, this is far from being the case in practice, due to the availability and quantity of data. It is important to remember that evaluation at constant prices is generally obtained and a very desaggregated level, and that the aggregates are subsequently obtained by adding. These aggregates are thus evaluated with the deflation criterion, errors in the constant price estimates are not only errors in the volume estimates , but also contain errors in the price estimates.

If we now consider the relative error in the estimate of a current price aggregate and make explicit its relationship with the relative error of the corresponding constant price aggregate, we have the following equation:

$$\frac{{}_pA_t - A_t}{A_t} = \frac{{}_pC_t - C_t}{C_t} + \frac{{}_pD_t - D_t}{D_t} + \frac{{}_pC_t - C_t}{C_t} \times \frac{{}_pD_t - D_t}{D_t}.$$

This equation is evidently valid for the mean relative error of a series of $n$ estimates, and we can obtain the following relative decomposition:

$$1 = \bar{e}(C)/\bar{e}(A) + \bar{e}(D)/\bar{e}(A) + \bar{e}(C;D)/\bar{e}(A), \tag{19}$$

where

$$\bar{e}(A) = \frac{1}{n} \Sigma \left( \frac{{}_pA_t - A_t}{A_t} \right)$$

($\bar{e}(C)$ and $\bar{e}(D)$ are defined in a similar manner, and

$$\bar{e}(C;D) = \frac{1}{n} \Sigma \left( \frac{{}_pC_t - C_t}{C_t} \right) \left( \frac{{}_pD_t - D_t}{D_t} \right).$$

If we leave aside the interaction component, we will then have an approximately additive decomposition: the mean relative error in the current price aggregate is equal to the sum of mean relative errors of the constant price aggregate and the implicit deflationary factor (or better still, if we take into account the estimation process: the mean relative error of the implicit deflationary factor can be approximately obtained by the difference between the mean relative errors of the current price aggregate and constant price aggregate respectively).

It is important to emphasize that the decomposition is between $\bar{e}(C)$ and $\bar{e}(D)$, and not between the "error in the quantities" and the "error in the prices." We have already seen that, in general, ${}_pC_t - C_t$ and $\bar{e}(C)$ also reflect any errors in the provisional price estimates. On the other hand, by definition, ${}_pD_t - D_t$, and $\bar{e}(D)$, are a function of errors in the provisional estimates both of prices and of quantities. Thus, at the end, both components, "the error in the constant price estimate" and "the error in the estimate of the implicit deflationary factor" incorporate errors in the estimates of price and quantities; and the possibility of interpreting ${}_pC_t - C_t$ and ${}_pD_t - D_t$ as "error in the quantities" and "error in the prices" respectively can be relegated only to extreme cases.

We can identify the following extreme cases:

(a) ${}_pC_t - C_t$ can be interpreted as "an error in quantities" only when the evaluation of the constant price aggregate has been carried out with the direct criterion or by extrapolation using an index of quantity;

(b) ${}_pD_t - D_t$ can be interpreted as "an error in the prices" only in the absence of a revision in the quantities. In this case, the error in the constant price estimate is evidently zero, and the relative current price error can be expressed as a linear combination of estimation errors in the prices of $k$ elementary goods and services:

$$({}_pA_t - A_t)/A_t = ({}_pD_t - D_t)/D_t = \sum q_t({}_pp_t - p_t)/\sum p_t q_t = \sum a_t({}_pp_t - p_t),$$

where

$$a_t = q_t/\sum p_t q_t .$$

(c) If there are no revisions in the prices, we have:

$$({}_pA_t - A_t)/A_t = \sum p_t({}_pq_t - q_t)/\sum p_t q_t = ({}_pC_t/C_t)({}_pD_t/D_t) - 1.$$

The absence of revision in the prices is not enough to distinguish between "errors in quantity" and "errors in prices," since the revision of quantities modifies the implicit deflationary factor.

## 5. SOME EMPIRICAL ANALYSES ON REVISIONS OF NATIONAL ACCOUNTS DATA IN CANADA, THE UNITED STATES AND ITALY

### 5.1 The Process Used to Revise National Accounts Data in the Three Countries and the Analyses Carried Out

In this last section we will summarize some of the results of empirical analyses on the revisions of national accounts in Canada, the United States and Italy.

The comparison of the characteristics of the process of revision used in the three countries contains a certain number of inevitable simplifications in the description and analyses of the revisions. For recent in-depth studies on each of the countries under consideration, refer to Lefrançois (1988) for Canada, Parker (1986) and Mork (1987) for the United States, and Di Fonzo, Rettore and Trivellato (1986) for Italy.

Charts of the processes used to revise national accounts data on a yearly and quarterly basis in the three countries under consideration are shown in Tables 2, 3 and 4.

The empirical analyses were carried out on the following data:

(a) Canada: quarterly non-seasonally-adjusted estimates and annual current price estimates, for the period between 1953 and 1982. This is the usual database used by Statistics Canada to analyze revisions.

(b) United States: quarterly seasonally-adjusted estimates for the period between 1968 and 1983. In this case also. this is the base normally used by the Bureau of Economic Analysis (BEA) to analyze revisions; however, contrary to the Canadian data, these do not coincide with the published data, because the BEA introduces adjustments and corrections to eliminate the effect of changing definitions.

(c) Italy: annual estimate series, for the period between 1961 and 1985 (we did not take into account quarterly estimates, because publication was started in 1976, but was discontinued in 1982). These are data published by the *Instituto Centrale di Statistics (ISTAT)*.

By taking into account available data, it is possible to carry out comparisons between Canada and the United States for the quarterly process, and between Canada and Italy for the annual process.

**Table 2**

Publication Plan for Quarterly and Annual Estimates of National Accounts Aggregates in Reference to Time $t$ in Canada (Current Revisions Only)[a]

| Period of Publication | | Period of Reference | | | | |
|---|---|---|---|---|---|---|
| | | Quarterly Data | | | | Yearly Data |
| Year | Month and Quarter | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | | | | | |
| | 4 | | | | | |
| | 5 | $_1p_{t.I}$ | | | | |
| | 6 | | | | | |
| | 7 | | | | | |
| | 8 | $_2p_{t.I}$ | $_1p_{t.II}$ | | | |
| | 9 | | | | | |
| | 10 | | | | | |
| | 11 | $_3p_{t.I}$ | $_2p_{t.II}$ | $_1p_{t.III}$ | | |
| | 12 | | | | | |
| $t+1$ | 1 | | | | | |
| | 2 | $_4p_{t.I}$ | $_3p_{t.II}$ | $_2p_{t.III}$ | $_1p_{t.IV}$ | $_1p_t$ |
| | 3 | | | | | |
| | II | $_5p_{t.I}$ | $_4p_{t.II}$ | $_3p_{t.III}$ | $_2p_{t.IV}$ | $_2p_t$ |
| | III | | | | | |
| | IV | | | | | |
| $t+2$ | I | | | | | |
| | II | $r_{t.I}$ | $r_{t.II}$ | $r_{t.III}$ | $r_{t.IV}$ | $r_t$ |
| | III | | | | | |
| | IV | | | | | |

[a] Series shown are those used to carry out the empirical analysis on the revisions of quarterly data.

In order to carry out comparative analyses between the processes used to revise quarterly estimates in Canada and the United States, it was necessary first to establish which of the U.S. series could be used in parallel with the Canadian series. By taking into account the temporal shift between successive estimates, the available data, and the characteristics of the revision process, we chose the series in Table 3. In this way, we have, for the two countries, and for each reference quarter, three estimates that will be identified by the same letter: $P$ is the first quarterly estimates series, $RT$ is the chosen series of quarterly revisions; and $A1$ is the first annual revisions series. To these three estimates, we added the final estimates; these coincide with the last published estimates, which we will identify with the letter $F$.

Subsequently, it seemed appropriate to carry out the following comparisons:

($P$,$RT$), which takes into account the effect of the quarterly revision (sub- annual). Nevertheless, it is important to point out that, in Canada, only the provisional data for the first three quarters of any year are subject to strictly quarterly revisions, since the revisions to

the fourth quarter data ($_2p_{t,iv}$ in Table 2) is made up of both a quarterly and an annual revision. Among other things, this results in a very obvious difference in the behaviour of the revisions (see Lefrançois, 1988). We took this characteristic into account simply by limiting the analysis to the first three quarters of every year.

($RT$,$A1$), which provides information on the contribution made by the annual benchmark to the quarterly process of data revision;

**Table 3**

Publication Plan of Quarterly and Annual Estimates of National Accounts Aggregates in the United States[a]

| Period of Publication | | Period of Reference | | | | |
|---|---|---|---|---|---|---|
| | | Quarterly Data | | | | Yearly Data |
| Year | Month and Quarter | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | $_0p_{t.I}$ | | | | |
| | 4 | $_1p_{t.I}$ | | | | |
| | 5 | $_2p_{t.I}$ | | | | |
| | 6 | $_3p_{t.I}$ | $_0p_{t.II}$ | | | |
| | 7 | | $_1p_{t.II}$ | | | |
| | 8 | | $_2p_{t.II}$ | | | |
| | 9 | | $_3p_{t.II}$ | $_0p_{t.III}$ | | |
| | 10 | | | $_1p_{t.III}$ | | |
| | 11 | | | $_2p_{t.III}$ | | |
| | 12 | | | $_3p_{t.III}$ | $_0p_{t.IV}$ | |
| $t+1$ | 1 | | | | $_1p_{t.IV}$ | |
| | 2 | | | | $_2p_{t.IV}$ | |
| | 3 | | | | $_3p_{t.IV}$ | |
| | II | | | | | |
| | III | $_4p_{t.I}$ | $_4p_{t.II}$ | $_4p_{t.III}$ | $_4p_{t.IV}$ | $_1p_t$ |
| | IV | | | | | |
| $t+2$ | I | | | | | |
| | II | | | | | |
| | III | $_5p_{t.I}$ | $_5p_{t.II}$ | $_5p_{t.III}$ | $_5p_{t.IV}$ | $_2p_t$ |
| | IV | | | | | |
| $t+3$ | I | | | | | |
| | II | | | | | |
| | III | $r_{t.I}$ | $r_{t.II}$ | $r_{t.III}$ | $r_{t.IV}$ | $r_t$ |
| | IV | | | | | |
| $t+5$ | IV | $_br'_{t.I}$ | $_br'_{t.II}$ | $_br'_{t.III}$ | $_br'_{t.IV}$ | $_br'_t$ |
| $t+10$ | IV | $_br''_{t.I}$ | $_br''_{t.II}$ | $_br''_{t.III}$ | $_br''_{t.IV}$ | $_br''_t$ |

[a] The series shown are those used to carry out the empirical analysis on the revisions of quarterly data.

**Table 4**

Publication Plan for Quarterly and Annual Estimates of National Accounts Aggregates in Italy
(Current Revisions Only)

| Period of Publication | | Period of Reference | | | | |
|---|---|---|---|---|---|---|
| | | Quarterly Data | | | | Yearly Data |
| Year | Month and Quarter | I | II | III | IV | |
| $t$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | | | | | |
| | 4 | | | | | |
| | 5 | | | | | |
| | 6 | $_1p_{t.I}$ | | | | |
| | 7 | | | | | |
| | 8 | | | | | |
| | 9 | | $_1p_{t.II}$ | | | |
| | 10 | | | | | |
| | 11 | | | | | |
| | 12 | | | $_1p_{t.III}$ | | |
| $t+1$ | 1 | | | | | |
| | 2 | | | | | |
| | 3 | $_2p_{t.I}$ | $_2p_{t.II}$ | $_2p_{t.III}$ | $_1p_{t.IV}$ | $_1p_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |
| $t+2$ | I | $_3p_{t.I}$ | $_3p_{t.II}$ | $_3p_{t.III}$ | $_2p_{t.IV}$ | $_2p_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |
| $t+3$ | I | $r_{t.I}$ | $r_{t.II}$ | $r_{t.III}$ | $r_{t.IV}$ | $r_t$ |
| | II | | | | | |
| | III | | | | | |
| | IV | | | | | |

($A1,F$), on the basis of which we were able to evaluate the effect of five-year benchmark estimates in the United States and compare them to those determined by the extraordinary revision process in Canada. In this respect, we should also point out that the comparison must be carried out with a great deal of caution, since the Canadian historical revision are mainly caused by changes in the definitions, and are not strictly comparable with the five-year data adjustment process carried out in the United States.

As far as the comparison between current revisions of annual data in Italy and Canada is concerned, we still have some minor problems, because in the two countries the data available describe a process of revision that takes place in two stages. We identified the three series of

estimates under consideration in the usual way, as follows, ${}_1p$, ${}_2p$ and $r$; in this case, the interesting comparisons are: $({}_1p,r)$, to evaluate the overall effect of the current revision process; $({}_2p,r)$, to verify how much of the revision is carried out in the second stage (and indirectly also to derive information on the first stage); and the simultaneous comparison $({}_1p,{}_2p,r)$ to evaluate the convergence of the series of estimates.

The analyses were carried out on the following four aggregates:

*PL* = Gross Domestic Product for Italy; Gross National Product for Canada and the United States.

*CF* = Final domestic consumption of households and private administrations for Italy; personal expenses for goods and services for Canada and the United States.

*CP* = Collective consumption of public administrations for Italy; Acquisition of goods and services by public administration for Canada and the United States.

*IN* = Gross Fixed Capital Formation.

It should be kept in mind that the differences between the reference accounting systems used (SNA for Canada and the United States, SEC for Italy) mean that the configuration of the aggregates is not completely homogeneous in the three countries.

The main results of the analysis of errors in estimates of the level of the aggregates are shown in Tables 5 to 8. More complete results are nevertheless available and we will refer to them as the occasion arises.

### 5.2 A Tentative Comparative Summary of the Results of the Analyses

A comparison of the three plans shown in Tables 2, 3, and 4 leads to the following conclusions.

**Table 5**

Indices of the Accuracy of Provisional Estimates of the Level of Current Price Aggregates - Canada

| Aggregates | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparison between $P$ and $RT$ (1953-83; $T = 89$) | | | | |
| *PL* | −.0018 | .0034 | .0042 | .0046 | .1567 |
| *CF* | −.0009 | .0025 | .0039 | .0040 | .0560 |
| *CP* | −.0041 | .0106 | .0193 | .0197 | .0425 |
| *IN* | −.0047 | .0133 | .0206 | .0211 | .0498 |
| | Comparison between $RT$ and $A1$ (1953-83; $T = 69$) | | | | |
| *PL* | −.0046 | .0064 | .0065 | .0080 | .3372 |
| *CF* | −.0054 | .0069 | .0086 | .0102 | .2877 |
| *CP* | −.0008 | .0176 | .0262 | .0262 | .0009 |
| *IN* | −.0075 | .0205 | .0270 | .0280 | .0721 |
| | Comparison between $A1$ and $F$ (1953-70; $T = 72$) | | | | |
| *PL* | −.0556 | .0556 | .0197 | .0590 | .8882 |
| *CF* | −.0617 | .0617 | .0192 | .0646 | .9110 |
| *CP* | −.0673 | .0727 | .0585 | .0892 | .5699 |
| *IN* | −.0228 | .0311 | .0326 | .0398 | .0328 |

**Table 6**

Indices of the Accuracy of Provisional Estimates of the Level of Current Price Aggregates Annual Data - Canada

| Aggregates | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparison between ${}_1p$ and $r$ (1953-83; $n = 20$) | | | | |
| *PL* | −.0084 | .0093 | .0058 | .0102 | .6720 |
| *CF* | −.0080 | .0086 | .0077 | .0112 | .5191 |
| *CP* | −.0069 | .0118 | .0136 | .0153 | .2052 |
| *IN* | −.0173 | .0183 | .0114 | .0208 | .6966 |
| | Comparison between ${}_2p$ and $r$ (1953-83; $n = 24$) | | | | |
| *PL* | −.0032 | .0041 | .0044 | .0054 | .3444 |
| *CF* | −.0029 | .0030 | .0048 | .0056 | .2619 |
| *CP* | −.0038 | .0063 | .0087 | .0096 | .1627 |
| *IN* | −.0040 | .0042 | .0075 | .0085 | .2274 |
| | Comparison between $r$ and $F$ (1953-70; $n = 18$) | | | | |
| *PL* | −.0464 | .0464 | .0188 | .0500 | .8588 |
| *CF* | −.0544 | .0544 | .0200 | .0580 | .8809 |
| *CP* | −.0523 | .0523 | .0348 | .0629 | .6928 |
| *IN* | −.0130 | .0147 | .0140 | .0192 | .4637 |

**Table 7**

Indices of the Accuracy of Provisional Estimates of the Level of Current Price Aggregates - United States

| Aggregates | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparison between $P$ and $RT$ (1968-83; $T = 64$) | | | | |
| *PL* | −.0017 | .0031 | .0040 | .0044 | .1459 |
| *CF* | −.0014 | .0041 | .0054 | .0055 | .0607 |
| *CP* | −.0013 | .0050 | .0064 | .0065 | .0433 |
| *IN* | −.0047 | .0104 | .0138 | .0146 | .1052 |
| | Comparison between $RT$ and $A1$ (1968-83; $T = 52$) | | | | |
| *PL* | −.0054 | .0064 | .0071 | .0089 | .3706 |
| *CF* | −.0049 | .0073 | .0078 | .0092 | .2854 |
| *CP* | .0001 | .0088 | .0109 | .0109 | .0002 |
| *IN* | −.0074 | .0176 | .0237 | .0249 | .0878 |
| | Comparison between $A1$ and $F$ (1968-83; $T = 52$) | | | | |
| *PL* | −.0093 | .0097 | .0070 | .0117 | .6411 |
| *CF* | −.0040 | .0070 | .0088 | .0096 | .1662 |
| *CP* | .0039 | .0070 | .0088 | .0096 | .1662 |
| *IN* | −.0408 | .0409 | .0253 | .0480 | .7222 |

**Table 8**

Indices of the Accuracy of Provisional Estimates of the Level of Current Price Aggregates
Annual Data - Italy

| Aggregates | $\bar{e}$ | $\bar{e}'$ | $s_e$ | $d_e$ | $U_e^b$ |
|---|---|---|---|---|---|
| | Comparison between ${}_1p$ and $r$ (1963-85; $n = 13$) | | | | |
| *PL* | −.0095 | .0095 | .0093 | .0133 | .5091 |
| *CF* | −.0067 | .0068 | .0057 | .0088 | .5781 |
| *CP* | −.0111 | .0122 | .0129 | .0170 | .4249 |
| *IN* | −.0062 | .0078 | .0091 | .0110 | .3141 |
| GR.1 | −.0032 | .0113 | .0140 | .0144 | .0489 |
| 2 | −.0253 | .0282 | .0283 | .0380 | .4443 |
| 3 | −.0107 | .0379 | .0462 | .0475 | .0508 |
| 4 | −.0133 | .0208 | .0271 | .0302 | .1937 |
| 5 | −.0056 | .0101 | .0113 | .0126 | .1994 |
| 6 | −.0044 | .0172 | .0206 | .0210 | .0429 |
| 7 | −.0143 | .0146 | .0150 | .0207 | .4750 |
| 8 | −.0061 | .0132 | .0170 | .0181 | .1121 |
| Total | −.0106 | .0106 | .0106 | .0150 | .4987 |
| | Comparison between ${}_2p$ and $r$ (1963-85; $n = 17$) | | | | |
| *PL* | −.0030 | .0037 | .0054 | .0062 | .2299 |
| *CF* | −.0015 | .0017 | .0024 | .0029 | .2900 |
| *CP* | −.0013 | .0049 | .0063 | .0064 | .0426 |
| *IN* | −.0021 | .0026 | .0052 | .0056 | .1445 |
| GR.1 | .0002 | .0026 | .0038 | .0038 | .0018 |
| 2 | −.0054 | .0096 | .0172 | .0180 | .0884 |
| 3 | −.0015 | .0096 | .0177 | .0177 | .0073 |
| 4 | −.0058 | .0060 | .0105 | .0120 | .2304 |
| 5 | .0008 | .0016 | .0026 | .0027 | .0954 |
| 6 | .0011 | .0055 | .0085 | .0085 | .0167 |
| 7 | −.0039 | .0069 | .0093 | .0101 | .1459 |
| 8 | −.0008 | .0008 | .0030 | .0031 | .0673 |
| Total | −.0026 | .0038 | .0058 | .0063 | .1648 |

Gr.1: Agriculture, food and beverage products, tobacco.
Gr.2: Textiles, clothing and shoes, skins and leather.
Gr.3: Wood and wood furniture.
Gr.4: Non-metal minerals, chemical products, extraction industries, energy and water.
Gr.5: Buildings and civil engineering.
Gr.6: Transportation and communications.
Gr.7: Other areas.
Gr.8: Building rentals.

The system used to produce and revise national accounts estimates in the United States seems to be much more structured and strongly oriented in favour of quarterly accounts; in this respect, it is important to emphasize that the preliminary estimates are published extremely fast. By the time Canada (and Italy) produce the first quarterly estimate, in the United States, this estimate has already been revised two or three times using a remarkable amount of additional direct information (Parker 1986). On the contrary, less importance is given to the annual estimate, which comes out with a delay of nine months.

On the other hand, speed seems to be a distinctive characteristic of Italian annual estimates, which come out during the first quarter of the next year; while in Canada, the first annual estimate is only available in the second quarter (the preliminary estimate published in the first quarter is a simple aggregation of quarterly data). However, this characteristic is not necessarily a good thing, and it is reasonable to think that the fast production time of the initial annual estimates can also be explained on the basis of gaps present in the Italian quarterly accounts system.

Finally, it is important to emphasize that the United States is the only country that carries out not only regular annual benchmark revisions for quarterly estimates, but also more long-term benchmark revisions based on input-output tables, and the results of wider surveys, for example, a census. This ensures greater coherence (and not only from an accounting point of view) in the estimates of the various aggregates. In Canada and in Italy there are similar revisions; however, these are only carried out occasionally and are almost always accompanied by modifications in the classification or evaluation criteria.

On the basis of Tables 5 to 8, we are able to make the following observations:

(a) In the Canada-U.S. comparison, the pattern of current revisions of quarterly estimates is substantially homogeneous, at least as far as the aggregates in question and the limits of the revision process described by the three available series of estimates are concerned. On the average, sub-annual revisions are rather modest (in both cases the *IN* aggregate is an exception) and their behaviour is rather irregular, even though the trend is to revise upwards. The effect of the annual benchmark revision seems to be relatively more marked and systematic, especially in terms of the *PL*. There is also a tendency to correct the previous evaluation upwards. In this second stage of the process of revision, aggregate *IN* also shows higher and more irregular size errors.

(b) In relation to annual estimates, we also found substantial behavioural similarities in the Italy-Canada comparison. On the average, the preliminary *PL* estimate underestimates the final current data by about 0.9%; this trend is very systematic and has a sufficiently stable order so that ${}_1p \leq {}_2p \leq r$. This leads to an underestimation of the rates of variation when they are calculated (as is generally done) on the basis of "horizontal" comparisons. We should add that the weight of the second revision is relatively more marked for Canada.

On the whole, the empirical evidence shows that the provisional estimates are valid, especially if we look at quarterly evaluations in Canada and the United States. On the other hand, the systematic character of the underestimations of preliminary annual estimates, both in Canada and in Italy, raises certain questions (even though the size of the error is certainly not significant).

This phenomenon is widespread in many countries (see Glejser and Schavey 1974). There are nevertheless some aspects of the Italian case that deserve further consideration. The first are the extreme differences in the behaviour of errors at the disaggregated level by area of economic activity, which points to situations that are certainly weak from the point of view of information: in some areas, there is a systematic underevaluation of 6-7% on the average (Marliani 1986).

The other aspect is related to the marked re-evaluation effect due to the extraordinary revisions, which generally affects all aggregates (see Di Fonzo, Rettore, and Trivellato 1986). The trend to revise the previous evaluations upwards when introducing statistical modifications in the basic statistical surveys and/or the estimation methods, is certainly not unique to Italy. The Canadian data (and indirectly, the U.S. data, through the effects of the five- year benchmark adjustments) show a similar trend.

These are signals that emphasize the weakness of national accounts systems, and lead us to consider with less optimism the small average size of the errors in provisional estimates generated on the basis of the currently used revision process.

## ACKNOWLEDGEMENTS

The results discussed in this paper were obtained within the framework of a research project financed by the Italian Ministry of Education. The data were generously provided by the Current Economic Analysis Section of Statistics Canada, for Canada; by Robert P. Parker, Assistant Director for National Accounts, BEA, for the United States; and by ISTAT, for Italy. We would like to thank Silvano Bordignon, Tomasso Di Fónzo, Gianni Marliani, and Enrico Rettore (who form part of the research team), and the two critics who, through their comments and advise, greatly improved the quality of this study. We would also like to, acknowledge the help of Cristina Pozzato and Claudio Palmieri for their help with the empirical analyses, and especially that of Gianni Marliani, who allowed us to use some of the results of his recent work (Marliani 1987). An initial version of this paper was presented at the Journées de Statistique of ASU, which were held at Lausanne from 18 to 20 May, 1987.

## REFERENCES

BIGGERI, L. (1984). Caratteristiche e analisi dei processi di revisione nelle valutazioni di aggregati ed indici economici: alcuni confronti internazionali. *Note Economiche*, 4, 18-63.

BORDIGNON, S., and TRIVELLATO, U. (1989). On the optimal use of provisional data in forecasting with dynamic models. *Journal of Business & Economic Statistics*, 7, 275-286.

BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

DI FONZO, T., RETTORE, E., and TRIVELLATO, U. (1986). L'accuratezza delle stime provvisorie degli aggregati di contabilità nazionale annuale. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 113-145.

GLEJSER, H., and SCHAVEY, P. (1974). An analysis of revisions on national accounts data for 40 countries. *The Review of Income and Wealth*, 20, 317-332.

HATANAKA, M. (1974). A simple suggestion to improve the Mincer-Zarnowitz criterion for the evaluation of forecasts. *Annals of Economic and Social Measurement*, 3, 521-524.

HEMPENIUS, A.L. (1980). *Forecast accuracy analysis applied to forecasts of the Dutch Central Planning Bureau, 1964-1978*. Katholieke Hogeschool Tilburg, Subfaculteit der Econometrie.

KENDALL, M.G. (1973). *Time Series*. London: Griffin.

LEFRANÇOIS, B. (1988). Application des séries chronologiques à l'étude des révisions. *The Canadian Journal of Statistics*, 16, 83-96.

MALINVAUD, E. (1969). *Méthodes statistiques de l'économétrie*. Paris: Dunod.

MANKIW, N.G., RUNKLE, D.E., and SHAPIRO, M.D. (1984). Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics*, 14, 15-27.

MANKIW, N.G., and SHAPIRO, M.D. (1986). News or noise? An analysis of GNP revisions. *Survey of Current Business*, 66, 20-25.

MARAVALL, A., and PIERCE, D.A. (1983). Preliminary data error and monetary aggregate targeting. *Journal of Business & Economic Statistics*, 1, 179-186.

MARLIANI, G. (1986). L'accuratezza delle stime provvisorie del valore aggiunto per settore di attività economica. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 113-146.

MARLIANI, G. (1987). Stime provvisorie e revisioni dei dati di contabilità nazionale: il caso italiano ed alcuni confronti internazionali. *Attendibilità e tempestività delle stime di contabilità nazionale* (Ed. U. Trivellato). Padova: Cleup, 35-78.

McNEES, S.K. (1986). Estimating GNP: The trade-off between timeliness and accuracy. *New England Economic Review*, January/February, 3-10.

MINCER, J., and ZARNOWITZ, V. (1969). The evaluation of economic forecasts. *Economic Forecasts and Expectations* (Ed. J. Mincer). New York: NBER, Columbia University Press, 3-46.

MORK, K.A. (1987). Ain't behavin': Forecast errors and measurement errors in early GNP estimates. *Journal of Business & Economic Statistics*, 5, 165-175.

NOVAK, G.J. (1975). Reliability criteria for national accounts. *The Review of Income and Wealth*, 21, 323-344.

PARKER, P.R. (1986). Revisioni nelle stime iniziali del prodotto nazionale lordo trimestrale degli Stati Uniti. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 181-217.

PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.

RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Estimation of level and change using current preliminary data. *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley, 457-479.

THEIL, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.

TRIVELLATO, U. (1986a) (Ed.). *Errori nei dati preliminari, previsioni e politiche economiche*. Padova: Cleup.

TRIVELLATO, U. (1986b). Sulla valutazione dell'accuratezza di stime provvisorie di aggregati economici. *Studi in onore di Silvio Vianelli*. Palermo: Università degli Studi, 1587-1620.

TRIVELLATO, U. (1987). Problemi e metodi di valutazione dell'attendibilità delle stime di contabilità nazionale. *Statistica*, 47, 365-388.

TRIVELLATO, U., DI FONZO, T., and RETTORE, E. (1986). L'accuratezza delle stime provvisorie di aggregati e di indici economici: orientamenti metodologici. *Errori nei Dati Preliminari, Previsioni e Politiche Economiche* (Ed. U. Trivellato). Padova: Cleup, 61-86.

TRIVELLATO, U., and RETTORE, E. (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equation models. *Journal of Business & Economic Statistics*, 4, 445-453.

WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11. *Journal of the Royal Statistical Society*, series A, 145, 74-85.

WILTON, D.A., and SMITH, P.M. (1974). The efficiency of the G.N.P. revision process: An historical analysis. Current Economic Analaysis Division, Statistics Canada, Ottawa (internal paper).

ZARNOWITZ, V. (1982). On functions, quality, and timeliness of economic information. *Journal of Business*, 55, 87-119.

# GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

## 1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.

1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.

1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.

1.4 Acknowledgements should appear at the end of the text.

1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.

3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.

3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.

3.4 Write fractions in the text using a solidus.

3.5 Distinguish between ambiguous characters, (e.g., w, $\omega$; o, O, 0; l, 1).

3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.

4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).

5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.