Statistics Canada Statistique Canada

SURVEY METHODOLOGY

1982

Volume 8

Numbers 1 & 2

A Journal produced by
Methodology Staff
Statistics Canada

Canada

SURVEY METHODOLOGY

1982 Vol. 8 Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada

CONTENTS

8-3200-501
Reference No.
Z - 079

ISSN: 0714-0045

SURVEY METHODOLOGY

1982 Vol. 8 Nos. 1 & 2

A Journal produced by Methodology Staff, Statistics Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. 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 Boad or the Department.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.

THE ROLE OF THE QUESTIONNAIRE IN SURVEY DESIGN

R. Platek and D. Royce [1]

The modern statistical survey is an effective method of meeting the ever-increasing demand for timely and accurate data. One important component of the statistical survey is the questionnaire. This article discusses the role of the questionnaire in meeting the needs of users, the relationship of the questionnaire to the other components of survey design, and the effect of the questionnaire on the quality of survey data. The importance of viewing the questionnaire as an integral part of the total survey design is stressed.

1. INTRODUCTION

The escalating demand for appropriate and timely information of various kinds and from various sources calls for an organized approach to the entire process of data collection. The past forty years have seen the emergence of the statistical survey as an important tool to meet this need.

One important component of the statistical survey is the questionnaire. In the sections which follow, we describe the role of the questionnaire in meeting information needs, the relationship of the questionnaire to the other components of survey design, and the effect of the questionnaire on the quality of survey data. Although the discussion is presented mainly in the context of the household survey conducted by personal interview, many of the comments are relevant to questionnaires and surveys of all types.

2. INFORMATION NEEDS AND THE ROLE OF THE QUESTIONNAIRE

The simplest definition of a questionnaire is that of a group or sequence of questions designed to elicit information upon a subject from a respondent. Within the range of techniques in questioning, the questionnaire may range from a list of undefined topics to a highly structured set of questions with no options for response other than those listed.

[1] R. Platek and D. Royce, Census and Household Survey Methods Division Statistics Canada.

The questionnaire plays a central role in a complex process (the interview) in which information is transferred from those who have it (the respondents) to those who need it (the users). The questionnaire is the means through which the information needs of the users are expressed in operational terms which can be presented to a respondent in such a way that he will supply the required information. For this transfer of information to be effective, the questionnaire must meet the requirements of both users and respondents.

The expression of information needs, which a user may initially only vaguely understand, in terms suitable to the respondent is not something that can be accomplished in one step. Instead, the questionnaire design evolves and is refined as part of the overall survey development process.

For example, the user may begin with a need for information on "the housing conditions of the poor". He develops this into survey objectives by asking questions such as:

(a) What is the problem we are trying to solve?

(b) What specific items of information are needed?

(c) How will the information be used?

(d) How accurate and timely does the information have to be?

In answering these questions, his thinking becomes more quantitative, and he expresses his information needs in terms of specific survey concepts. The survey concepts describe both what is to be measured and the units for which measurements are required. He may describe "housing conditions" in terms of the number of rooms, the presence of plumbing and electricity, or the state of repair of the dwelling. He may define "the poor" in terms of income level or in terms of assets and debts.

It is important to emphasize that specific question wording is not at issue in the development of survey concepts. The first step for the user in expressing

his information needs is to decide what should be measured, not how it will be measured. The user should choose the concepts based on their relevance to his information needs. He should consider, for example, what concepts are most appropriate for the uses to be made of the data and whether the concepts are compatible with other sources of information.

Once information needs have been expressed in terms of specific survey concepts, the questionnaire becomes the instrument by which these concepts are measured. Through specific questions and accompanying instructions, the user specifies precisely how the survey concepts are to be measured in operational terms. Several questions may be required to measure complex concepts. In the Canadian Labour Force Survey, for example, as many as ten questions are needed to measure the concept "unemployed".

The questionnaire often serves as the document for recording of measurements as well. This is mainly of benefit to the interviewer or respondent, since it is convenient to record the answers immediately following the question. In theory, however, there is no reason why the questions and answers cannot be on two separate forms.

In the more structured types of surveys, the questionnaire is an important method of standardizing and controlling the data collection process. In statistical surveys, in contrast to other methods of investigation, the researcher usually cannot do his own data collection but must rely on interviewers hired for the job. Without specific question wordings and instructions to follow, interviewers would inevitably change the meaning or emphasis of questions and quite possibly the responses. The questionnaire helps ensure that the researcher measures what he wishes to measure with every respondent. It is, in effect, a "program" for the interviewer and respondent to follow in order to produce the desired result.

The questionnaire cannot be too rigid, however. It must be flexible enough to adapt to respondents of different age/sex groups, languages and social backgrounds. Different words or groups of words may be needed in order to convey the desired meaning to all respondents. The questionnaire must also

anticipate all of the possible answers that could be given. This is especially true in the initial, exploratory stages of research where an unstructured collection of data may be the most appropriate approach.

It must be recognized that the questionnaire is a complex and often imprecise measuring instrument. The subjects of measurement are human beings, and the process of measurement is based on language. As well as being a measuring instrument, the questionnaire is also a form of communication involving the researcher, the interviewer and the respondent. It transmits a request for information to the respondent, and it transmits the respondent's answer back to the researcher in a form useable to him. Warren Weaver, in The Mathematical Theory of Communication (1949), identifies three problems that must be faced in the design of any communication system:

A. How accurately can the symbols of communications be transmitted? (The technical problem).

B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem).

C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

All three problems are directly relevant to the construction of questionnaires, and all three problems are closely linked. Within the context of statistical surveys, the way in which the questionnaire solves these problems plays a major role in determining how well the information needs of the user are met.

3. THE QUESTIONNAIRE AND THE COMPONENTS OF SURVEY DESIGN

The process of making the survey concepts operational in a specific document forces the researcher to consider not only question wording, sequencing and layout, but nearly every other aspect of the survey as well. The questionnaire design must take into account elements such as the type of population

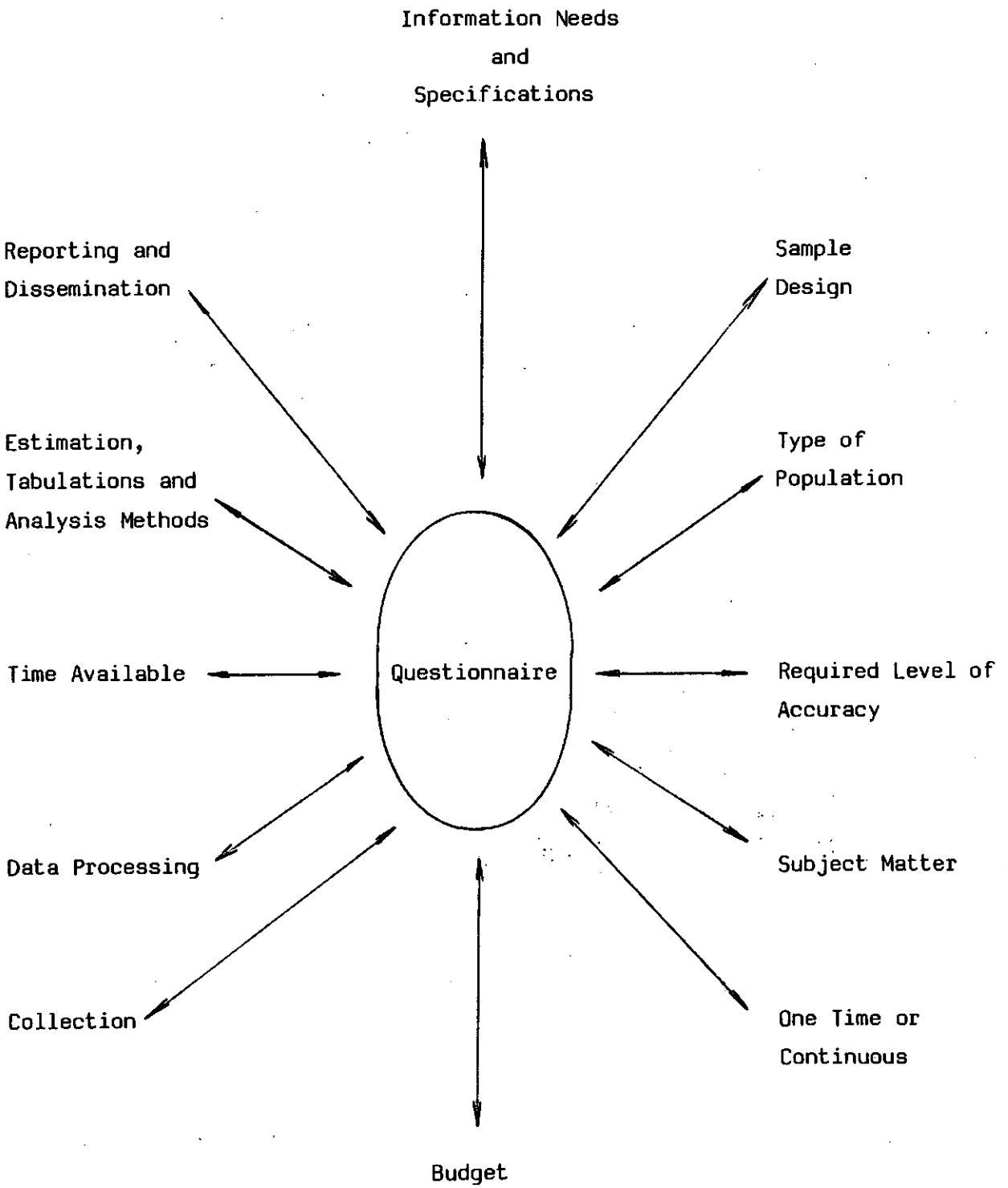
being surveyed, the sample design and sample size, the subject matter of the survey, the interviewing method, the data processing techniques to be used, and the budget and time available.

Figure 1 illustrates the questionnaire's relationships to some of the other elements which make up the total survey design. These interrelationships form a complex network; changes to one component of the design often require changes in several other components as well. Virtually any component of survey design could be placed at the centre of this network, but for the purpose of discussion we have chosen to focus attention on the questionnaire.

Elements such as the type of population, the sample design and the required level of accuracy are closely interrelated with questionnaire design. For example, the heterogeneous nature of many survey populations results in a need for cross-classified data. These needs affect the sample size, the type and degree of stratification, and the reliability of the information. This in turn will affect the questionnaire through the types of questions asked and the level of detail requested. This will further have an effect on the cost and timeliness of the information, the amount of respondent burden, and so on.

The questionnaire design is closely linked to the method of data collection and the survey's subject matter. Each method of data collection, such as personal interviewing, telephone interviewing and mail surveys, creates its own survey conditions which may be more or less appropriate to a given subject matter. These conditions will in turn affect the questionnaire's style of questioning, content, format, length and so on. In personal interviews, for example, it is often possible for the interviewer to collect certain data, such as type of dwelling and sex of respondent, by direct observation rather than questions. In addition, the questionnaire can be designed for the use of flash cards or other visual aids by the interviewer. The element of face-to-face communication is also a powerful motivating factor for the respondent. A personal interview is often the only choice when a complex, long and demanding questionnaire is involved. In telephone interviews, much of the social interaction between interviewer and respondent is lost and the respondent's

Figure 1: Elements Affecting the Questionnaire



co-operation may be affected. The questionnaire must rely entirely on verbal communication for its success, and the subject matter may have to be less demanding. However, with certain sensitive surveys, (e.g. criminal victimization surveys), the extra distance between interviewer and respondent may actually make it easier to answer questions. In mail surveys, the questionnaire itself assumes the role of interviewer. It must introduce the survey, motivate the respondent to co-operate and guide the respondent in completing the interview. It is a particularly demanding role which must be taken into account in designing the questionnaire.

Whether the survey is one-time or continuing also has an effect on questionnaire design. With a continuing survey, there is often more scope for learning from experience and refining the questionnaire over time. Experiments in question wording, programs to monitor response errors, and other methods of evaluating and improving the questionnaire design may only be feasible with a continuing survey. However, the ability to improve a questionnaire must be balanced against the disadvantages of change: for example the inability to make comparisons over time, the necessity to retrain interviewers, and the necessity to change expensive computer software.

In many continuing surveys, such as the Canadian Labour Force Survey, the same respondents are interviewed several times. The questionnaire must take into account the total response burden during the respondent's stay in the survey. The questionnaire may also have to adapt to different collection methods: for example in the LFS the first interview is conducted in person while in urban areas most subsequent interviews are conducted by telephone. Questionnaires designed for continuing surveys must be developed with the longer term view in mind.

The questionnaire is also interrelated with data processing and budgetary concerns. The format of questions, for example open or closed, has direct implications for operations such as coding, data capture, editing and tabulations. The presence of many open-ended questions increases the time and effort during coding operations, and the programs to edit and tabulate the data become more difficult and costly to write and test.

The questionnaire as an operational expression of user needs thus involves the total survey design itself. Survey design is a combination of intricate components, among which the questionnaire plays a central role. The questionnaire neither determines the form of the other components, nor is its form determined by the others. The process of questionnaire design must flow from and be a part of the total survey design process.

4. THE QUESTIONNAIRE AND ERRORS

All survey-taking is subject to errors from various sources, and in recent years non-sampling error has received increasing attention as a major component of the total survey error (see, for example, Anderson et al (1979), Bailar (1976), Hansen, Hurwitz and Bershad (1961), Koch (1973), and Platek and Singh (1980)). The control of non-sampling errors is an integral and vital part of survey design, requiring specific programs for the diagnosis, measurement and prevention of errors. Further, each program will have its own costs and benefits which must be taken into account in the design of controls (Platek and Singh (1980)).

The questionnaire is both an important source of non-sampling errors, and an important part of programs for their prevention and measurement. The scientific development of data collection has lagged behind that of sample design and estimation; improvements in sampling techniques often deal in fractions of a percent while experiments in question wording may reveal variations of 20 percent or more (Payne (1951)). This section discusses the relationship of the questionnaire to a few of the more important sources of non-sampling errors and illustrates the role of the questionnaire in minimizing these errors.

4.1 Non-response errors

Non-response is one important source of non-sampling error. If the characteristics of interest differ from respondents to non-respondents, bias will almost certainly be introduced into the results. Non-response is basically of two types: the "no contact" type, (e.g. no one home, temporarily absent, bad

weather, etc.) and the "refusal" type. The latter may be either a complete non-response or only non-response to some questions. The questionnaire can do little to eliminate the "no contact" type of non-response but it does play an important role in preventing the refusal.

To understand how the questionnaire does this, it is important to first understand why respondents do or do not respond. Many different psychological forces motivate people to respond to surveys, including an interest in the topic, a desire to be helpful, a belief in the importance of the survey, a feeling of duty, or even a belief in their own importance. Other forces influence people to refuse: for example difficulty in understanding questions, fear of strangers, the feeling of one's time being wasted, difficulty in recalling information, and embarrassing or personal questions. All of these forces will have an effect on the questionnaire design through the way in which survey topics are introduced, the question wording, the questionnaire's appearance and length, assurances of confidentiality, and so on. At the same time, these forces interact with the survey's subject matter, the type of population and the data collection method, which in turn influence the design of the questionnaire.

One must also consider the ability of respondents to respond. Unrealistic demands on the respondent's knowledge or memory, the use of overly difficult and technical language, or excessive demands on the respondent's patience are all sources of non-response which have their roots in the questionnaire. It must be said, however, that the patience of respondents often amazes even hardened survey designers. Chinnappa and Wills (1978) describe an interesting study of non-response to the physical measures component of the Canada Health Survey, where respondents were asked to submit to blood pressure tests, skinfold measurements, exercise tests, and were even asked to donate blood samples.

A more thorough discussion of the causes and treatments of non-response is given in Platek (1980).

4.2 Response errors

Response errors are a second category of non-sampling errors to which the questionnaire is closely related. Response errors can occur anywhere during the question-answer-recording process, and may be either systematic (response bias) or random (response variance).

Questions on sensitive topics, such as amounts and sources of income, use of alcohol and tobacco, illegal activities or mental illness are subject to large response errors. It is often felt, for example, that the respondent may distort the answer to avoid embarrassment or to appear to conform to societal norms (Warwick and Lininger (1975)). Many questionnaire design techniques have been devised to counter this "social desirability bias", including the anonymous questionnaire, the use of projective questionning techniques,[1] or randomized response techniques in which the respondent chooses which of two (or more) questions he answers by the random choice. However, in a recent study which compared questionnaire responses to external criterion information (e.g. official records or test results), Marquis et al (1981) found, rather surprisingly, that for most items which they studied the response bias was almost negligible, but that the response variance was quite large. This conclusion, if supported by other studies, indicates that measuring and reducing response variance may also be important in sensitive topic surveys. This might involve techniques such as reinterviews, internal consistency checks during the interview, and the collection of other information correlated with the variables of interest. This kind of emphasis has direct implications for questionnaire design.

Questions which depend on the respondent to remember events, such as the taking of a trip or the occurrence of a crime, are another source of response errors. Events may be forgotten, or events which occurred before the reference period may be incorrectly included. Bushery (1981), in an experiment

[1] An example of projective questionning might be the sequence:
1. What do you think most people feel about smoking marijuana?
2. How do you yourself feel about it?
The first question asks for the respondent's view of the societal norm and the second asks for his own view.

with the U.S. National Crime Survey, found that victimization rates with a 3-month reference period were much higher than those reported under a 6-month reference period, which were in turn higher than the victimization rates reported with a 12-month reference period. The bias due to recall loss with the longer reference periods was a much more serious source of error than sampling variability. The choice of an appropriate reference period for questions involving recall has been examined in a number of different subject matter areas (Sudman (1980), National Center for Health Statistics (1972)). Bounded recall, where respondents are interviewed at the end of the reference period, or the use of prominent dates (e.g. Christmas) and calendar aids to jog respondents' memories have shown to be of some value in reducing under-reporting (Neter and Waksberg (1965), Ashraf (1975)). With some topics, however, the only possible way to collect the information is to make the questionnaire into a form of diary, where the respondent records the event during, or shortly after, it happens. Questionnaires of this type are used for the Food Expenditure Survey and the Fuel Consumption Survey of Statistics Canada.

Although questions demanding recall and sensitive topics are important sources of response errors, there are many other causes. For example, an important component of response error is that due to the interviewer, the so-called correlated response error. Each interviewer exerts, to some degree, a common influence on all of the respondents in his/her assignment through the way in which the questions are asked, they way in which the respondent's replies are interpreted and recorded, and so on. The contribution of this component of error to the total survey error is directly related to the size of the interviewer's assignment. In telephone surveys, which may have quite large assignments, the correlated component can be a much more serious error than in personal interviews (Groves and Kahn (1979)). In turn, the correlated response error is more serious in personal interviews than in mail surveys or other surveys of the "self enumeration" type. This consideration was a major reason why the Census of Population and Housing has adopted the drop-off-mailback as the standard technique since 1971. The choice of data collection method in turn has a direct influence on the questionnaire design.

Numerous other examples of response errors could be given. They depend on what question is asked, how the interviewer asks it, the way in which the respondent interprets and answers the question, and the way in which the interviewer interprets and records the answer. The interview is a dynamic, interactive process of communication between interviewer and respondent. How it is handled determines whether or not the interview produces the desired information in an accurate and efficient fashion. In the heat of the interview, it is the questionnaire, through its content, question wording, instructions and layout, which must play the major role in controlling the situation.

4.3 Data processing errors

Once the interview is completed, the questionnaire becomes primarily a data processing document. Errors can occur at all phases of processing including coding, data capture, editing, imputation, estimation and tabulation. The way in which the questionnaire was designed will have a significant impact on the number and type of errors at this stage of the survey.

By including data capture codes right on the questionnaire, for example, data capture errors are usually reduced significantly. The data are captured directly from the questionnaire without first being transcribed onto another form. A step beyond this is the Computer Assisted Telephone Interview. The questionnaire is stored in a computer program, which controls the entire interview process. The questions appear one at a time on a video display terminal in front of the interviewer, who then asks the question and types the respondent's reply directly into the computer. The data can be edited immediately and errors corrected while the respondent is still on the telephone. The process also reduces the incidence of questions missed or of incorrect application of skip instructions.

Editing and imputation errors are also closely related to the questionnaire design. Problems of missing or inconsistent data can often be traced back to faulty questionnaire design. The ability to reconstruct or impute for missing values often depends on what concomitant variables were included on the questionnaire and what kind of fail-safe mechanisms were built in. For example,

in a survey which requests information on several detailed components of income, cases where the information is not given or is incorrect can often be salvaged by including a question asking for total income.

Non-response errors, response errors, and data processing errors are a few of the non-sampling errors which are closely linked to the questionnaire and to the other components of the overall survey design. The questionnaire is inevitably a cause of non-sampling error, but it must also go as far as possible in preventing errors. The degree to which the questionnaire succeeds at this task depends largely on the survey designer's knowledge of the various sources of errors and on his skill in integrating the design of the questionnaire with that of the entire survey. Each new survey may present new problems and pitfalls and as such they must be anticipated and taken into account in developing questionnaires.

5. CONCLUSION

The preceding sections have illustrated the questionnaire's role as both an expression of the user's information needs and as an important determinant of the quality of survey data. In both roles, the questionnaire is closely linked to all of the components of survey design. The total survey design, and in particular the questionnaire, must try to maximize both the relevance of the data to the user and the accuracy of the data. Successful questionnaire design incorporates both; we must ask the right question, and we must ask it in the right way.

It is important to underline that users' needs and the requirements of accuracy often conflict. The process of questionnaire design involves tradeoffs. A user may have to ask a simpler question than he would like simply for the respondent to be capable of answering. On the other hand, the questionnaire designer should not avoid asking complex questions simply because the answers may contain errors.

Questionnaire development is not simply a laboratory process. Although guidelines exist and research is possible, the skill of questionnaire design is learned to a large extent by practical experience and by trial and error. It is learned through discussions with users, interviewers and respondents. Questionnaire design is undoubtedly an interactive process which cannot be carried out in isolation and independent of other factors in survey development. It interrelates with them and, in fact, it forms an integral part of the total survey design.

REFERENCES

[1] Andersen, R., Kasper, J., Frankel, M.R. and Associates (1979), Total Survey Error, Jossey-Bass Publishers, San Francisco.

[2] Ashraf, A. (1975), "The Methodology of the Canadian Travel Survey", Survey Methodology, Vol. 1, No. 2, 208-227.

[3] Bailar, B. (1976), "Some Sources of Error and Their Effect on Census Statistics", Demography, Vol. 13, No. 2, 273-286.

[4] Bushery, J.M. (1981), "Recall Biases for Different Reference Periods in the National Crime Survey", paper presented at the 141st Annual Meeting of the American Statistical Association, Detroit.

[5] Carson, E.M. (1973), Questionnaire Design: Some Principles and Related Topics, Statistics Canada internal report.

[6] Chinnappa, B.N. and Wills, B. (1978), "A Study of Refusal Rates to the Physical Measures Component of the Canada Health Survey", Survey Methodology, Vol. 4, No. 2, 100-114.

[7] Dillman, D.A. (1978), Mail and Telephone Surveys: The Total Design Method, Wiley, New York.

[8] Fellegi, I. (1979), "Data, Statistics, Information - Some Issues of the Canadian Social Statistics Scene", Survey Methodology, Vol. 5, No. 2, 130-161.

[9] Groves, R.M. and Kahn, R.L. (1979), Surveys by Telephone, Academic Press, New York.

[10] Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961), "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, Vol. 38, Part II, 359-374.

[11] Jabine, T.B. (1981), Guidelines and Recommendations for Experimental and Pilot Survey Activities in Connection with the Inter-American Household Survey Program, Inter-American Statistical Institute report 7679 a - 5/7/81 - 100, Washington, D.C.

[12] Kahn, R.L. and Cannell, C.F. (1957), The Dynamics of Interviewing, Wiley, New York.

[13] Koch, G. (1973), "An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means", Journal of the American Statistical Association, Vol. 68, No. 344, 906-913.

[14] Marquis, K.H., Marquis, M.S. and Polich, J.M. (1981), "Survey Response Errors for Sensitive Topics: The Problem is Noise Rather than Bias", paper presented at the 141st Annual Meeting of the American Statistical Association, Detroit.

[15] National Center for Health Statistics (1972), "Optimum Recall Period for Reporting Persons Injured in Motor Vehicle Accidents", Vital and Health Statistics, Series 2, No. 50, U.S. Government Printing Office, Washington, D.C.

[16] Neter, J. and Waksberg, J. (1965), Response Errors in Collection of Expenditures Data by Household Interviews: an Experimental Study, Bureau of the Census Technical Paper No. 11, U.S. Government Printing Office, Washington, D.C.

[17] Payne, S.L. (1951), The Art of Asking Questions, Princeton University Press, Princeton, N.J.

[18] Platek, R. (1980), "Causes of Incomplete Data, Adjustments and Effects", Survey Methodology, Vol. 6, No. 2, 93-132.

[19] Platek, R. and Singh, M.P. (1981), "Cost Benefit Analysis of Controls in Surveys", Current Topics in Survey Sampling, Academic Press, New York.

[20] Shannon, C.E. and Weaver, W. (1949), The Mathematical Theory of Communication, University of Illinois Press, Urbana.

[21] Statistics Canada (1979), Basic Questionnaire Design, Second edition, Ottawa.

[22] Sudman, S. (1980), "Reducing Response Errors in Surveys", Statistician, Vol. 29, 237-273.

[23] Warwick, D.P., and Lininger, C.A. (1975), The Sample Survey: Theory and Practice, McGraw - Hill, New York.

EVALUATION OF SMALL AREA ESTIMATION TECHNIQUES FOR THE CANADIAN LABOUR FORCE SURVEY [1]

J.D. Drew, M.P. Singh, G.H. Choudhry [2]

Estimates from sample surveys are sometimes required for domains whose boundaries do not coincide with those of design strata. Taking the Canadian Labour Force Survey as an example of a survey utilizing a clustered sample design, some alternative small area estimation techniques available in the literature are evaluated empirically including synthetic, domain (simple and post-stratified) and composite estimators which are linear combinations of synthetic and post-stratified domain estimators. A sample dependent estimator which attaches weight to the post-stratified domain estimate depending on the amount of sample in the domain is proposed and its performance is also evaluated.

1. INTRODUCTION

With increasing emphasis on planning, administering and monitoring social and fiscal programs at local levels, there has been demand for more and good quality data at these levels from various municipal, provincial and federal government departments as well as from private institutions. The type of data required ranges from simple population counts to complex socio-economic variables such as employment, unemployment, income, houseing, proverty indices, health conditions and facilities etc. However, until recently not much attention had been paid to the development of sound statistical estimation techniques for small area data, with the notable exception of statistical demographers who for some time have been investigating the particular problem of small area population estimates, and who have identified several competing methods based on the use of administrative data and other sources.

[1] Presented at the Annual meetings of the American Statistical Association, Cincinnati, August 1982

[2] J.D. Drew, M.P. Singh and G.H. Choudhry, Census and Household Survey Methods Division, Statistics Canada.

A comprehensive review of existing small area (domain) estimation techniques along with their limitations is given by Purcell and Kish (1979). From the research done to date it is clear that there is not a unique best solution to the small area estimation problem. The choice of a particular method for small area estimation will depend on the data needs and on the richness and availablility of data sources, which differ from country to country, and within countries from one subject matter to another. Therefore, the classification of the type of small areas (domains) and examination of the data sources available in a particular context, followed by thorough investigation of the alternative small area estimation techniques for given situations, seems to be the most appropriate approach to development of small area data. In this context, we shall use the following classification of domains suggested by Purcell and Kish (1979) and point out the type of domain to which developments in this articles primarily refer.

(a) Planned domains - for which separate samples have been planned, designed, and selected. In the Canadian context, such domains for example may be economic or planning regions within a province or the province itself.

(b) Cross Classes - which cut across the sample design and the sample units (may also be referred to as characteristic domains); e.g., age/sex, occupation, industry.

(c) Unplanned Domains - that have not been distinguished at the time of sample design and thus may cut across the design strata or the primary sampling units (PSU's) within the strata. Examples of these in the Canadian context include Federal Electoral Districts, and Census Divisions or subdivisions, counties and manpower planning regions.

It should be noted that both types (a) and (c) refer to areal domains.

We consider this distinction of the domains into the above types important since the form of the estimator as well as its efficiency would depend upon the particular type of application. As pointed out by Purcell and Kish most of the developments in small area estimation techniques in the United States

and elsewhere have concentrated on the domains of types (a) and (b). In Canada however, type (a) and (b) domains are not so problematic due to the type of design and the sizes of the national surveys, and the main emphasis has been on the data for the domains of type (c), with the possible exception of population counts using symptomatic data.

Investigations into the application and evaluation of small area estimation techniques for variables other than population started with the publication of synthetic estimates from the National Center for Health Statistics (1968). Since then a series of investigations (Gonzalez (1973), Gonzalez and Waksberg (1973), Schaible, Brock and Schnack (1977), Gonzalez and Hoza (1978) and others) have been carried out usingdata from the Current Population Survey in the application and evaluation of a particular synthetic estimator. Using a synthetic estimator whose form is different, studies were carried out by Purcell and Linacre (1976) aimed at production of estimates for Census divisions in Australia and by Ghangurde and Singh (1976, 1977, 1978) in the evaluation of synthetic estimates in the context of Canadian Labour Force Survey (LFS).

As remarked by Purcell and Kish (1979), the nature of the design in relation to the domains of interest has an important role to play in the choice of an estimator. The estimators considered in this article are geared to the Canadian LFS where the domains are unplanned domains (typec) and are of a size such that, had they been planned domains (type a), the reliability of regular unbiased survey estimates would be satisfactory without having to resort to small areas estimation techniques. Also in the LFS, primary sampling units are small (populations from 2,000 - 5,000) relative to the sizes of the domains of interest. This differs from the situation in the United States where the sizes of primary sampling units for most of the large scale surveys are larger, comparable in size to the small areas for which the estimates are desired.

In this article estimators are evaluated in the context of producing Census Division level estimates from the Labour Force Survey, using data from the 1971 and 1976 Censuses of Population and Housing in an auxiliary fashion. In

addition to synthetic estimators, we evaluate post-stratified domain estimators which were considered earlier by Singh and Tessier (1976), and composite estimators which are linear combinations of the synthetic and the post-stratified domain estimators, similar to those considered by Schaible (1979) and Schaible, Brock and Schnack (1977). Also we propose and evaluate a new estimator which we call a sample dependent estimator, which is of the same form as the composite estimator, except the weight given to the synthetic component is a decreasing function of the amount of sample falling into the domain upto a critical point after which the estimator relies totally on the post-stratified domain component. Efficiencies of the small area estimators relative to the direct (or simple domain) estimator for the characteristics employed and unemployed were obtained in an empirical (Monte Carlo) study in which the LFS design was simulated using census data. The situations where both the design and the auxiliary information are up-to-date and where both are out-of-date were considered. We have also evaluated the bias of synthetic estimators for the characteristics employed and unemployed for Federal Electoral Districts.

2. DESCRIPTION OF ESTIMATION PROCEDURE

Consider a finite population consisting of N units, (e.g. households or dwellings in household surveys), divided into L design strata labelled 1, 2, ..., h, ...L. The stratification has been carried out on the basis of geographic and/or certain socio-economic characteristics, and the sample allocation ensures certain precision for estimates from individual strata. The problem considered is that of estimating the total of an x-variate for all those unites belonging to an unplanned areal domain (type c). We denote by 'a' the set of units belonging to the small area or domain of interest, thus the parameter to be estimated is the total of the x-variable in the domain 'a', which we denote by ${}_aX$.

Let a_h be the set of those units belonging to the domain which are in stratum h, then

$$a = \bigcup_{h=1}^{L} a_h . \tag{2.1}$$

In practice the domain 'a' will have a non-null intersection with a certain number of design strata and if we denote by $\underset{\sim}{h}$ the set of such strata, then we have

$$a = \bigcup_{h \in \underset{\sim}{h}} a_h . \tag{2.2}$$

The particular design under consideration follows a multi-stage clustered sample design which is self-weighting within each stratum with weight W_h for stratum h.

For a particular given sample we can obtain the quantities:

t_h = sample total of x-variate in stratum h,

and

${}_a t_h$ = sample total of x-variate in a_h

for h=1, 2, ..., L. Note that ${}_a t_h = 0$ for $h \notin \underset{\sim}{h}$. Then the direct (or also referred to as design based or simple domain) estimator for the total of x-variate for those units in 'a' say ${}_a\hat{X}$, is given by:

$$\hat{X}_a = \sum_{h \in \underset{\sim}{h}} W_h \cdot {}_a t_h . \tag{2.3}$$

It should be noted that the direct estimator (2.3) does not utilize any auxiliary information - all it requires is the identification of those sampled units which belong to the domain. Due to the clustered nature of the design,

the sample falling in the domain may on occasion be very small or non-existent, generally resulting in high variance for this estimator.

The other estimators in this section rely in different fashions on auxiliary information for a variable y, which is often taken as the count of persons by population sub-groups (defined on the basis of age/sex etc.) from a recent census. These estimators are:

1) Post-stratified domain
2) Synthetic
3) Composite
4) Sample Dependent

Additionally estimators (2) - (4) rely to differing degrees on sample external to the domain.

For each of the above estimators, the adjustments based on the auxiliary information can be made either be applying separate adjustments to each stratum intersecting the domain, or by applying an overall adjustment for all strata intersecting the domain. Thus the estimators will be further classified as <u>separate</u> or <u>combined</u> depending on the level at which the adjustment is made. These estimators are denoted by ${}_a\hat{X}_{uv}$, where u is the level of adjustment with values:

u = s : separate
= c : combined

and v is the type of estimaror taking the following values:

v = p : post-stratified domain
= s : synthetic
= c : composite
= d : sample dependent

For example, ${}_a\hat{X}_{cs}$ denotes the <u>combined synthetic</u> estimator, etc.

2.1 Post-Stratified Domain Estimator

Define

Y_{hg}= total of the auxiliary y-variable for population sub-group g in group g in stratum h, and

${}_aY_{hg}$= total of the auxiliary y-variable for population sub-group g in a_h.

Further let ${}_a\tilde{Y}_{hg}$ be an unbiased estimated of ${}_aY_{hg}$ which would be formed analogously to the direct estimate defined in (2.1), except the characteristic being estimated in this case would be the auxiliary y-variable whose value is known for the set of sampled units (s) at some stage of sampling (whereas (2.1) is defined on the x-variate for the sample of ultimate units). In practice provided auxiliary y-variable information is available for them, sampling units at any stage down to the penultimate stage could be used.

Then the separate post-stratified domain estimator (for which adjustments are applied at the stratum level) is:

$$\hat{X}_{asp} = \sum_g \sum_{h \in \tilde{h}} (W_h \cdot {}_at_{hg}) \frac{{}_aY_{hg}}{{}_a\tilde{Y}_{hg}} \qquad (2.4)$$

where ${}_at_{hg}$ is the sample total of the x-variate for population sub-group g in the intersection of domain 'a' and stratum h.

Similarly the combined post-stratified domain estimator (for which adjustments are applied at the domain level) is:

$$\hat{X}_{acp} = \sum_{g} \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{\sum_{h \in \tilde{h}} {}_aY_{hg}}{\sum_{h \in \tilde{h}} {}_a\tilde{Y}_{hg}} \tag{2.5}$$

The post-stratified domain estimator is unbiased except for the effect of ratio estimation bias, provided ${}_aY_{hg}$ is obtained at the same time as ${}_a\tilde{Y}_{hg}$ and using the same source such as census.

Estimators of the above type have been considered earlier by Singh and Tessier (1976) with a different choice of post-strata.

2.2 Synthetic Estimators

We consider separate and combined synthetic estimators defined respectively as follows:

$$\hat{X}_{ass} = \sum_{g} \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{{}_aY_{hg}}{Y_{hg}} \tag{2.6}$$

$$\hat{X}_{acs} = \sum_{g} \sum_{h \in \tilde{h}} (W_h \cdot t_{hg}) \frac{\sum_{h \in \tilde{h}} {}_aY_{hg}}{\sum_{h \in \tilde{h}} Y_{hg}}, \tag{2.7}$$

where t_{hg} is the sample total for the x-variable for population sub-group g in stratum h.

The above synthetic estimator has been considred by Purcell and Linacre (1976) and also by Ghangurde and Singh, (1976, 1977, 1978) who developed expressions for its variance and bias and evaluated the estimator using census data and a super-population model. A different form of syntheticestimator was proposed earlier by the National Centre for Health Statistics (1968) and investigated by Gonzalez (1973), Gonzalez and Waksberg (1973) and Gonzalez and Hoza (1975, 1978) using data form the Current Population Survey.

The difference between the synthetic and post-stratified domain estimators can be readily seen by comparing (2.4) and (2.6). The post-stratified domain estimator uses only the sample falling into the domain (i.e., ${}_a t_{hg}$) and the adjustment factor is the ratio of the true to the estimated values for the y-variable for the domain and hence can take on values greater than or less than 1 (its expected value being unity). On the other hand the synthetic estimator uses the estimate from entire strata intersected by the domain (i.e. $w_h \cdot t_{hg}$ for $h \in \underset{\sim}{h}$) which is then deflated by adjustment factors specific to population subgroups. (i.e. the ratio of the y-variable for the domain to the y-variable for the entire stratum).

The synthetic estimator will suffer from bias depending on the degree of departure from the assumption of homogeneity for the x-variate between the domain and the larger area, namely $\underset{\sim}{h}$, withing sub-groups of the y-variable. In defining the above synthetic estimator, the larger area was restricted to those strata which form part of the domain as it was believed that such a choice would lead to less bias. In general however, $\underset{\sim}{h}$ need not be so restricted but it may include other neighbouring ares which are believed to satisfy the homogeneity assumption. Bias and mean square error of such estimators have been reported by some of the earlier referenced authors.

2.3 Composite Estimators

A composite estimator using the direct estimator and the synthetic estimator as the two components was suggested by Royall (1973) and others, and has been studied by Schaible (1978). Such an estimator minimises the chances of extreme situations (both in terms of bias and mean square error) and therefore may be preferred over either of its components. Synthetic estimators have a low variance by virtue of their use of data from a larger area to derive estimates for small are (domain), but for the same reason this introduces bias which could be quite large if as noted earlier, the assumption of homogeneity is not satisfied. On the other hand the simple domain estimator, which is unbiased, may have large variance particularly if the sample falling in the domain is very small. Empirical evidence of such relative performances of synthetic and direct estimators are available from Gozalez and Waksberg (1975)

Schaible, Brock and Schnack (1977), and Ghangurde and Singh (1977). The composite estimator considered here is obtained by replacing the direct estimator (2.3) by the post-stratified domain estimator which may be slightly biased but is generally more efficient than the direct estimator.

The two types of composite estimators: namely, separate and combined are formed as linear combinations of the corresponding post-stratified domain and synthetic estimators; viz,

$$ {}_a\hat{X}_{sc} = \alpha_1 \; {}_a\hat{X}_{sp} + (1 - \alpha_1) \; {}_a\hat{X}_{ss} \tag{2.8} $$

and

$$ {}_a\hat{X}_{cc} = \alpha_2 \; {}_a\hat{X}_{cp} + (1 - \alpha_2) \; {}_a\hat{X}_{cs} \tag{2.9} $$

The optimum values for α_1 and α_2 for minimum mse's are given by

$$ \overset{*}{\alpha}_1 = \frac{\text{mse} \left[{}_a\hat{X}_{ss}\right] - E \left[{}_a\hat{X}_{ss} - {}_aX\right] \left[{}_a\hat{X}_{sp} - {}_aX\right]}{\text{mse} \left[{}_a\hat{X}_{ss}\right] + \text{mse} \left[{}_a\hat{X}_{sp}\right] - 2\, E \left[{}_a\hat{X}_{ss} - {}_aX\right] \left[{}_a\hat{X}_{sp} - {}_aX\right]} \tag{2.10} $$

and a similar expression for $\overset{*}{\alpha}_2$.

Further, neglecting the covariance term in (2.10) under the assumption that this term will be small relative to mse $[{}_a\hat{X}_{ss}]$ and mse $[{}_a\hat{X}_{sp}]$, then the optimal weight $\overset{*}{\alpha}_1$ can be approximated by

$$ \overset{**}{\alpha}_1 = \frac{\text{mse} \left[{}_a\hat{X}_{ss}\right]}{\text{mse} \left[{}_a\hat{X}_{ss}\right] + \text{mse} \left[{}_a\hat{X}_{sp}\right]} \tag{2.11} $$

with a similar expression for $\overset{**}{\alpha}_2$, which was the approach to defining weights followed by Schaible (1978).

2.4 Sample Dependent Estimators

In practice the true values of α_1^* (or α^*_2) used as the weight in the composite estimator will not be available as they involve population variances and covariances, which would have to be estimated from the sample. Further calculation of the covariance term in (2.10), in particular, may be quite complex and thus one may have to resort to an approximate value α_1^{**} (or α_2^{**}) which would require simply the estimated mse's of the two component estimators or an estimate of the ratio of the two mse's. In either case there estimates would introduce a certain amount of instability in the weight used, thus affecting the performance of the composite estimator.

The sample dependent estimator (Drew and Choudhry, (1979)) which is a particular case of a composite estimator, depends on the outcome of the given sample and is quite simple to compute. It is constructed using the result that the performance of the post-stratified domain estimator depends upon the proportion of the sample falling in the domain. If the proportion of the sample within the domain is 'reasonably large' then the sample dependent estimator is the same as the post-stratified domain estimator, otherwise it becomes a composite estimator with gradual increasing reliance (in the sense of increasing weight) on the synthetic estimator as the size of the sample in the domain decreases. Thus the separate sample dependent estimator (i.e., constructed at the stratum level) is given by

$$\hat{X}_{asd} = \sum_g \sum_h \left[\delta_{hg} W_h \cdot t_{ahg} \frac{{}_a\tilde{Y}_{hg}}{{}_aY_{hg}} + (1 - \delta_{hg}) W_h \cdot t_{hg} \frac{{}_aY_{hg}}{{}_aY_{hg}} \right] \quad (2.12)$$

where

$$\delta_{hg} = 1, \text{ if } {}_a\tilde{Y}_{hg} / {}_aY_{hg} \geq K_o ,$$

$$= \frac{1}{K_o} \frac{{}_a\tilde{Y}_{hg}}{{}_aY_{hg}} , \qquad \text{otherwise.}$$

Similarly the combined sample dependent estimator (i.e. constructed at the domain level) is given by

$$\hat{X}_{acd} = \sum_g \Big[\delta_g \Big(\sum_{h \in \underset{\sim}{h}} W_h \cdot t_{ahg} \Big) \frac{\sum_{h \in \underset{\sim}{h}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg}} + (1 - \delta_g) \Big(\sum_{h \in \underset{\sim}{h}} W_h \cdot t_{hg} \Big) \frac{\sum_{h \in \underset{\sim}{h}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{h}} Y_{hg}} \Big] \qquad (2.13)$$

where

$$\delta_g = 1, \text{ if } \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg} \geq K_o$$

$$= \frac{1}{K_o} \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg}, \qquad \text{otherwise}$$

The ratios

$$\tilde{Y}_{ahg} / Y_{ahg} \quad \text{and} \quad \sum_{h \in \underset{\sim}{h}} \tilde{Y}_{ahg} \Big/ \sum_{h \in \underset{\sim}{h}} Y_{ahg}$$

indicate the over- or under-representation of the population sub-group at the individual stratum or domain level with respect to auxiliary information for the y-variable, conditional upon the selected sample.

Values of ratios greater than or equal to 1 signify that, conditional onthe given sample (s), the representation of the population sub-groups for the auxiliary y-variable is better than or as good as its unconditional representation had the domain been sampled independently at the same rate as the stratum.

The value of K_o may be appropriately chosen. In this study the efficiency of sample dependent estimator has been investigated for two specific values of K_o namely 1.0 and 0.5.

Holt, Smith and Tomberlin (1979) under the prediction approach derived an estimator (which relies on synthetic and direct estimates) where the weight attached to the direct component depends only on the sample falling into the domain. Sarndal (1981) proposed an alternative estimator in which the weight attached to the direct component depends on the sample in the domain relative to the sample in the larger area.

3. DESCRIPTION OF THE EMPIRICAL STUDY

3.1 Simulation of the LFS Design

The LFS follows a multi-stage area sampling design (see Platek and Singh, (1976)). Within each of the 10 provinces of Canada, two principal area types are identified - the Self-Representing Units (SRU's) which correspond to cities generally of 15,000 or more population, and the Non Self-Representing Units (NSRU's) which correspond to smaller urban centers and rural areas. In the SRU's, cities are divided into compact areal strata with populations of 15,000 each, within which a two stage sample of clusters (similar to blocks) and dwelling is selected.

In NSRU's, Economic Regions, of which there are from 1-10 per province, form the starting point. These are stratified into 1-5 strata with populations from 30,000 to 80,000 using census data for 7 broad industryclassifications. Within strata, primary sampling units (PSU's) from 2,000 - 5,000 in population are formed. The second stage in the rural portions of PSU's corresponds to 1971 Census Enumeration Areas (i.e., EA's), with populations of roughly 500, whereas in urban portions all urban centers are selected with certainty. The last two stages correspond to clusters and dwellings.

In simulating the LFS design two cases were examined: (i) the case whereboth the sample design and the auxiliary information are up-to-date, and(ii) the case where both are out-of-date.

For (i), the sample design, the auxiliary information, and the study variables were all based on 1971 census data. Counts of persons (15+) cross-classified by age/sex, and Labour Force status were retrieved at the EA level. In NSRU's, for each replication in the Monte Carlo study independent samples of primaries and secondaries were selected based on census population or dwelling counts. Within rural EA's and urban centers, the final two stages of sampling were simulated by random samples of persons. In SRU's, EA's comprising the areal strata were known, but there after the LFS design was independent of the census. Hence for the purposes of the study, EA's were randomly partitioned into 'clusters' having a size distribution corresponding to that for LFS clusters. For each replication, a sample of 'cluster' and a random sample of persons within were selected.

3.2 Choice of Population Sub-Groups

The estimators defined in section 2 utilize auxiliary information for population sub-groups. Since the LFS is redesigned only decennially, it would be desirable to base the population sub-groups on information collected in the mid-decade as well as decennial census, so that the auxiliary information could be updated mid-way through the life of the survey. This ruled out such variables as industry or occupation, leaving various cross-classifications of basic demographic variables as the possible choices for population sub-groups.

For the variables marital status, age and sex, the Automatic Interaction Detection (AID) procedure, due to Sonquist and Margan (1964) was used ona sample of census data from across Canada to derive optimal population sub-groups, separately for each Labour Force characteristic. Results ofthe AID analysis showed that for unemployed, no population sub-groups accounted for more than 2% of the variation, while for the characteristics employed and not in Labour Force the following sub-groups accounted for approximately 25% of the variation: (i) age 15-16 and 65+; (ii) age17-64, sex female; (iii) age 17-64, sex male. Further splitting of these sub-groups did not result in significant additional gains.

In addition to estimators based on the above population sub-groups, estimators based on total population 15+, and on dwelling counts were also considered. Dwelling count data were included due to the possibilities which exist for up-to-date dwelling information being available intercensally at the required level of detail. It might be noted that the estimators using population 15+ and dwelling counts are both special cases of the general formulation where the number of population sub-groups equals 1.

3.3 Evaluation of Efficiency of Small Area Estimators

In the Monte Carlo study, we have considered 16 Census Divisions (CD's) and 11 Federal Electoral Districts (FED's) in the province of Nova Scotia and 7 FED's from elsewhere in Canada. (There are altogether 18 CD's in the province of Nova Scotia, but two of 18 CD's correspond to complete LFS strata and therefore were omitted from the study). Due to the multi-stage nature of the design and larger number of domains in the study, the computational costs involved were high and it was decided to use only 100 replications.

Census Divisions and Federal Electoral Districts, it should be noted, comprise networks of geo-statistical and geo-political areas respectively across Canada. There are approximately 300 of each, with the populations of Federal Electoral Districts being fairly uniform in the range 80,000 to 120,000, while those of Census Divisions, which often correspond to local levels of government or counties, vary greatly.

We have reported results only for the 16 Census Divisions in Nova Scotia. Results were similar for other unplanned domains considered.

If we let ${}_a\hat{X}_{m(r)}$ be the estimate of total ${}_aX$ (i.e. the total for the x-variable for the domain 'a') for the r'th replicate, for small area estimation method m, then the average mean square error for the method m over the 16 domains in the study was calculated as:

$$\text{Avg mse } (m) = \frac{1}{16} \sum_{a} \sum_{r=1}^{100} ({}_a\hat{X}_{m(r)} - {}_aX)^2/100 \quad . \qquad (3.1)$$

The efficicency of the small area estimator (m) relative to the direct estimator, say method m_0 was obtained as:

$$\text{Eff } (m \text{ vs } m_0) = \frac{\text{Avg mse } (m_0)}{\text{Avg mse } (m).} \tag{3.2}$$

3.4 Evaluation of Bias of Synthetic Estimators

Since the composition of the LFS frame and the Federal Electoral Districts were known for all of Canada in terms of both 1971 and 1976 census units, it was possible to compute exact biases of the synthetic estimators based on census data. The following cases were considered: (i) design and auxiliary information up-to-date (in which case the design, adjustment factors and x-variables were all based on the 1971 census); and (ii) design and auxiliary information out-of-date (in which case the design and adjustment factors were based on the 1971 census, but the x-variables were based on the 1976 census).

Let ${}_aB_{ss}$ and ${}_aB_{cs}$ denote the biases of the separate and combined synthetic estimates for unplanned domain 'a', then we have

$$B_{a\,ss} = \sum_g \sum_{h \in \underset{\sim}{h}} \left(X_{hg} \frac{{}_aY_{hg}}{Y_{hg}} - {}_aX_{hg} \right) \tag{3.3}$$

and

$$B_{a\,cs} = \sum_g \left(\sum_{h \in \underset{\sim}{h}} X_{hg} \frac{\sum_{h \in \underset{\sim}{h}} {}_aY_{hg}}{\sum_{h \in \underset{\sim}{h}} Y_{hg}} - \sum_{h \in \underset{\sim}{h}} {}_aX_{hg} \right) \tag{3.4}$$

where ${}_aY_{hg}$ and Y_{hg} are defined as in section 2, and where X_{hg} and ${}_aX_{hg}$ are similarly defined for the x-variable (based on the census.

Relative absolute biases at the province level were obtained by summing the absolute biases over individual FED's and dividing by the provincial total for the x-variable.

4. ANALYSIS OF RESULTS

4.1 Efficiency considerations: Auxiliary Information up-to-date

In this part of the empirical (Monte Carlo) study, data used for simulation of the design and the auxiliary variables used in estimation refer to the same period as those of the study variable; i.e., to the 1971 census. Efficiencies of the four small area estimators are presented relative to the direct estimator in Table 1, for separate and combined levels of construction, and for each of the following auxiliary variables - dwellings, total population (15+), and population by age/sex groups. Census Divisions in the province of Nova Scotia whose populations range from 3,885 to 39,260 were used as the unplanned domains (type c) for the purpose of the study. The following observations can be made:

(i) Separate vs Combined Estimator: The level of construction of estimator does not have much impact on the efficiencies of synthetic estimators for both the characteristics employed and unemployed. For the post-stratified domain estimator for employed, however, the combined form is approximately twice as efficient as the separate. This is likely due to the effect of the clustering in the sample design being more accentuated with the separate estimator.

Since the post-stratified domain estimator was less efficient in its separate form, a similar result was anticipated for the composite estimator and hence, only the combined composite estimator was considered. On the other hand, the separate form of the sample dependent estimator was found to rely slighlty more on the synthetic component, leaving the efficiencies unaffected by the level of construction.

(ii) Effect of Auxiliary Information: The performance of population by age/sex as an auxiliary variable is uniformly superior, although only marginally so, to the total (15+) population for all four estimators using auxiliary information. Further, both these variables out-perform the dwelling count as an auxiliary variable.

In actual survey situations, the choice of population by age/sex as the auxiliary variable may be desirable also from the point of viex of correcting estimates for biases due to non-response and undercoverage as both factors may be dependent on age and sex.

(iii) Compararison among the estimators: For unemployed, performance of composite estimator with optimum α_2^* chosen for the characteristic unemployed is marginally superior to the other estimators irrespective of the level of construction, and the choice of auxiliary variable does not seem to have appreciable impact on any of the estimators. For employed, the situation is not that clear, however the sample dependent estimator shows an edge over other estimators and particularly so with population by age/sex as the auxiliary variable.

4.2 Efficiency Considerations: Auxiliary information out-of-date

In this part of the study whereas the design and auxiliary information were based on 1971 census results, the study variable was based on the 1976 census. As can be seen from table 2, although for unemployed the use of small area estimation techniques showed larger gains relative to the direct estimator (than in the up-to-date case), considerably smaller gains were observed for employed, which would likely be due to the reduced correlation between the study variable and the auxiliary information as both design and auxiliary information become out-of-date. Also in this case, the efficiency of the synthetic estimator is higher for both of the characteristics measured.

4.3 Consideration of Bias

Given that the post-stratified domain estimator will generally have negligible bias, the bias of both the composite and sample dependent estimators would

generally be smaller than that of the synthetic estimator, i.e. stemming only from the degree of reliance on the synthetic component. Hence the bias of synthetic estimator was investigated in detail. Using the total population (15+) as the auxiliary variable, the relative bias for the characteristics employed and unemployed were computed and are given in Table 3 for the ten provinces. Theses biases refer to the case where the unplanned domains are Federal Electoral Districts and the study variables are based on 1976 census data, while the survey design and adjustment factors (synthetic weights) are based on the 1971 census. Biases were also computed using age/sex sub-groups as the auxiliary variable and were found to follow similar trends while being marginally smaller. It is observed from this table, with the exception of the two smaller provinces, namely P.E.I. (for unemployed) and N.B. (for employed), that the relative bias of separate synthetic estimator is smaller than that of the combined synthetic estimator for both the characteristics under study. This confirms the intuitive feeling that the higher the level at which synthetic estimator is constructed, the higher would be the resultant bias in general, due to weakening of the assumption of homogeneity.

Biases were also computed for the case when both the study variable and the auxiliary information referred to the 1971 census. Biases for this case while slightly lower, followed similar trends to those in Table 3.

While the bias of the synthetic estimator was fairly small on average, it can be observed from Table 4 that it exceeded 10% in 13 and 19 (out of 279) FED's when the auxiliary information was up-to-date and out-of-date respectively. Further, in about half the instances for which the bias exceeded 10% for the up-to-date case the bias also exceeded 10% for the later time period when the auxiliary information was out-of-date. This suggests that for domains with a known high bias at the time to which the auxiliary information refers, less use should be made of the synthetic estimator. For instance, with the sample dependent estimator the value of K_0 could be set lower in such cases. However there is still the danger of bias in the synthetic estimator from category (ii) type cases in Table 4 which cannot be identified when deriving current estimates during the intercensal period.

4.4 Efficiency vs Bias in Overall Choice of Estimator

The synthetic estimator is generally highly biased and at the same time highly efficient. Therefore, in the search for a reasonable estimator for small areas, the question is to what extent one can reduce the effect of the synthetic estimator's bias, without sacrificing too much on its efficiency, in order to obtain a 'reasonable level of confidence' in the final estimate. At the same time it is also important to determine the reliance on the synthetic estimator without introducing too many computational complexities. Looking from this perspective in the context of the Labour Force Survey, one should strive for small area estimators whose performance for unplanned domains is comparable to that of simple survey estimates for planned domains, and amongst estimators meeting this criterion, more emphasis should be on reducing bias than on improving efficiency, especially if the differences in efficiencies are minor.

Average variances of the unbiased design estimator for the planned domains (say $\hat{X}$), comparable in size to the unplanned domains were obtained analogously to the average mse defined in (3.1). The efficiencies of the synthetic, composite and sample dependent estimators relative to the usual survey estimate for the planned domain i.e. X were also obtained. These efficiencies ranged from 1.08 to 1.17 for unemployed, and 1.22 to 1.47 for employed, hence all three estimators meet the above mentioned criterion. Since the sample dependent estimator makes use of the synthetic estimators whenever there is not 'sufficient' sample in the domain, its bias would depend upon the weight attached to the synthetic estimator component and this can be controlled by a proper choice of K_0. Table 5 presents the $(1-\delta)$ values, averaged over 100 replicates with $K_0 = 0.5$ and $K_0 = 1.0$ for the separate sample dependent estimator using total population 15+ as the auxiliary variable for each of the Census Divisions (unplanned domains) in this study. These average $(1-\delta)$ values indicate the degree of reliance of the sample dependent estimator on the synthetic component. As expected, domains consisting primarily of partial strata tend to place increased reliance on the synthetic component. Nevertheless, that reliance remains quite small. For example, with $K_0 = 1$ the highest value it assumes is .28 for Census Division 218.

Also as expected, the average $(1-\delta)$ values for $K_o = 0.5$ are lower than those for $K_o = 1.0$, implying the lower the value of K_o chosen, the lower would be the value of $(1-\delta)$ and consequently less reliance (weight) on the synthetic component of the sample dependent estimator. However as illustrated in Table 1, a trade-off between bias and efficiency is involved since lower choices of K_o also result in reduced efficiency. The above values of K_o provide a reasonable degree of confidence for the type of domains discussed here. In general, however, other values of K_o may be chosen depending upon e.g. the size of the domain, sample size, strata sizes and their geographical configurations with respect to the domain.

4.5 Concluding Remarks

1. The use of population by age/sex fares uniformly better than the other auxiliary variables, although gains over total population (15+) are mariginal.

2. The post-stratified domain estimator although more efficient as compared to the simple domain estimator, performs poorly as compared to the other three small area estimators investigated.

3. From the point of view of bias, the separate estimator has smaller relative bias as compared to the combined synthetic estimator. Further while average biases tend to be fairly small and tend to increase only slightly when the auxiliary information became out-of-date, biases for individual domains can be very high and change dramatically, frustrating efforts to identify 'outliers' where reduced reliance on syntheic estimators should be made.

4. The combined composite estimator constructed as a linear combination of post-stratified and synthetic estimators is more efficient than either of its component estimators although only marginally so, as compared to the synthetic component, for optimum value of α. Its bias would depend upon the weight attached to the synthetic component since the bias of the post-stratified estimator would generally be

negligible. Further, as the computation of the optimum α is quite involved, in practice only an estimated value of α may be used, resulting in a decrease in efficiency of this estimator.

5. The synthetic, composite and sample dependent estimator with $K_0 = 1$ are all more or less equally efficient, and out-perform the un biased design based estimator for planned domains.

6. Since the bias of the separate synthetic estimator is smaller than that of the combined synthetic estimator, the separate sample dependent estimator would result in smaller relative bias as compared to the combined sample dependent estimator. The bias of the separate to the combined sample dependent estimator. The bias of the separate post-stratified domain component can be controlled by collapsing those strata for which the intersection with the domain is very small. Thus considering all the three aspects, bias, mean square error and the computational complexities, the sample dependent estimator constructed at the stratum level using population by age and sex would seem to be a better choice.

5. FUTURE DIRECTION OF INVESTIGATION:

The study reported in this paper has focussed on evaluation of certain small area estimation methods using only census and survey data, in the context of the LFS, primarily for unplanned domains (type c). The estimators examined made use of synthetic and post-stratified domain estimators in different ways in an attempt to strike a balance between bias and mean square error. Below we point to directions which future investigations might take in efforts to develop statistically sound techniques for small area data in the Canadian context.

In the context of the Labour Force Survey, since the small area estimation methods for the unplanned domains have out-performed the unbiased design based estimates for comparable planned domains, it would be desireaable to extend this investigation to certain small planned domains (type a) as well. In par-

ticular the sample dependent estimator considered here and other similar estimators discussed in the literature will be further investigated for the Labour Force characteristics. In addition these investigations should also be extended to other smaller surveys conducted by Statistics Canada for which small area data are in demand. Further work on development of methods of variance estimation to be used in practice for these estimators is also needed.

Other estimators which seem to be promising are the Structure Preserving Estimators (SPREE) suggested by Purcell and Kish (1980). In this approach the estimation process, specified by the association structure (i.e. the relationship between y and x variables at some previous time at domain level) and the allocation structure (i.e. the current relationship at the larger area level), preserves the earlier relationship present in the association structure without interfering with current information in the allocation structure. In the Canadian context, for characteristics for which large scale surveys (such as the Labour Force Survey) are undertaken regularly, it would seem the short term demand for data for domains of the size of FED's or Census Divisions may be met through the use of refined estimation techniques (and pooling of estimates over a period of time) utilizing census and servey data alone. However, for meeting such demands in the longer term and for other types of data based on smaller surveys and other types and sizes of domains, all three sources of data namely census, surveys and administrative files would have to be fully explored. Multi-variate linear regression estimators of the type considered by Ericksen (1974) and Gonzalez and Hoza (1978) using data from all three sources should be studied in detail for their bias, mean square error and the computational complexities. Each of the three sources, with limitations of their own, when put together offer considerable potential for improvements in the sense that the weaknesses of one source can be the strengths of another. Hence there is reason for optimism that statistically sound techniques exploiting the strengths of data from different sources in an integrated fashion hold the future key to good quality small area data for a large variety of subject matters.

ACKNOWLEDGEMENT

Discussions with Mr. R. Platek have been beneficial in the finalization of this paper.

Table 1. Efficiencies of Small Area Estimators Relative to Direct Estimator - Nova Scotia Census Divisions (Auxiliary data up-to-date).

Characteristic	Auxiliary Variable	Level of Construction	ESTIMATOR: Post-Stratified Domain	Synthetic	Composite (α^*_2=0.223)	Sample Dependent K_0=0.5	Sample Dependent K_0=1.0
Employed	Dwelling	combined	4.58	10.17	10.92	9.17	10.42
	Population	"	4.92	10.75	10.58	10.50	11.67
	Population by age/sex	"	5.08	10.83	11.25	11.17	12.25
	Dwelling	separate	2.75	10.50	-	9.58	10.50
	Population	"	2.83	10.92	-	10.58	11.42
	Population by age/sex	"	2.83	11.00	-	11.00	11.75
Unemployed	Dwelling	combined	1.33	1.70	1.75	1.40	1.55
	Population	"	1.36	1.70	1.75	1.43	1.58
	Population by age/sex	"	1.36	1.70	1.75	1.43	1.58
	Dwelling	separate	1.30	1.69	-	1.48	1.58
	Population	"	1.33	1.69	-	1.51	1.61
	Population by age/sex	"	1.33	1.69	-	1.51	1.61

Table 2. Efficiencies of Small Area Estimators Relative to Direct Estimator - Nova Scotia Census Divisions (Auxiliary data out-of-date)

Characteristic	Auxiliary Variable	Level of Construction	ESTIMATOR: Post-Stratified Domain	ESTIMATOR: Synthetic	ESTIMATOR: Sample Dependent (K_0=1.0)
Employed	population	combined	3.47	4.73	4.07
	population by age/sex	"	3.60	4.73	4.23
Unemployed	population	combined	1.44	2.19	1.68
	population by age/sex	"	1.46	2.21	1.69

Table 3. % Average Relative Absolute Bias of Synthetic Estimators for FED's (using out-of-date, 15+ population as auxiliary information

Characteristic	Employed		Unemployed	
Level of Construction	Separate	Combined	Separate	Combined
Province				
Newfoundland	1.20	2.19	1.75	3.17
Prince Edward Island	3.54	5.62	3.71	2.80
Nova Scotia	0.87	1.64	1.25	1.87
New Brunswick	2.95	2.54	6.35	7.04
Quebec	2.53	3.87	3.87	4.51
Ontario	2.17	3.52	3.01	4.25
Manitoba	1.42	2.28	2.22	3.44
Saskatchewan	2.41	2.65	4.35	4.85
Alberta	5.24	7.08	5.12	10.33
British Columbia	1.73	2.71	2.72	4.20

Table 4. FED's with Biases of Separate Synthetic Estimator for Unemployed Exceeding 10%

Category (i)			Category (ii)			Category (iii)		
	% rel bias			% rel bias			% rel Bias	
FED	Up-to-date	Out-of-date	FED	Up-to-date	Out-of-date	FED	Up-to-date	Out-of-date
102	12.25	-3.90	414	1.57	17.78	301	12.71	25.85
104	-12.52	6.23	426	-7.38	-11.78	304	-11.29	-17.57
411	-13.10	-0.37	450	1.93	-11.35	412	-10.07	-15.80
436	10.48	-0.48	455	2.67	15.46	438	12.15	14.22
474	18.35	-6.04	460	-7.74	20.80	501	10.52	16.83
806	15.15	-0.20	504	7.19	41.05	579	15.50	10.59
			527	9.59	11.76	818	10.83	39.41
			605	3.63	14.29			
			701	4.90	17.74			
			804	0.78	15.85			
			813	-0.48	-15.46			

Category (i): bias exceeds 10% only when auxiliary information is up-to-date.

(ii): bias exceeds 10% only when auxiliary information is out-of-date.

(iii): bias exceeds 10% in both the cases.

Table 5. Average Reliance of Separate Sample Dependent Estimator on Synthetic Component: Nova Scotia Census Divisions.

Census Division	Reliance $(1-\delta)$ on Synthetic Component		Proportion of Census Division Population	
	$K_0=0.5$	$K_0=1.0$	Partial Strata	Complete Strata
201	.04	.15	1.00	-
202	.15	.20	.70	.30
203	.01	.12	1.00	-
204	.12	.22	1.00	-
205	.04	.14	1.00	-
206	.03	.08	.37	.63
207	.05	.07	.26	.74
210	.06	.09	.35	.65
211	.04	.05	.13	.87
212	.06	.10	.52	.48
213	.03	.16	1.00	-
214	.04	.16	1.00	-
215	.11	.21	1.00	-
216	.05	.15	1.00	-
217	.01	.01	.03	.97
218	.18	.28	1.00	-

REFERENCES

[1] Drew, J.D. and Choudhry, G.H. (1979), "Small Area Estimation", Technical Report, Census and Household Surveys Methods Division, Statistics Canada.

[2] Ghangurde, P.D. and Singh, M.P. (1976), "Synthetic estimation in the LFS", Technical Report, Household Surveys Development Division, Statistics Canada.

[3] Ghangurde, P.D. and Singh, M.P. (1977), "Synthetic Estimates in Periodic Household Surveys", Survey Methodology, Vol. 3, No. 1, 152-181, Statistics Canada.

[4] Ghangurde, P.D. and Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 53-61.

[5] Gonzalez, M.E. (1973), "Use and Evaluation of Synthetic Estimates", Proceedings of the American Statistical Association, Social Statistics Section, 33-36.

[6] Gonzalez, M.E. and Waksberg, J. (1973), "Estimation of the Error of Synthetic Estimates", paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

[7] Gonzalez, M.E. (1975), "Small Area Estimation of Unemployment", Proceedings of the American Statistical Association, Social Statistics Section, 437-460.

[8] Gonzalez, M.E. and Hoza, C. (1978), "Small Area Estimation with Application to Unemployment and Housing Estimates", Journal of the American Statistical Association 73, 7-15.

[9] Holt, T., Smith, T.M.F. and Tomberlin, T.J. (1979), "A Model Based Approach to Estimation for Small Sub-groups of a Population", Journal of

[10] National Center for Health Statistics (1968), "Synthetic State Estimates of Disability", P.H.S. Publication No. 1759, U.S. Government Printing Office, Washington, D.C.

[11] Platek, R. and Singh, M.P. (1976), "Methodology of the Canadian Labour Force Survey," Catalogue No. 71-526, Statistics Canada.

[12] Purcell, N.J. and Linacre, S. (1976), "Techniques for the Estimation of Small Area Characteristics", paper presented at the 3rd Australian Statistical Conference, Melbourne, Australia.

[13] Purcell, N.J. and Kish, L. (1979), "Estimation for Small Domains", Biometrics 35, 365-384.

[14] Royall, R.M. (1973), "Discussion of two papers on Recent Developments in Estimation of Local Areas", Proceedings of the American Statistical Association, Social Statistics Section, 43-44.

[15] Royall, R.M. (1978), "Prediction models in Small Area Estimation", NIDA Workshop on Synthetic Estimates, Princeton, N.J.

[16] Sarndal, C.E. (1981), "When Robust Estimation is not an obvious answer: The case of the Synthetic Estimator versus Alternatives for Small Areas", Proceedings of the American Statistical Association, Survey Research Section.

[17] Schaible, W.L., Brock, D.B. and Schnack, G.A. (1977), "An Empirical Comparison of the Simple Inflation, Synthetic and Composite Estimators for Small Area Statistics", Proceedings of the American Statistical Association, Social Statistics Section, 1017-1021.

[18] Schaible, W.L. (1978), "Choosing Weights for Composite Estimators for Small Area Statistics", Proceedings of the American Statistical Association, Survey Research Section, 741-746.

[19] Schaible, W.L. (1979), "A Composite Estimator for Small Area Statistics", in Synthetic Estimates for Small Areas (J. Steinberg, Ed.) National Institute on Drug Abuse Research Monograph No. 24, U.S. Government Printing Office, Washington, D.C., 36-53.

[20] Singh, M.P. and Tessier, R. (1975), "Some Estimators for Domain Totals", Journal of The American Statistical Association 71, 322-325.

[21] Sonquist, J.N. and Morgan J.A. (1964), "The Detection of Interaction Effects", Monograph no. 35, Survey Research Center, Institute for Social Research, University of Michigan.

CHARACTERISTICS OF RESPONDENT AND NON-RESPONDENT HOUSEHOLDS IN THE CANADIAN LABOUR FORCE SURVEY

Elizabeth Clayton Paul and Murray Lawes [1]

This article presents findings from a study to characterize responding and non-responding households in the LFS. This study was motivated by two projects associated with the LFS Redesign, namely, the family estimation project and evaluation of non-response compensation procedures. However, the results of the study are of general interest in the assessment of the quality of data emanating from the LFS.

1. INTRODUCTION

Non-response is the lack of complete information for all selected units in a sample or census. The occurrence of non-response poses special problems for the producers and users of survey data. Non-response affects the quality of survey data in two basic ways. First, it reduces the effective sample size, resulting in loss of precision of the survey estimates. Second, to the extent that differences in the characteristics of respondent and non-respondent units are not properly accounted for in the estimation strategies, it may introduce a bias into the survey estimates. This paper focuses on the latter aspect of quality, specifically the characterization of respondent and non-respondent units in the Canadian Labour Force Survey (LFS). This information will provide some insight into the potential effect of non-response on the survey estimates and will suggest some variables which should be considered when compensating for non-response. Units were characterized by the variables size of household, economic family type, length of time in the survey, location, age of household members and labour force status of household members. This study is based on data derived from the LFS longitudinal data files. A statement of

[1] Elizabeth Clayton Paul, Economic Characteristics Staff, Statistics Canada and Murray Lawes, Census and Households Survey Methods Division, Statistics Canada.

major findings from this analysis is found in Section 2 followed by a brief description of the LFS, of the longitudinal files and the methodology used to characterize non-respondent households in Section 3. Section 4 then presents the derived data and resulting analysis. The final section briefly discusses the impact of the findings of this study on the quality of LFS data at the individual, family and household levels and suggests potential methods of dealing with non-response to alleviate or minimize deficiencies in the survey data arising due to non-response.

2. STATEMENT OF MAJOR FINDINGS

Within the LFS, non-response compensation procedures are based on the assumption that the characteristics of non-respondent households are similar to the characteristics of respondent households. Should this assumption prove incorrect, the non-response adjustment procedure will contribute to a bias in the survey estimates. It is impossible to determine the exact extent of this non-response bias. However, by examining longitudinal data on the survey life of a household, a profile of respondent and non-respondent households may be determined and the extent of differences evaluated.

Of the many variables examined in the characterization of respondent and non-respondent households, the variables month in sample, household size and labour force status of household members exhibited a definite trend in relation to response status. With respect to month in sample, the levels of non-response decreased as month in sample increased. Between months one and two the percentage of non-respondent households decreased sharply, and then gradually continued to decrease until month six, implying survey tenure is a critical factor in the determination of survey response. Thus any estimates by rotation number based on a non-response adjustment across all rotation groups may impart a slight bias to estimates on a rotation number basis.

Regarding household size, non-response decreased as household size increased. On a distributional basis there were almost twice as many households of size one for non-respondent households as for respondent households; and conversely, for households of size 5 and over, there were over twice as many

households for respondent households. The implication is that a non-response adjustment which does not take household size into consideration will, on average, represent non-respondent households by households which contain more household members than the non-respondent households.

The response patterns exhibited by household size and month in sample remained unchanged when the two characteristics were jointly examined. Since the analysis of these two variables, household size and month in sample, has shown a strong functional relationship with non-response, a non-response adjustment incorporating household size and month in sample should do much to alleviate discrepancies by rotation number in sample survey estimates of household and economic family units, and of characteristics dependent on these variables.

In addition to household size and month in sample, a relationship between non-response and labour force status was also exhibited, with particular reference to unemployment. For non-respondent households, the percentage of individuals classified as unemployed increased as month in sample increased, while the percentage for respondent households remained relatively stable. When the added dimension of household size was examined, a definite relationship was exhibited for households of size one with a slightly more variable pattern being exhibited for households of size two or more. For households of size one, the percentages of individuals classified as employed and unemployed were substantially greater for non-respondent households than for respondent. Also, the percentage of employed individuals decreased as month in sample increased; however, the percentage of unemployed increased. For households of size two or greater, the differences in the labour force distributions for respondent and non-respondent households were less pronounced than those for size one households, but the percentage of unemployed individuals in non-respondent households of size two or more did generally increase as month in sample increased.

Although there may be advantages in utilizing some variables relating to labour force activities in addition to household size and month in sample in the non-response adjustment process, and thus improving the labour force estimates; the desire for a general weight adjustment, the small sample size

at this level of aggregation, and the relatively low level of non-response currently experienced in the LFS may preclude the implementation of a non-response adjustment based on labour force status related variables. However, a non-response adjustment on the basis of household size and month in survey should have some benefits for the labour force estimates. Consequently, it may be feasible to consider adjustments for two groups of households, namely size one and size two or more, and for two survey tenures, namely one month and two months or more, in evaluating any improvements to the current LFS non-response adjustment process.

3. DATA SOURCE

3.1 The Labour Force Survey

The LFS is a multi-stage stratified random sample with stratification occurring within the economic region level for each province. The final unit of sample selection is the dwelling. Each selected dwelling remains in the survey sample for six months. At the end of that time, these dwellings are replaced by another group of dwellings in such a manner that every month one-sixth of the sample is replaced or rotated. This implies that in any given month, there are six panels of dwellings in the LFS with each panel at various stages of aging. That is, one panel is in the survey for the first occasion (i.e., the birth rotation group), one panel for the second occasion,..., and one panel for the sixth occasion.

During one week each month, Survey Week[1], LFS interviewers contact selected dwellings to obtain information on the composition, demographic variables and labour market activities of household members who are part of the survey universe[2]. For various reasons, interviewers are unable to obtain information from all selected dwellings. These dwellings where no interview is conducted are classified as vacant dwellings or non-respondent households[3], depending on their occupancy status. For vacant dwellings, no response is obtainable or expected; whereas, for non-respondent households, survey information is missing. An adjustment[4] for non-response to compensate for this missing information is made at the data processing stage based on the assumption that

households which have been interviewed, i.e., respondent households, typify households which should have interviewed, i.e., non-respondent households. Should this assumption be false, then a bias is introduced into the survey estimates by this adjustment for non-response. This bias will increase as the rate of non-response increases. For this reason, it is important that the characteristics of non-respondent and respondent households be similar, and for this reason much effort is expended (successfully) in minimizing non-response.

3.2 Longitudinal Data File

Estimates based on monthly cross-sectional LFS data provide a static snapshot of the population and labour market for each month; however, by linking respondent information over the survey lifetime, a dynamic view of labour market activities is observed. In any given month, dwellings in one of the six rotation panels complete their six-month tenure in the survey. For dwellings in this panel, it is possible to trace the household composition and response pattern over the previous five months. This tracing is done by means of the Longitudinal Data File. The Longitudinal Data File is formed by concatenating the information on a given household over its six months of survey life.

In the LFS, dwellings and individuals are assigned unique identification codes. This affords a method of linking individual, household, and dwelling information over the six months a dwelling is in the survey, thus creating the Longitudinal Data File.

Initially, longitudinal records containing the six monthly response status codes are created for each dwelling. If a dwelling is respondent for one or more months, then individual records containing information on the household members who were living in the household at the time it was respondent are also included on the longitudinal file. However, if no response is indicated over the six months, only basic dwelling information is available for the dwelling. Thus, every individual who was a household member at some time over the six-month survey period is associated with a Longitudinal Data File

record. From this record, labour market activity and demographic information can be obtained for the months the individual was a responding household member. Based on this formulation of longitudinal data, examination of responding and non-responding households can occur and the characteristics of each response type evaluated.

3.3 Methodology for Deriving Estimates

In examining the characteristics of responding and non-responding households, the type of household response for each month was required. On a monthly basis, there are three types of dwelling responses: respondent, non-respondent, and vacant. Responding households are those where the LFS questionnaire is completed for all or some eligible household members. Non-respondent households are occupied by individuals who should be included in the survey but, for some reason, choose not to participate or are unable to participate due to existing circumstances. Vacant dwellings, on the other hand, are not occupied, or are occupied by individuals not included in the survey universe.

Thus, in determining the characteristics of responding and non-responding households, dwellings labelled as vacant were ignored.

To obtain the characteristics of responding households, the characteristics of individual household members who responded in the survey were examined; however, to obtain the characteristics of non-responding households, an imputation strategy was implemented. The characteristics of a non-responding household should be identical to or closely approximated by characteristics of individuals in that household in a month of response.

For those households who did respond at least once during the six months the household was in the survey, the months of response were the information donors for any months of non-response during the six months. In this manner, the characteristics of non-responding households were estimated. To impute for non-response by this method, it was imperative that a given household be respondent for at least one month; however, the household could have been

respondent for more than one month. If this latter situation occurred, the month of response closest to the month of non-response provided the donor information. If two months of response were equally close to a month of non-response, the month prior to the month of non-response was chosen as the donor month. The following algorithm summarizes this technique.

Month of Non-Response	Ordering of months to check for donor information
1	2, 3, 4, 5, 6
2	1, 3, 4, 5, 6
3	2, 4, 1, 5, 6
4	3, 5, 2, 6, 1
5	4, 6, 3, 2, 1
6	5, 4, 3, 2, 1

If there was no month of response available, then no imputation wasperformed and this household was excluded from this study.

3.4 Cautionary Note

If non-response rates based on this study are compared to non-response rates by rotation groups from the monthly LFS, they will differ in magnitude. The main source of difference is the exclusion of certain non-respondent households from this study of longitudinal data. As previously indicated, the ability to characterize a household in a month of non-response depended on the availability of respondent data in an alternative month for that household. That is, there had to be at least one month of response for a non-respondent household to be characterized. This implies that a household which was non-respondent, or a combination of non-respondent and vacant, for each of the six months it was part of the survey sample was excluded from this study. Thus, some non-respondent households which contributed to the monthly LFS measurement of non-response did not contribute to this longitudinal study of non-

response. Approximately 1.4% of the total sampled households were excluded on this basis.

Exclusion of some non-respondent households is the main reason for differences in data from this study and any other study on non-response which is based on the monthly LFS data. In addition to this source of discrepancy, the weighting technique applied may cause estimates to vary. For this report records were weighted by a product of the inverse sampling ratio, the sub-sampled cluster weight, and the stabilization weight[5]. In examining and interpreting the results in Section 4, or comparing these results to any other study on non-response, it is necessary to remember that the data source was the Longitudinal Data File, only records with at least one month of response contribute to the estimates, and the weighting structure was based on sample design weights only.

4. ANALYSIS

The methodology in the previous section documented the procedures used to derive estimates of characteristic totals from the longitudinal file. In this section a number of variables (separately and jointly) are examined with respect to their characterizations between respondents and non-respondents. A particular variable or cross-classification of variables is dealt with in each of the following subsections. The motivation for examining the variables, tables containing relevant tabulations and a summary of the essential results are presented for the various subsections.

4.1 Month in Sample

As noted in the introduction the LFS is based on a rotating panel design with each panel of dwellings remaining in the sample for a period of six months. At the sample design stage, considerable effort is taken to ensure that the sample associated with each rotation number (i.e. dwellings by panel) is a representative one-sixth subsample of the full LFS sample. In the past a number of references have been made to the phenomenon of rotation group bias, i.e. that the expected value of estimates based on a single panel differs

depending on number of months in the sample. For this reason the composition of the sample by month in sample and by response status were examined. Weighted estimates of the number of households at the Canada level by month in sample and by response status were obtained based on averages over 1980 and 1981 and are presented in Table 1. Due to design efforts to ensure representativeness of the sample by rotation number, it was expected that the total weighted counts would be equally distributed by month in sample. Examination of the data revealed that very close to one-sixth (or 16.67%) of the total households fall into each month in sample class. In all cases the differences in percentage distribution for a cell were within one-half of 1%.

When distributions of households by month in sample were examined by response status, deviations from a uniform distribution were observed, particularly for non-respondent units. The non-response rates by month in sample exemplified this fact. As illustrated in Table 1, the rate of non-response decreased as the number of months in the survey increased. The largest decrease occurred between the first and second months in the sample when the rate in the second month was approximately one-half of the rate in the first month. Further reductions in the non-response rates were observed as the number of months in the sample increased. Decreases in the rates between the second and sixth months were 21.1% and 34.2% for 1980 and 1981 respectively.

The percentage distribution of non-respondent households exhibited a similar decreasing trend as number of months in the sample increased. On a distributional basis, there are substantially more non-respondent households in the first month in sample than there were respondent households; however, this number decreased with increasing tenure in the survey. Thus any estimates by rotation number based on a non-response adjustment across all rotation numbers may impart a slight bias to estimates on a rotation number basis.

4.2 Household Size

In the LFS, non-response generally occurs at the household level, i.e. the rate of partial non-response within households is very low. The household is the unit at which non-response occurs. Thus the characterization of house-

holds is necessary for the determination of the effects of non-response on estimates from the survey - be they at the level of household, family, or individual units. Perhaps the most basic household attribute, in relation to deriving demographic/socio-economic estimates from the survey, is household size. From a data collection point of view it is reasonable to assume that difficulties of contacting households decrease with increasing household size.

To evaluate the potential effect of household size on the non-response rate, Table 2 presents the percentage distribution of households by size and response status based on averages over the calendar years 1980 and 1981. For both years the non-response rate decreased dramatically as household size increased. Non-response rates by household size ranged from a high of 7.48% for households of size 1 to a low of 1.89% for households of size 5 or more in 1980 and correspondingly from 6.58% to 1.69% in 1981 for households of sizes 1and 5 or more, respectively. An examination of the distribution of responding and non-responding households by size of household revealed a substantial difference in the distribution of households by size depending on the response status. On a distributional basis there were almost twice as many households of size one for non-respondent households as for respondent households. For respondent households there were slightly more than 50% which were of size 3 or more, whereas for non-respondent households only about 30% were of size 3 or more. The distributional differences in household size between respondent and non-respondent households was also reflected in the average household size for each response type. For 1980 the average household size for respondent and non-respondent households was 2.93 and 2.26, respectively; while for 1981 the corresponding sizes were 2.88 and 2.19. The implication is that with the adjustment for non-response at the LFS data processing stage, non-respondent households are represented by households which, on average, contain more household members than the non-respondent household. This leads one to question the assumption that respondent households typify non-respondent households, at least with respect to household size.

4.3 Household Size by Month in Sample

In the previous two subsections substantial variations in the response rates were noted depending on the number of months in sample and also depending on the size of household. The next table was obtained to determine whether the noted variations in non-response rates were also observed when either household size or month in sample was held constant. Based on annual averages for 1980 and 1981, Table 3 presents percentage distributions of respondent and non-respondent households by household size and month in sample as well as the corresponding non-response rates for 1980 and 1981, respectively.

These tables show that the decreasing trends in non-response rates observed in Tables 1 and 2 for the full populations also hold true when the rates are examined holding one of the variables constant and letting the other vary. For example, in Table 1 non-response rates for all household sizes combined were shown to decrease as month in sample increased. Table 3 generally shows the same phenomenon when one examines the pattern of response rates by month in sample for each of the household size groupings separately. As when months in the survey alone were examined, the non-response rate decreased sharply from month one to month two. Similarly, the non-response rate decreased from month one to month two by approximately one half for each given household size. For households of size one and two the non-response rate continued to decrease in subsequent months in the survey; however, for households of size 3 and greater the non-response rate tended to stabilize during the second month in the survey.

Holding the number of months in the survey constant and examining the non-response rate as the household size varied, revealed a pattern similar to that exhibited in Table 2, where household size alone was considered. The non-response rate decreased with increasing household size. Table 3 likewise shows that for a given number of months in the survey (from one to six), there is a decreasing trend in the non-response rate as household size increases.

Combining these two trends, there was an expectation that the highest non-response rate would be observed in households of size one during the first

month in the survey. Similarly, there was an expectation that the lowest non-response rate would be observed in households with five or more members during the final month in the survey (i.e., in month six). Based on annual averages for 1980 and 1981, this expectation was verified. In 1980 and 1981 the non-response rates of highest magnitude were 13.39% and 12.81% respectively. Each of these rates applied to households of size one during the initial survey month. The non-response rate of least magnitude in 1980 was 1.54%. This applied to households containing five or more members during the third month in the survey; however, a non-response rate of 1.59% also applied to households containing five or more members for month 6. In 1981, the non-response rate of least magnitude was 1.37%. This occurred in households having five or more members during month 3, while the non-response rate for month 6 was 1.39%. Thus, although the lowest non-response rate did not uniquely occur in households containing five or more members during the final survey month, the non-response rate for households in this cell was not significantly different.

The distributions of household size by survey duration by response status indicated the potential for non-response bias in survey estimates. A non-response adjustment which does not take into account household size, will implicitly compensate for non-respondent households on the basis of the distribution of respondents, i.e., underestimating households of size 1 and 2 and over-estimating households of size 3 or more. It can be seen on a distributional basis that there were substantially more households of sizes 1 and 2 among non-responding households than there were among responding households and, of course, conversely fewer households of larger sizes (3, 4 and 5+) among the non-responding households than among the responding households. This discrepancy in distributions became more exaggerated when months in sample, or rotation groups, were considered, particularly for months one and two. After month two, the non-response rate tended to stabilize for households of size greater than two, whereas for households of size 1 or 2, the non-response rate continued to vary over the survey lifetime. This suggested that household size and rotation number are important characteristics to consider when methods for non-response adjustment are being evaluated.

4.4 Family Composition of Household

In Section 4.2 there were substantial differences in the distribution of households by size between respondent and non-respondent households. To further evaluate household size discrepancies between respondent and non-respondent households, tabulations of households in terms of their composition of family types were obtained. The family type compositions were based on the number of economic families in the household, the size of the family units, the presence of children, and the marital status and age of the head of the family unit. The specific variables are indicated in Table 4a with corresponding percentage distributions and non-response rates by type by response status in Table 4b.

The higher non-response rates for households of size one were again evident from these tabulations. The rates were particularly high for households containing only an unattached individual aged less than 65 years of age. Households containing a married couple with other members present in the household (children or non-children) i.e., codes 6, 7 and 8 had low non-response rates relative to other types of households. In other words, there were proportionately more of these types of households among the responding than among the non-responding households. Households containing only unattached individuals (either one or more) and households containing a married couple only formed a higher percentage of non-responding households than of responding households. Thus in addition to household size, the composition of the household in terms of family types appeared to have some influence on the rate of non-response. Thus certain types of family units may not be properly compensated for in various weight adjustment strategies for non-response. This is particularly a crucial issue in the production of family estimates.

4.5 Age of Individuals

Although the unit of potential response is generally the household, Table 5 presents percentage distributions by age group and response status at the individual level. Also presented are the distributions of the non-respondents

as percentages of the total population, or these could be referred to as individual level non-response rates.

The rate of non-response for all individuals combined were 3.13% and 2.63% for 1980 and 1981 respectively. These rates corresponded to household level non-response rates of 4.02% and 3.43% respectively for 1980 and 1981. The lower rates at the individual level were indicative of the inverse relationship between the size of household and the level of non-response as pointed out in Section 4.2. Since larger households had lower non-response rates, a greater proportion of individuals fell into the responding category. The relationships on a distributional basis between individual respondents and non-respondents bore out the results of the previous section with respect to the generally lower household non-response rates in households which contained children. For the age groups 0-14 and 15-19, the non-response rates in 1980 were 2.50% and 2.42% respectively, while in 1981 they were 2.12% and 1.92%. The highest non-response rates were observed in the age groups 65+ and 20-24. This again reflected the inverse relationship between household size and the non-response rate. Households of size 1 and 2 had the highest non-response rates. Individuals within the age groups 65+ and 20-24 were more likely to live alone or as a couple; hence, the non-response rates for these individuals were expected to be high. The variation in non-response rates by individual age groups indicates a potential effect on the quality of survey based estimates. In particular, age groups with a lower non-response rate than the over-all individual non-response rate will be over-estimated by a weight adjustment factor which does not take into account age variables. The opposite occurs when the non-response rate for the age group is greater than the overall individual non-response rate. To some extent any distortions introduced at the provincial level are corrected by the application of the ratio adjustment procedure.

4.6 Age of Individuals by Size of Households

Continuing from the previous section the distributions of individuals by age groupings and response status were obtained within various household size breakdowns. These distributions as well as non-response rates, are presented

in Table 6a based on 1980 annual averages and Table 6b based on 1981 annual averages.

The distributions of individuals by age group were relatively similar by household size between respondents and non-respondents in households of sizes 2, 3, 4 and 5+; however, for households of size 1 there were substantial differences in the distributions. Within size 1 households the primary differences were for age group 25-44 in which there were substantially more individuals (on a distributional basis) in non-responding than responding households (39.6% compared with 28.8% for 1981 and 35.5% compared with 27.9% for 1980) and for age group 65+ in which there were substantially fewer individuals in non-responding households than in responding households (22.3% compared with 34.3% for 1981 and 22.4% compared with 34.3% for 1980). This latter observation is particularly important as about 28% of the population 65+ reside in households of size 1 whereas less than 5% of individuals in the age group 25-44 reside in households of size 1. Thus, it is differences in the distributions by age groups between respondents and non-respondents which merit special attention in any procedures to compensate for non-response in households of size 1.

The non-response rates in Tables 6a and 6b show that individual non-response rates within age groups exhibit the same pattern across household size measures as was observed in Section 4.2, namely that non-response rates decrease as household size increases. Within a particular size of household the relationships of non-response rates by age group were very different than non-response rates by age groups for all household sizes combined. Perhaps most notable was the fact that for each household size group separately (except size 4 in 1980), individuals 65+ exhibited the lowest level of non-response whereas the non-response rate for individuals 65+ in all households combined was the largest of any age group. This phenomenon resulted from the fact (mentioned earlier in this section) that the majority of individuals of age 65+ live in households of size 1 or 2, where the non-response rate was the greatest.

These tables indicate that non-response is very much dependent on household size and that age is not an important factor apart from the fact that there is a relationship between household size and the age of individuals residing in the household.

4.7 Age of Individuals by Month in Survey

The distribution of individuals by age group for varying numbers of months in the survey, separately for respondents and non-respondents, are presented in Tables 7a and 7b for 1980 and 1981 respectively, as well as the corresponding non-response rates.

From Tables 7a and 7b it can be noted that distributions by age group for respondents were virtually identical regardless of the number of months in sample. Although the distributions for non-respondents showed a higher degree of variability for differing months in sample, there remained a degree of stability in the distributions. The pattern between distributions for respondents and for non-respondents was similar for each month in sample breakdown as it was for totals across months in sample.

A study of individual non-response rates again indicated in general a decreasing trend as number of months in sample increased. This occurred for individual age groups as well as for the total population. As expected the pattern over time was not as pronounced for individuals as it was on a household basis. This can be attributed to changes in the response pattern for various sized households; that is, there is a tendency for larger sized households to become non-respondents in the later survey months while smaller sized households tend to become respondent (refer to Table 3).

4.8 Labour Force Status

In this subsection attention is turned from the basic demographic characteristics of households by response status to the characteristics of labour force activity. This evaluation was motivated by the desire to assess potential non-response bias in the survey estimates of these characteristics.

Section 4.2 presented substantial differences in the distributions of respondent and non-respondent households by household size, while Section 4.1 presented similar findings for month in sample. For this reason, the distributions of individuals by labour force status within each category defined by household size, month in sample, and response status were examined. They are presented in Table 8a.

Examination of these distributions by labour force status for all individuals regardless of size of household, showed that the distributions for respondent households differ in some important ways from the distributions of non-respondents and the pattern of differences was not consistent over time. The percentages of individuals unemployed showed perhaps the most interesting changes. For respondents, this percentage was relatively constant for each number of months in the sample; whereas, for non-respondent households, there was an increase in the percentage of individuals unemployed as the number of months in sample increased. The percentage of the population (aged 15 and over) unemployed for respondent households ranged from a low of 4.7% in months 3 to 6 to a high of 5.0% in month 1 for 1980, and a low of 4.6% in months 4 and 5 to a high of 4.9% in month 1 for 1981. For non-respondent households, the corresponding range of percentages was 4.5% in month 1 to 6.4% in month 6 for 1980, and a low of 4.0% in month 1 to a high of 6.2% in month 5 for 1981. A comparison of the percentage unemployed for each response status over time shows that there were fewer unemployed persons among non-respondent than respondent households for households in the sample for the first occasion and more unemployed persons among non-respondent than respondent households for households in the sample for four to six months. The relationship was variable for months two and three. A comparison of the percentage distribution patterns of labour force activities for respondent households over time indicated a relatively stationary distribution; however, the pattern for non-respondent households varied. For non-respondent households there were greater fluctuations in the percentage distributions for each labour force status across months. No distinct pattern of change was exhibited except with unemployment where representation increased with survey duration. This variation among non-respondents was at least partly attributable to small sample sizes of non-respondents relative to sample sizes for respondents.

Since unemployment is more sensitive to sample fluctuations than the other labour force statuses and exhibits a definite trend over time, compensating for non-response over rotation groups would distort this characteristic. Adjusting over rotation groups would result in an overestimation of unemployment in month 1, and an underestimation of unemployment in months 4 to 6. Since the divergence between responding and non-responding households in the percentage distribution of unemployment was more pronounced in the later survey months, the overall effect would be an underestimation of unemployment. Since the non-response adjustment occurs at the household level, not at the individual level, and the size of the household has proven to be an important response determinant (see Section 4.2), it is essential to consider household size as an additional component for the evaluation of non-response with respect to the labour force status.

When distributions by labour force status and month in sample were examined by household size breakdowns, the patterns or relationships noted above did not hold. For households of size 1, the proportions of individuals employed and unemployed were substantially higher for non-respondents than for respondents. For respondents the proportion of individuals employed and the proportion unemployed were relatively constant for varying number of months in the sample. For non-responding households, there was a general decrease in the proportion of individuals employed as the duration in sample increased; whereas, there was a substantial increase in the proportion of unemployed as the number of months in sample increased.

For households of other sizes (2,3,4 and 5+), the differences between labour force status distributions for respondent and non-respondent households were much smaller. Also patterns between distributions for respondents and non-respondents were not nearly as strong or consistent as for the case of household of size 1. On a distributional basis, there were generally fewer unemployed individuals in non-respondent households for the first survey occasion and more unemployed individuals in non-respondent households for the fourth and subsequent months in the sample, than for responding households. For households in the survey for two or three months the pattern was variable.

The percentage of individuals "not in the labour force" differed between responding and non-responding households by household size. In households of size 1 and 2 there were fewer individuals "not in the labour force" in non-responding households than in responding; whereas, no definite pattern existed for households of size 3 or more. As the employed constituted the majority of the group "in labour force", generally the relationship on a distributional basis between respondent and non-respondent households was the complement of that noted for the characteristic "not in the labour force".

Table 8b presents unemployment rates by household size and month in sample by response status for 1980 and 1981 respectively. These results are related to those in the previous tables and observations may be similar in that the relationship between unemployment rates for respondents and non-respondents are the result of the relationships between proportions employed and unemployed between respondent and non-respondent units.

For all individuals (i.e., regardless of household size) the rate of unemployment for non-respondents was less than the rate for respondents for the first month and greater than the rate for respondents in months 4 to 6. The relationship between the rates for months 2 and 3 varied by year. For non-responding households, there was a substantial increase in the unemployment rate as the number of months in the survey increased. This phenomenon was not observed for respondents where the first month in sample had the highest rate but the pattern for subsequent months was somewhat variable.

For households of sizes 2, 3, 4, and 5+ the same general relationship in unemployment rates between respondent and non-respondent households was observed as for the full set of individuals (i.e., regardless of household size). There was no definite pattern in unemployment rates over time for non-respondent households when various household sizes were considered. For households of size 1 the unemployment rate for non-respondents was generally higher than the rate for respondents.

4.9 Type of Area

Results presented in Section 4.3 showed that there were substantial differences in distributions of households by size and month in sample between responding and non-responding households. This section further examines these results within broad types of area determined generally on the basis of population concentration and density; namely, self-representing areas (SRU), non-self representing urban areas (NSRU urban), and non-self-representing rural areas (NSRU rural). Although a more precise definition of area types is available, for this study it is sufficient to note that SRU's consist of the larger cities in the country, NSRU urban areas consist of smaller cities and towns, and NSRU rural areas are composed of the more sparsely populated portions of the country, including small villages and farm land. Due to the very small sample sizes, special areas were not considered. In very general terms, the patterns observed in Section 4.3 for all area types combined, were similar to those observed for the three broad area types; however, there were different distributions by household size for respondents depending on type of area. In SRU areas, on a distributional basis, most households were smaller sized whereas there were fewer smaller sized households in NSRU rural areas. The opposite was observed for larger sized households. The relationship between respondent and non-respondent households, however, was relatively the same regardless of type of area. From Tables 9a and 9b it can be noticed that there were approximately twice as many households of size 1 in non-responding households as in responding households and approximately one-half as many larger sized households (5+) in non-responding as in responding households.

Non-response rates, although levels differ by type of area, showed the same pattern of decreases by number of months in sample as was observed for all units combined (i.e., as compared with results presented in Section 4.3). Again there were substantial decreases in levels of non-response between the first and second months with decreases of lesser magnitude occurring in subsequent months.

The rates of non-response for all households (i.e., regardless of household size) were the highest for SRU areas, followed by NSRU urban areas and were

the lowest for NSRU rural areas. These differences were a function of the distributions of households by size across area types. Within specific size of household groupings, the patterns between respondent and non-respondent households are generally the same as when examined for comparable size groupings for all area types combined. The type of area variable is an important factor in compensation procedures as it differentiates between areas with different levels of non-response. However, in addition to size of households and month in sample variables the type of area variable does not provide much additional information in the characterization of survey units by response status.

5. SUMMARY

The previous section presented characterizations of responding and non-responding households with respect to a wide range of variables. The households and/or individuals displayed somewhat different characterizations depending on their response status. On the assumption that responding and non-responding households exhibit similar characteristics, it would seem to be important to incorporate some of the variables examined in Section 4 into non-response compensation procedures for the survey.

The method of compensating for non-respondent households in the LFS is carried out within small geographic areas (balancing units) by an inflation of the design weight by the inverse of the household response rate. These adjustments are made on the basis of household counts independent of any characteristics of the household. Unless there is a high degree of correlation among households within balancing units, one would expect very little reduction in non-response bias by the present adjustment procedure.

An indication of the magnitude of non-response bias under the current procedure for compensation for non-response would be desirable. An explicit imputation of missing information due to non-response on the LFS file can be obtained using procedures similar to those used in this study. After adjustments for complete non-response (i.e., non-response for all six months) survey estimates based on these comprehensive imputation strategies can be obtained.

Comparison of these resulting estimates with official survey estimates would provide added support to assessments of response bias which have been alluded to in this report.

This report has provided justifications for considering various additional variables in the adjustment for non-response: month in sample, household size and labour force status. As there are substantial variations in the response rate by rotation number (month in sample) it is advisable to adjust for non-response within each rotation number separately. As the pattern of labour force characteristics for non-respondents exhibits a degree of variation over months in sample, an adjustment on the basis of rotation number should have some benefits for labour force estimates as well. As the greatest differences are between the first month and subsequent months in sample, an adjustment for these two classes may be sufficient.

Among the non-responding households there are substantially more households of size one (and to a lesser extent for size two) than in responding households. Thus, household size is an important variable to be incorporated in any adjustment procedures for non-response. The analysis has shown that discrepancies are the greatest for households of size one. It may thus be feasible to consider adjustments for two groups of households only, namely households of size one and households of size two or more. Incorporation of household size into compensation procedures for household non-response necessitates having some information available about the size of non-responding households. This may be explicit, as for example the household size on a previous survey occasion, or implicit, as for example a distribution of non-responding households by size from previous surveys, or a distribution by household size from an independent source such as the Census. In either situation, adjustments incorporating considerations of household size in conjunction with adjustments by rotation number, should do much to alleviate discrepancies by rotation number in sample survey estimates of household and economic family units.

As noted in Subsection 4.9, even within household size and month in sample, there are differences in the distributions of respondents and non-respondents

by labour force status. For the LFS there may be advantages in utilizing some variables relating to labour force activities in the adjustment process. There are two factors which tend to preclude this as being viable in practice. Namely, there is a desire for a general weight adjustment, not only for the LFS but also for the various supplementary surveys, and secondly, information at this level of disaggregation would be very unstable and necessitate adjustments at higher levels of aggregation. This new level of adjustment would negate any advantages which may currently be experienced due to local labour market phenomenon. Any compensation procedures must bear in mind the relatively low level of non-response currently experienced for the LFS. This has implications on the level of sophistication warranted, the potential for impact on the estimates, and the reliability of non-response information which would form a key part of the procedure.

There are a range of possible alternatives to the present method of compensating for non-response. Further work in the development of other feasible compensation strategies is a two-staged process. The first stage is the simulation and evaluation of monthly labour force estimates based on the imputation strategy suggested in this report. The second stage is the development of other non-response adjustment strategies followed by their empirical evaluation. Such work is in fact under way.

FOOTNOTES

[1] The estimates provided by the Labour Force Survey refer to the specific week covered by the survey each month, Reference Week, normally the week containing the 15th day. Survey Week, when all interviews are conducted, is the week immediately following Reference Week.

[2] The survey universe for the Labour Force Survey is all persons in the population aged 15 years of age or over residing in Canada, with the exception of the following: residents of the Yukon and the Northwest Territories, persons living on Indian Reserves, inmates of institutions and full-time members of the Armed Forces.

[3] Each month the interviewer is required to indicate whether a complete interview was obtained, that is, a complete Labour Force Survey questionnaire was completed for each eligible household member; a partial interview was obtained, that is a questionnaire was completed for some but not all eligible household members; or no interview was obtained. When no interview occurs, the interviewer must indicate the reason for this. Non-respondent households include those where no one was home (after several calls), the household refused to respond, the household was temporarily absent, or the interview was prevented by weather conditions, death, sickness, a language problem or other unusual circumstances in the household. Vacant dwellings include unoccupied dwellings, seasonal dwellings, dwellings under construction, dwellings occupied by persons not to be interviewed, and dwellings demolished, converted to business premises, moved, abandoned (unfit for habitation), or listed in error.

[4] For further detail on the LFS non-response adjustment see "Methodology of the Canadian Labour Force Survey, (1976)", Statistics Canada, Catalogue 71-526 Occasional, October 1977, pp. 67-68.

[5] For further detail on the LFS weighting process see "Methodology of the Canadian Labour Force Survey, (1976)", Statistics Canada, Catalogue 71-526 Occasional, October 1977, pp. 65-74.

TABLE 1. Percentage Distributions for Respondent and Non-respondent Households by Month in Sample for 1980 and 1981, Canada

Month in sample	Total	Respondent	Non-respondent	Non-response rate
1980				
1	16.6	16.1	28.6	6.94
2	16.6	16.7	15.9	3.84
3	16.7	16.8	14.4	3.47
4	16.7	16.8	14.3	3.45
5	16.7	16.8	14.2	3.42
6	16.8	16.9	12.6	3.03
Total	100	100	100	4.02
1981				
1	16.6	16.0	32.1	6.66
2	16.7	16.7	16.6	3.42
3	16.7	16.8	14.4	2.96
4	16.7	16.8	13.9	2.83
5	16.7	16.9	12.1	2.48
6	16.7	16.9	11.0	2.25
Total	100	100	100	3.43

TABLE 2. Percentage Distributions of Respondent and Non-respondent Households and Non-Response Rates by Household Size for 1980 and 1981 Annual Averages, Canada

	1980				1981			
	Total	Respondent	Non-respondent	Non-response rate	Total	Respondent	Non-respondent	Non-response rate
Size of household								
1	19.0	18.3	35.4	7.48	19.9	19.2	38.1	6.58
2	29.3	29.2	32.8	4.50	29.8	29.7	32.6	3.76
3	17.8	18.2	13.4	3.00	17.6	17.8	11.9	2.33
4	18.8	19.1	11.4	2.44	18.8	19.1	10.4	1.90
5+	14.9	15.2	7.0	1.89	14.0	14.2	6.9	1.69
Total	100	100	100	4.02	100	100	100	3.43
Average household size	2.91	2.93	2.26		2.86	2.88	2.19	

TABLE 3. Percentage Distributions of Respondent and Non-respondent Households by Household Size for Month in Sample for 1980 and 1981 Annual Averages, Canada

Month in sample	Household size 1980						1981					
	1	2	3	4	5+	Total	1	2	3	4	5+	Total
Respondent households												
1	17.4	29.1	18.3	19.5	15.7	100.0	18.3	29.6	17.9	19.5	14.7	100.0
2	18.1	29.2	18.2	19.1	15.4	100.0	19.0	29.8	17.7	19.2	14.3	100.0
3	18.4	29.2	18.1	19.1	15.2	100.0	19.2	29.7	17.8	19.0	14.3	100.0
4	18.5	29.1	18.2	19.1	15.1	100.0	19.5	29.7	17.8	18.9	14.2	100.0
5	18.7	29.1	18.1	19.0	15.0	100.0	19.7	29.7	17.7	18.9	14.0	100.0
6	18.9	29.2	18.1	19.0	14.9	100.0	19.8	29.8	17.7	18.8	13.9	100.0
Non-respondent households												
1	36.1	32.3	13.2	11.2	7.2	100.0	37.7	32.9	12.8	10.3	6.3	100.0
2	38.9	32.7	12.1	9.6	6.7	100.0	40.5	32.7	11.5	9.1	6.2	100.0
3	35.9	32.2	13.5	11.8	6.6	100.0	41.5	33.0	9.6	9.4	6.5	100.0
4	34.9	33.2	13.4	11.7	6.9	100.0	37.6	31.7	12.5	11.0	7.1	100.0
5	32.9	33.9	13.8	12.5	6.9	100.0	36.5	32.6	12.0	10.9	8.0	100.0
6	32.1	33.0	14.9	12.4	7.7	100.0	34.1	32.4	12.3	12.7	8.5	100.0
Non-response rates												
1	13.39	7.64	5.10	4.10	3.30	6.94	12.81	7.34	4.85	3.63	2.97	6.66
2	7.90	4.28	2.59	1.97	1.71	3.84	7.03	3.74	2.25	1.65	1.51	3.42
3	6.56	3.81	2.61	2.17	1.54	3.47	6.19	3.28	1.62	1.49	1.37	2.96
4	6.32	3.92	2.56	2.15	1.61	3.45	5.32	3.02	2.01	1.67	1.44	2.83
5	5.87	3.97	2.63	2.28	1.61	3.42	4.50	2.72	1.70	1.45	1.43	2.48
6	5.05	3.41	2.51	2.00	1.59	3.03	3.82	2.45	1.58	1.53	1.39	2.25
Total	7.48	4.50	3.00	2.44	1.89	4.02	6.58	3.76	2.33	1.90	1.69	3.43

TABLE 4a. Determination of Family Type Composition Variable

Code	Number of economic family units in the household	Size of economic family unit	Age of head of family unit	Presence of children in the household	Head is a member of a married couple
1	1	1	25		
2	1	1	25-64		
3	1	1	65+		
4	1	2	45	No	Yes
5	1	2	45	No	Yes
6	1	2+	45	Yes	Yes
7	1	2+	45	Yes	Yes
8	1	2+		No	Yes
9	1	2+		No	No
10	1	2+		Yes	No
11	2+	all of size 1			
12	2+	all of size 2+			
13	2+	mixed			

TABLE 4b. Percentage Distribution of Respondent and Non-respondent Households by Economic Family Type for 1980 and 1981 Annual Average, Canada

	1980			1981		
Economic family type	Non-respondent households	Respondent households	Non-response rate	Non-respondent households	Respondent households	Non-response rate
1	5.8	2.3	9.80	5.7	2.4	7.82
2	21.0	9.5	8.51	23.4	9.9	7.71
3	8.6	6.6	5.14	9.1	7.0	4.47
4	9.5	8.0	4.72	9.5	8.0	4.05
5	15.5	13.6	4.57	14.4	13.8	3.57
6	18.4	28.1	2.67	16.3	27.0	2.10
7	4.9	9.9	2.04	4.7	9.0	1.83
8	4.6	8.2	2.32	4.1	8.6	1.65
9	3.1	4.3	2.99	3.1	4.3	2.45
10	3.9	4.8	3.30	5.0	4.9	3.47
11	3.4	2.7	5.05	3.5	2.8	4.25
12	0.0	0.1	1.25	0.1	0.1	2.29
13	1.2	2.1	2.47	1.2	2.1	2.06
Total	100.0	100.0		100.0	100.0	

TABLE 5. Percentage Distributions of Individuals by Age Groups by Response Status for 1980 and 1981 Annual Averages, Canada

	1980			1981		
Age group	Respondent	Non-respondent	Non-response rate	Respondent	Non-respondent	Non-response rate
0-14	24.3	19.4	2.50	23.6	18.9	2.12
15-19	9.7	7.5	2.42	9.5	6.9	1.92
20-24	9.1	10.4	3.54	9.4	10.9	3.04
25-44	29.2	31.6	3.38	29.6	33.0	2.91
45-64	18.9	20.9	3.45	19.1	19.9	2.74
65+	8.7	10.2	3.65	8.9	10.5	3.08
Total	100.0	100.0	3.13	100.0	100.0	2.63

TABLE 6a. Percentage Distribution of Individuals by Age Group and Non-response Rates for Household Size and Response Status for 1980 Annual Averages, Canada

Age group	Household size					
	1	2	3	4	5+	Total
Respondent						
0-14	0.0	2.6	20.5	35.1	37.1	24.3
15-19	1.9	3.5	8.2	9.7	16.7	9.7
20-24	10.5	14.2	11.4	6.0	6.7	9.1
25-44	27.9	25.3	31.3	35.4	25.2	29.2
45-64	25.4	31.6	23.5	12.4	11.8	18.9
65+	34.3	22.8	5.2	1.3	2.6	8.7
Total	100.0	100.0	100.0	100.0	100.0	100.0
Non-respondent						
0-14	0.3	2.6	22.6	37.0	40.8	19.4
15-19	3.0	3.8	8.1	8.8	15.7	7.5
20-24	13.6	14.4	12.0	4.8	5.7	10.4
25-44	35.5	27.3	33.5	37.4	26.6	31.6
45-64	25.2	33.0	19.6	10.7	10.2	20.9
65+	22.4	19.1	4.1	1.4	1.1	10.2
Total	100.0	100.0	100.0	100.0	100.0	100.0
Non-response rates						
0-14	-	4.52	3.30	2.57	2.05	2.50
15-19	11.07	4.83	2.93	2.22	1.76	2.42
20-24	9.49	7.94	3.16	1.95	1.60	3.54
25-44	9.33	4.83	3.21	2.58	1.98	3.38
45-64	7.45	4.68	2.52	2.11	1.62	3.45
65+	5.01	3.80	2.41	2.47	0.85	3.65
Total	7.48	4.50	3.00	2.44	1.87	3.13

TABLE 6b. Percentage Distribution of Individuals and Non-response Rates by Age Group for Household Size and Response Status for 1981 Annual Averages, Canada

Age group	Household size					
	1	2	3	4	5+	Total
Respondent						
0-14	0.0	2.7	19.8	34.4	36.9	23.6
15-19	1.8	3.4	8.6	9.9	16.1	9.5
20-24	10.6	14.1	11.4	6.4	7.1	9.4
25-44	28.8	25.9	31.8	35.2	25.8	29.6
45-64	24.4	31.7	23.5	12.6	11.6	19.1
65+	34.3	22.2	5.0	1.5	2.5	8.9
Total	100.0	100.0	100.0	100.0	100.0	100.0
Non-respondent						
0-14	0.1	3.8	23.6	37.2	39.4	18.9
15-19	2.0	3.1	7.5	8.8	15.6	6.9
20-24	13.0	15.1	12.0	5.9	5.8	10.9
25-44	39.6	28.6	34.3	37.2	27.7	33.0
45-64	22.9	30.1	19.8	10.1	10.2	19.9
65+	22.3	19.3	2.8	0.9	1.2	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0
Non-response rates						
0-14	-	5.16	2.77	2.05	1.77	2.12
15-19	7.15	3.47	2.03	1.69	1.61	1.92
20-24	7.94	4.02	2.47	1.77	1.37	3.04
25-44	8.85	4.14	2.51	2.01	1.78	2.91
45-64	6.20	3.57	1.97	1.53	1.46	2.74
65+	4.38	3.27	1.31	1.15	0.83	3.08
Total	6.58	3.76	2.33	1.90	1.66	2.63

TABLE 7a. Percentage Distribution of Individuals and Non-response Rates by Age Group for Month in Sample and Response Status for 1980 Annual Averages, Canada

Age group	Month in sample						
	1	2	3	4	5	6	Total
Respondent							
0-14	24.2	24.1	24.3	24.4	24.5	24.5	24.3
15-19	9.9	9.8	9.7	9.7	9.7	9.6	9.7
20-24	9.2	9.2	9.2	9.1	9.1	9.0	9.1
25-44	29.1	29.1	29.2	29.3	29.2	29.1	29.2
45-64	18.9	18.9	18.9	18.9	18.9	19.0	18.9
65+	8.7	8.8	8.7	8.7	8.7	8.7	8.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Non-respondent							
0-14	18.9	19.0	19.6	19.5	20.0	19.9	19.4
15-19	7.6	6.3	7.8	7.7	7.6	7.9	7.5
20-24	10.6	10.3	10.1	10.3	10.5	10.3	10.4
25-44	31.9	33.1	30.8	30.6	31.4	31.5	31.6
45-64	20.4	20.6	21.8	21.5	20.9	21.0	20.9
65+	10.6	10.8	10.0	10.4	9.6	9.5	10.2
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Non-response rates							
0-14	4.22	2.25	2.18	2.16	2.24	2.00	2.50
15-19	4.12	1.84	2.15	2.16	2.17	2.03	2.42
20-24	6.12	3.16	2.94	3.04	3.16	2.78	3.54
25-44	5.83	3.21	2.83	2.81	2.93	2.65	3.38
45-64	5.74	3.08	3.09	3.06	3.00	2.71	3.45
65+	6.40	3.50	3.06	3.22	3.01	2.69	3.65
Total	5.34	2.84	2.69	2.70	2.73	2.46	3.13

TABLE 7b. Percentage Distribution of Individuals and Non-response Rates by Age Group for Month in Sample and Response Status for 1981 Annual Averages, Canada

Age group	Month in sample						
	1	2	3	4	5	6	Total
Respondent							
0-14	23.4	23.4	23.5	23.6	23.7	23.8	23.6
15-19	9.7	9.6	9.5	9.5	9.4	9.3	9.5
20-24	9.4	9.4	9.4	9.4	9.4	9.3	9.4
25-44	29.5	29.7	29.7	29.6	29.6	29.6	29.6
45-64	19.1	19.1	19.0	19.1	19.1	19.1	19.1
65+	8.9	8.9	8.9	8.9	8.9	8.9	8.9
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Non-respondent							
0-14	18.1	18.7	18.3	19.8	19.8	20.1	18.9
15-19	6.9	6.2	6.8	6.6	7.3	7.6	6.9
20-24	10.9	10.5	11.1	11.0	11.2	10.7	10.9
25-44	33.5	33.2	32.0	32.6	33.1	32.6	33.0
45-64	20.2	20.5	20.8	19.2	18.9	19.3	19.9
65+	10.4	11.0	11.0	10.8	9.8	9.7	10.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Non-response rates							
0-14	3.96	2.03	1.71	1.85	1.64	1.57	2.12
15-19	3.67	1.66	1.57	1.55	1.54	1.50	1.92
20-24	5.83	2.80	2.55	2.58	2.34	2.12	3.04
25-44	5.67	2.82	2.35	2.42	2.20	2.03	2.91
45-64	5.32	2.71	2.38	2.21	1.94	1.86	2.74
65+	5.88	3.10	2.68	2.66	2.15	2.01	3.08
Total	5.05	2.53	2.18	2.20	1.96	1.85	2.63

TABLE 8a. Percentage Distribution of Individuals by Labour Force Status for Month in Sample, Household Size and Response Status for 1980 and 1981 Annual Average, Canada

		Month in Sample																	
		1			2			3			4			5			6		
Size of Household	Response Status	E	U	N	E	U	N	E	U	N	E	U	N	E	U	N	E	U	N
1980																			
1	Respondent	51.0	3.2	45.8	52.3	3.2	44.5	52.2	3.2	44.6	52.2	3.2	44.5	52.4	3.3	44.5	52.2	3.2	44.6
	Non-respondent	65.0	4.4	30.6	62.8	4.9	32.3	60.2	6.8	33.0	60.0	6.3	33.8	60.9	6.2	32.9	57.0	8.1	34.9
2	Respondent	55.5	4.1	40.3	55.7	4.2	40.2	56.1	3.9	40.1	56.0	3.9	40.0	55.6	4.0	40.4	55.5	4.0	40.5
	Non-respondent	58.0	4.0	38.0	56.0	4.1	39.8	56.3	4.6	39.1	55.2	5.3	39.5	58.4	5.1	36.5	60.3	4.9	34.8
3	Respondent	61.7	5.5	32.9	61.9	5.2	32.9	61.9	5.2	33.0	61.9	5.2	32.9	61.6	5.4	33.0	61.9	5.2	33.0
	Non-respondent	62.4	4.5	33.1	57.9	6.7	35.4	59.5	5.0	35.5	58.4	5.3	36.3	61.6	5.9	32.4	58.7	6.3	35.0
4	Respondent	64.1	5.2	30.7	64.2	5.0	30.8	65.0	4.6	30.5	64.8	4.8	30.5	65.1	4.7	30.2	65.1	4.6	30.3
	Non-respondent	64.6	4.6	30.8	64.1	5.1	30.7	57.8	6.5	35.7	63.3	4.9	31.8	62.0	6.2	31.8	61.7	8.1	30.2
5+	Respondent	58.0	5.9	36.1	58.0	5.8	36.2	58.1	5.7	36.3	58.4	5.5	36.2	58.6	5.5	36.0	58.4	5.7	35.9
	Non-respondent	56.8	5.6	37.6	57.9	4.7	37.4	57.1	4.0	39.0	58.4	5.9	35.7	56.6	8.4	35.1	60.1	6.2	33.7
Total	Respondent	58.9	5.0	36.1	59.1	4.9	36.0	59.3	4.7	36.0	59.4	4.7	36.0	59.3	4.7	36.0	59.3	4.7	36.0
	Non-respondent	61.0	4.5	34.6	59.2	4.9	36.0	58.0	5.3	36.7	58.4	5.5	36.1	59.8	6.0	34.2	59.7	6.4	33.9
1981																			
1	Respondent	50.5	3.6	46.0	51.9	3.2	44.8	52.1	3.5	44.5	52.2	3.3	44.6	51.9	3.4	44.7	51.7	3.5	44.8
	Non-respondent	63.7	3.7	32.5	63.4	4.1	32.5	62.3	3.5	34.2	60.8	6.4	32.8	59.0	4.8	36.2	60.8	5.6	33.7
2	Respondent	55.7	3.9	40.4	56.1	3.8	40.1	56.1	3.7	40.3	56.0	3.7	40.3	55.8	3.8	40.4	55.7	3.9	40.4
	Non-respondent	60.4	3.5	36.1	56.9	4.2	38.9	58.4	4.5	37.0	56.4	4.7	38.9	59.2	5.9	35.0	56.5	5.0	38.5
3	Respondent	63.3	5.3	31.4	63.4	5.2	31.4	63.9	5.1	31.1	63.6	5.0	31.4	63.3	5.2	31.5	63.2	5.2	31.6
	Non-respondent	66.6	4.8	28.6	65.5	5.3	29.2	63.8	6.5	29.8	65.9	5.8	28.3	61.5	7.1	31.4	66.1	6.7	27.2
4	Respondent	64.8	5.1	30.1	65.1	4.9	30.0	65.1	5.1	29.8	65.4	4.9	29.6	65.5	4.8	29.6	65.6	5.0	29.4
	Non-respondent	63.4	4.2	32.4	66.4	4.3	29.3	61.6	3.9	34.6	63.3	5.6	31.2	62.3	6.8	30.9	63.2	6.9	30.0
5+	Respondent	59.5	6.1	34.4	59.6	5.8	34.6	59.8	5.8	34.5	60.2	5.6	34.2	60.4	5.4	34.2	60.2	5.9	34.0
	Non-respondent	62.3	4.8	32.9	60.1	5.6	34.3	60.2	5.4	34.3	58.5	5.3	36.2	61.6	7.5	31.0	61.8	6.7	31.5
Total	Respondent	59.7	4.9	35.3	60.0	4.8	35.3	60.1	4.7	35.1	60.2	4.6	35.2	60.1	4.6	35.2	60.0	4.8	35.2
	Non-respondent	62.8	4.0	33.2	61.3	4.5	34.2	60.7	4.5	34.8	60.2	5.5	34.4	60.3	6.2	33.5	60.7	5.9	33.4

TABLE 8b. Unemployment Rates by Household Size, for Month in Sample and Response Status for 1980 and 1981 Annual Averages, Canada

Household size	Response status	Month in sample 1	2	3	4	5	6
1980							
1	Respondent	5.88	5.74	5.77	5.83	5.99	5.79
	Non-respondent	6.36	7.16	10.16	9.44	9.29	12.45
2	Respondent	6.94	6.95	6.46	6.54	7.01	6.75
	Non-respondent	6.39	6.81	7.58	8.77	7.97	7.56
3	Respondent	8.14	7.76	7.71	7.71	8.04	7.69
	Non-respondent	6.75	10.37	7.74	8.27	8.79	9.65
4	Respondent	7.54	7.24	6.55	6.84	6.71	6.60
	Non-respondent	6.71	7.42	10.09	7.22	9.05	11.58
5+	Respondent	9.21	9.10	8.88	8.54	8.57	8.92
	Non-respondent	8.92	7.49	6.50	9.19	12.85	9.36
Total	Respondent	7.83	7.63	7.27	7.28	7.37	7.35
	Non-respondent	6.82	7.63	8.43	8.60	9.13	9.72
1981							
1	Respondent	6.56	5.86	6.22	5.87	6.16	6.38
	Non-respondent	5.52	6.07	5.29	9.46	7.53	8.39
2	Respondent	6.57	6.38	6.11	6.13	6.37	6.53
	Non-respondent	5.51	6.93	7.18	7.69	9.06	8.09
3	Respondent	7.77	7.59	7.35	7.31	7.57	7.60
	Non-respondent	6.76	7.46	9.19	8.12	10.33	9.25
4	Respondent	7.30	7.02	7.28	7.02	6.87	7.12
	Non-respondent	6.28	6.05	5.88	8.07	9.77	9.80
5+	Respondent	9.24	8.90	8.78	8.57	8.26	8.89
	Non-respondent	7.16	8.58	8.27	8.35	10.84	9.76
Total	Respondent	7.64	7.33	7.28	7.14	7.16	7.42
	Non-respondent	6.05	6.89	6.95	8.30	9.36	8.91

TABLE 9a. Percentage Distribution of Households by Size for Response Type, Type of Area, and Number of Months in Sample for 1980 Annual Averages, Canada

Month in sample	Household Size	SRU Respondent	Non-respondent	NSRU urban Respondent	Non-respondent	NSRU rural Respondent	Non-respondent	Total Respondent	Non-respondent
1	1	19.0	39.7	17.8	30.9	10.3	22.1	17.4	36.1
	2	29.4	31.9	29.6	32.8	27.6	34.0	29.1	32.3
	3	18.3	12.3	17.3	14.7	18.7	16.1	18.3	13.2
	4	19.1	10.3	19.0	12.5	21.6	15.2	19.5	11.2
	5+	14.2	5.9	16.3	9.2	21.9	12.7	15.7	7.2
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	1	19.8	42.7	18.1	32.3	10.7	24.7	18.1	38.9
	2	29.5	32.0	29.4	33.2	27.7	35.4	29.2	32.7
	3	18.2	11.0	17.5	13.8	18.6	16.0	18.2	12.1
	4	18.7	8.7	18.9	11.8	21.3	12.7	19.1	9.6
	5+	13.8	5.5	16.1	9.0	21.8	11.2	15.4	6.7
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
3	1	20.1	39.4	18.4	33.4	10.8	22.7	18.4	35.9
	2	29.6	32.1	29.6	30.2	27.7	33.8	29.2	32.2
	3	18.0	12.6	17.1	15.0	18.6	16.1	18.1	13.5
	4	18.7	10.9	19.1	13.7	21.2	14.9	19.1	11.8
	5+	13.6	5.1	15.8	7.8	21.6	12.5	15.2	6.6
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
4	1	20.3	37.9	18.6	35.0	11.0	22.0	18.5	34.9
	2	29.5	33.6	29.4	29.6	27.7	33.7	29.1	33.2
	3	18.1	12.4	17.3	12.8	18.7	17.4	18.2	13.4
	4	18.6	10.6	19.0	14.7	21.2	14.0	19.1	11.7
	5+	13.5	5.5	15.7	7.9	21.4	12.9	15.1	6.9
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	1	20.4	36.1	19.1	33.3	11.0	19.8	18.7	32.9
	2	29.5	34.3	29.1	31.7	27.7	33.9	29.1	33.9
	3	18.1	12.8	17.2	13.6	18.6	17.7	18.1	13.8
	4	18.4	11.5	19.1	13.8	21.4	15.7	19.0	12.5
	5+	13.5	5.2	15.6	7.7	21.2	12.9	15.0	6.9
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
6	1	20.6	34.7	19.2	33.3	11.2	21.3	18.9	32.1
	2	29.7	32.8	28.9	36.5	27.6	31.6	29.2	33.0
	3	18.1	14.1	17.1	13.0	18.6	19.3	18.1	14.9
	4	18.4	11.6	19.2	11.2	21.4	15.5	19.0	12.4
	5+	13.2	6.7	15.6	6.0	21.2	12.3	14.9	7.7
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLE 9b. Percentage Distribution of Households by Size for Response Type, Type of Area, and Number of Months in Sample for 1981 Annual Averages, Canada

Month in sample	Household size	SRU Respondent	SRU Non-respondent	NSRU urban Respondent	NSRU urban Non-respondent	NSRU rural Respondent	NSRU rural Non-respondent	Total Respondent	Total Non-respondent
1	1	20.1	40.4	17.5	34.1	11.2	22.2	18.3	37.7
	2	29.9	33.2	30.3	32.5	28.1	33.8	29.6	32.9
	3	17.7	11.8	17.6	15.2	18.2	18.5	17.9	12.8
	4	19.1	9.1	19.5	11.1	21.4	14.7	18.5	10.3
	5+	13.2	5.6	15.2	7.1	21.0	10.8	14.7	6.3
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	1	20.9	43.2	18.3	34.1	11.4	25.7	19.0	40.5
	2	30.2	33.6	30.3	32.8	28.1	32.5	29.8	32.7
	3	17.4	10.7	17.5	11.1	18.4	16.4	17.7	11.5
	4	18.7	7.2	19.1	13.9	21.3	16.3	19.2	9.1
	5+	12.8	5.4	14.7	8.2	20.8	9.2	14.3	6.2
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
3	1	21.1	44.2	18.5	38.2	11.5	26.0	19.2	41.5
	2	30.1	33.3	30.1	34.5	28.2	31.5	29.7	33.0
	3	17.5	8.7	17.7	8.2	18.4	15.2	17.8	9.6
	4	18.5	8.5	19.3	8.2	21.4	16.2	19.0	9.4
	5+	12.8	5.4	14.4	11.0	20.6	11.1	14.3	6.5
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
4	1	21.4	40.4	18.7	35.5	11.7	25.1	19.5	37.6
	2	30.2	31.0	30.1	33.1	28.1	34.3	29.7	31.7
	3	17.4	11.8	17.7	12.3	18.4	14.3	17.8	12.5
	4	18.3	10.5	19.2	11.1	21.4	15.9	18.9	11.0
	5+	12.7	6.3	14.4	8.0	20.5	10.4	14.2	7.1
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
5	1	21.6	39.2	18.8	31.1	11.7	24.7	19.7	36.5
	2	30.2	32.6	30.2	34.7	28.2	33.8	29.7	32.6
	3	17.4	10.7	17.5	13.8	18.3	15.9	17.7	12.0
	4	18.2	9.7	19.0	14.4	21.6	15.2	18.9	10.9
	5+	12.6	7.9	14.4	6.0	20.3	10.5	14.0	8.0
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
6	1	21.7	37.0	18.9	28.7	11.9	23.6	19.8	34.1
	2	30.2	31.4	30.1	40.0	28.2	33.6	29.8	32.4
	3	17.5	11.7	17.6	11.1	18.3	13.6	17.7	12.3
	4	18.2	11.7	19.2	12.9	21.5	18.2	18.8	12.7
	5+	12.4	8.2	14.3	7.3	20.1	11.0	13.9	8.5
	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

TABLE 9c. Household Non-response Rates by Type of Area, Household Size, and Month in Sample for 1980 and 1981 Annual Averages, Canada

		Type of area					
		1980			1981		
Month in sample	Household size	SRU	NSRU urban	NSRU rural	SRU	NSRU urban	NSRU rural
1	1	13.99	11.02	11.62	13.57	11.79	9.37
	2	7.82	7.31	7.05	7.98	6.84	5.90
	3	4.98	6.05	5.03	4.93	5.62	5.02
	4	4.03	4.48	4.15	3.59	3.76	3.45
	5+	3.13	4.01	3.45	3.21	3.10	2.61
	Total	5.32	6.65	5.79	7.25	6.42	4.95
2	1	8.21	6.79	7.20	6.30	6.32	6.56
	2	4.30	4.41	4.13	3.96	3.77	3.47
	3	2.44	3.12	2.82	2.24	2.24	2.69
	4	1.90	2.48	1.96	1.41	2.56	2.31
	5+	1.63	2.23	1.71	1.53	1.97	1.36
	Total	3.98	3.93	3.26	3.59	3.50	3.01
3	1	6.68	6.33	6.50	6.33	5.61	5.52
	2	3.83	3.67	3.88	3.44	3.20	2.81
	3	2.48	3.18	2.78	1.58	1.32	2.10
	4	2.08	2.60	2.27	1.46	1.21	1.93
	5+	1.35	1.81	1.87	1.35	2.16	1.37
	Total	3.53	3.60	3.20	3.13	2.81	2.52
4	1	6.41	6.58	6.04	5.50	5.26	5.02
	2	4.00	3.63	3.77	3.06	3.11	2.90
	3	2.44	2.71	2.91	2.05	1.99	1.87
	4	2.05	2.82	2.07	1.74	1.66	1.79
	5+	1.47	1.85	1.90	1.51	1.59	1.23
	Total	3.53	3.61	3.11	2.99	2.83	2.39
5	1	6.04	5.94	5.51	4.59	4.47	4.39
	2	4.05	3.81	3.81	2.79	3.15	2.55
	3	2.51	2.79	2.99	1.60	2.18	1.86
	4	2.22	2.56	2.34	1.40	2.10	1.52
	5+	1.41	1.75	1.94	1.64	1.17	1.12
	Total	3.51	3.50	3.14	2.59	2.75	2.14
6	1	5.12	5.54	5.30	3.92	3.48	4.24
	2	3.43	4.08	3.25	2.42	3.06	2.60
	3	2.45	2.51	2.95	1.58	1.48	1.63
	4	1.99	1.93	2.07	1.52	1.58	1.86
	5+	1.61	1.28	1.68	1.55	1.19	1.21
	Total	3.11	3.26	2.85	2.34	2.32	2.19

ROTATION GROUP BIAS IN THE LFS ESTIMATES[1]

P.D. GHANGURDE[2]

The paper attempts to evaluate the impact of non-response adjustment by rotation groups on rotation group bias in the estimates from the Canadian Labour Force Survey. Results on bias and non-response characteristics are presented and discussed. An index used to measure rotation group bias is given and some empirical results are analyzed.

1. INTRODUCTION

In the Canadian Labour Force Survey (LFS) sample design each month one-sixth of the households rotate out of the sample and one-sixth rotate in. The sample is thus composed of six panels or rotation groups. In any given month households in a rotation group have been in the survey from one to six months, including the current month. It is well-known that in household surveys with rotation sample designs estimates for the same characteristics from different rotation groups could have different expected values. This phenomenon, called rotation group bias, has been studied for the LFS and other household surveys with rotation sample designs (see [1], [5], [7] and [8]).

Rotation group bias can be attributed to several factors. In the LFS the non-response rates at household level are known to differ between rotation groups i.e. number of months a household is in the survey. It is also known that non-respondent households tend to have different characteristics as compared to respondent households. Both these factors can contribute to bias. Due to conditioning of the respondent or familiarity with the survey

[1] Presented at the American Statistical Association Meeting in Cincinnati, August 1982.

[2] P.D. Ghangurde, Census and Household Survey Methods Division, Statistics Canada.

over a period of six months, response bias in the data from successive months can be of different magnitude. There is some evidence from the LFS reinterview data of such differential bias over the period of six months. However, in the literature it has also been hypothesized that rotation group biases can be attributed to differences in non-response probabilities between rotation groups [7]. Although individual probabilities are not known, their averages can be estimated by non-response rates.

In this paper an attempt is made to evaluate the impact on rotation group bias of non-response adjustment by rotation groups. In section 2 some results on bias are introduced and their implications on the bias in the estimates from different rotation groups are discussed. Section 3 presents some data on nonresponse rates in the LFS and characteristics of respondents and non-respondents by months in the survey and their contribution to rotation group bias. Section 4 explains the adjustment of LFS weight for non-response by rotation groups and its impact on the rotation group bias and an index used as a measure of rotation group bias. In section 5, some data on the index for labour force status categories, based on 1981 surveys, are analyzed.

2. THE STATISTICAL MODEL

We introduce a model which provides expressions for contribution to bias of differences in non-response rates, differences in characteristics of respondents and non-respondents and response bias for any groups of the sample in which adjustment of weight for non-response can be done. Rotation groups can be considered as a particular case of these groups.

A population of size N is assumed to be divided into "strata" of respondents and non-respondents of sizes N_1 and N_2 respectively. A simple random sample of size n is drawn and responses are obtained from n_1 units and $(n-n_1)$ units are non-respondents.

Suppose the sample can be divided into K groups such that non-response rates and characteristics of respondents and non-respondents differ between the groups. The data collection methods used in these groups and the extent of conditioning of respondents or their familiarity with the survey could be

different leading to differences in non-response rates and characteristics and also possibly to different response biases. By an extension of a result in [2] and [6] to include response bias component, the bias of the sample mean $\bar{y}$ of n_1 units (without adjustment of weight for non-response within groups) is given by

$$B(\bar{y}) = \frac{1}{\bar{R}} \sum_{i=1}^{K} P_i \bar{Y}_{1i} (R_i - \bar{R}) + \sum_{i=1}^{K} P_i (1-R_i) (\bar{Y}_{1i} - \bar{Y}_{2i})$$

$$+ \frac{1}{\bar{R}} \sum_{i=1}^{K} \bar{P}_i R_i \beta_i, \qquad (1)$$

where $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$ are population means of respondents and non-respondents in the ith group, R_i, response rate for the ith group, P_i, proportion of total population in the ith group, $\bar{\beta}_i$, mean response bias in the ith group and $\bar{R} = \sum_{i=1}^{K} P_i R_i$, overall response rate.

The above expression shows the decomposition of bias into three components. The first shows contribution of differential response rates, the second due to differences in characteristics between respondents and non-respondents and the third due to response bias. For simplicity, we consider in this paper characteristics based on attributes, e.g., proportions of "employed" and "unemployed". We now consider the estimate $\bar{y}_a$, with adjustment for non response by inverse of response rate done within each group. Thus

$$\bar{y}_a = \frac{1}{n} \sum_{i=1}^{K} n_{.i} \bar{y}_i,$$

where $n_{.i}$ is sample size in the ith group and $\bar{y}_i$ is mean of n_{1i} units in the ith group. The bias of $\bar{y}_a$ is given by

$$B(\bar{y}_a) = \sum_{i=1}^{K} P_i (1-R_i) (\bar{Y}_{1i} - \bar{Y}_{2i}) + \sum_{i=1}^{K} P_i \bar{\beta}_i. \qquad (2)$$

The first component of bias in (1) due to differential response rates between groups is eliminated, the second component due to differences in characteristics remains the same and the third component due to response bias could be different from that in (1).

Based on a framework of response non-response error model involving response probabilities at unit level, the bias has been decomposed into components due to non-response and response errors [3]. The above decomposition of bias does not use response probabilities at the level of individual units but is simple enough for empirical evaluation of the components.

If response rates do not differ between the groups the first component is zero so that, (1) is identical to (2); hence non-response adjustment within the groups does not lead to reduction in bias. The difference in the bias of $\overline{y}$ and $\overline{y}_a$ is given by

$$B(\overline{y}) - B(\overline{y}_a) = \frac{1}{\overline{R}} \sum_{i=1}^{K} P_i \, (\overline{R}_i - \overline{R}) \, (\overline{Y}_{1i} + \overline{\beta}_i). \qquad (3)$$

Thus if response rates are different, and $\overline{Y}_{1i}$ and $\overline{\beta}_i$ do not differ between the groups, there is no change in the bias after non-response adjustment within the groups. If the means $\overline{Y}_{1i}$ and $\overline{\beta}_i$ differ between the groups there is a decrease in bias if the term on the right-hand side of (3) is positive and an increase, if it is negative. The change in absolute bias from $|B(\overline{y})|$ to $|B(\overline{y}_a)|$ as result of adjustment will depend upon the sign and magnitude of the term on the right hand side of (3).

The bias of estimate of mean for i^{th} rotation group, without adjustment and with adjustment of weight for non-response by rotation groups, is obtained from (1) and (2) by simple substitution of $P_i = 1$ and keeping the terms corresponding to the rotation group. Also, from (3) the difference in biases of estimate for i^{th} rotation group is given by

$$B(\overline{y}_i) - B(\overline{y}_{ia}) = \left(\frac{Ri - \overline{R}}{\overline{R}} \right) (\overline{Y}_{1i} + \overline{\beta}_i), \qquad (4)$$

where $\overline{y}_i$ and $\overline{y}_{ia}$ are estimates for i^{th} rotation group before and after adjustment. Assuming $(\overline{Y}_{1i} + \overline{\beta}_i) > 0$ for all i, if $R_i < \overline{R}$, the bias for i^{th} rotation group increases after adjustment and if $R_i > \overline{R}$, it decreases.

Since the population of respondents in a survey month is the same for various rotation groups, it may be argued that the proportions $\overline{Y}_{1i}$ could be the same for all rotation groups or months in the survey. However, the differences in exposure to survey or conditioning of the respondents can produce different response biases, $\overline{\beta}_i$, between rotation groups. Thus the difference in the bias of $\overline{y}$ and $\overline{y}_a$ is given by

$$B(\overline{y}) - B(\overline{y}_a) = \frac{1}{\overline{R}} \sum_{i=1}^{K} P_i \, (R_i - \overline{R}) \, \overline{\beta}_i. \qquad (5)$$

However, the difference in bias of estimates for rotation group i is given by (4).

It may also be noted that under the assumption of constant $\overline{Y}_{1i}$ and $\overline{\beta}_i$ for all i and differential response rates, non-response adjustment by rotation groups does not change the bias of estimate based on all rotation groups. However, the change in the biases of individual rotation groups after non-response adjustment are accounted for by different response rates.

The above results are useful in the evaluation of contribution of various factors to rotation group bias and the impact of adjustment of weight by rotation groups on the estimates of rotation group bias.

The LFS is a monthly national household survey with a sample size of 55,000 households. Each of the ten provinces in Canada is divided into economic regions, which consist of groups of counties with similar economic structure. The economic regions are divided into homogeneous strata on the basis of distribution of employed persons in various industry-occupation groups in the last Census. The sample design is stratified multi-stage sampling with two

stages in the self-representing (SR) urban areas and three or four stages in the non-self-representing (NSR) rural areas of the design. The sample selection in the initial stages is with probability proportional to population size and that in the last stage, where dwellings are selected from clusters, being systematic. The selected clusters are assigned six rotation numbers independently within each stratum. In any survey month one-sixth of the households have been in the survey from 1 to 6 months. Thus the entire sample is divided into six equally representative sub-samples of equal sizes [4]. The rotation numbers for six rotation groups can be converted to number of "months in the survey" by a simple transformation.

The adjustment of weight for non-response is done for the entire sample in balancing units by ratio of households in the sample to responding households. In the NSR areas each primary sampling unit (PSU) is divided into two balancing units consisting of urban and rural parts. In the SR areas of the design, strata (called sub-units) form balancing units. The number of balancing units thus exceeds 900 in NSR areas and 800 in SR areas.

In order to evaluate the rotation group bias in the LFS estimates, with and without adjustment, data on non-response rates $(1-R_i)$ and $\overline{Y}_{1i}$ and $\overline{Y}_{2i}$, proportions for the characteristics "employed" and "unemployed" for respondents and non-respondent respectively in twelve surveys in 1981 are presented and analyzed in Section 3. The "months in the survey" represents number of months (including the current month) a rotation group is in the survey. No data on response biases, $\overline{\beta}_i$, are presented.

3. ANALYSIS OF LFS DATA

Table 1 shows average non-response rates, $(1-R_i)$, by months in the survey for calendar months in 1981. It can be seen that the rates differ substantially between the two areas and between months in the survey for a given area. In both the areas and at Canada level, non-response rates are high in the first month, decrease substantially in the second month and decrease slowly over the succeeding months. The high non-response rates in the first month are contributed by "temporary absent" and "no-one-at home" type households. In the later months the rates reduce due to interviewer's knowledge about the

best time to call on these households. The rates are higher in SR areas, especially apartments (not shown in the table) as compared to NSR areas. During processing, for approximately 1/2% households data are carried forward from the previous month. The non-response rates presented in the tables are obtained by considering those households as respondent. It may be noted that difference of rates from their mean $(R_i - \overline{R})$, is negative in the first and in some cases in the second month in the survey and positive in the following months. The mean rate $\overline{R}$ is approximately equal to R_2. Thus from (4) relative bias for first month in the survey is expected to increase, if $(\overline{Y}_{1i} + \overline{\beta}_i)$ and population mean $\overline{Y}_i$ are assumed constant; for months 3 to 6, the relative bias is expected to decrease after adjustment of weight for non-response.

Table 2 shows estimated proportions, $\hat{\overline{Y}}_{1i}$ and $\hat{\overline{Y}}_{2i}$, of employed and unemployed heads of households by months in the survey for respondent and non-respondent households respectively. The estimates were obtained from LFS longitudinal files for the period March - August 1976 and are based on unweighted counts. The data on non-respondents, who responded at least once during the six month period, were obtained from months in which they responded. Non-respondent households tend to have greater proportion of employed heads and lesser proportion of unemployed heads as compared to respondent households. It is known that the difference of proportions between respondents and non-respondents for employed persons tends to be 0.10 and that for unemployed persons tends to be about 0.005, the signs of differences remaining the same. No particular trend over months in the survey can be observed in the proportions of employed and unemployed heads among respondent and non-respondent households.

The contribution of the first month to the first component is negative in all calendar months for both unemployed and employed. This indicates that the bias for the first month in the survey is expected to increase after adjustment for non-response.

The analysis in sections 2 and 3 isolates rotation groups as groups considered for non-response adjustment. For real data, the same relative changes may not be seen due to impact of differential response rates in other groups and changes in magnitude of $\overline{Y}_{1i}$ and $\overline{\beta}_i$ during the six month period. In section 5,

we analyze the impact of non-response adjustment by rotation groups on rotation group bias in the LFS estimates and attempt to explain the results on the basis of the model.

It may be noted that non-response adjustment in the present weighting of LFS data is done within balancing units which are much smaller than NSR and SR areas within a province. Thus the estimates of rotation group bias based on the present weighting and non-response adjustment are corrected for differential non-response rates between the two areas but not for those between rotation groups.

4. WEIGHT-ADJUSTMENT BY ROTATION GROUPS

The LFS final weight is composed of five factors: (1) mathematical weight, (2) rural-urban factor, (3) cluster sub-weight (4) balancing factor and (5) age-sex factor. The mathematical weight for a household is the inverse of overall sampling ratio for the household, based on the sample design. Within each province the weight is the same within urban (SR) and rural (NSR) strata except in a few cases, resulting in twenty areas at Canada level with the same mathematical weight. The cluster sub-weight is the inverse of sampling ratio within a cluster. The balancing factor adjusts the weight for non-response and age-sex factor is a ratio adjustment factor based on projected population within age-sex groups at province level.

As explained in section 2, adjustment of weight for non-response is done within balancing units for the sample of households. For the evaluation of impact of weight adjustment by rotation groups, it was decided to use progressively smaller areas (as balancing units) starting with rotation groups at province level. The adjustment of final weight within rotation groups in these areas was done by multiplying by adjustment factors:

$$R_{H(i)} = \frac{\text{respondent households in the sample}}{\text{respondent households in rotation group (i)}}$$

$$R_{p(i)} = \frac{\text{respondent persons in the sample}}{\text{respondent persons in rotation group (i)}}$$

The first factor weights up the estimate of households within a rotation group in a balancing unit to the level of sample of respondent households. The balancing factor weights it up to the level of sample of households within the balancing unit. The second factor, based on the count of respondent persons weights up the estimates to the level of the entire sample of respondent persons and thus corrects the estimates for different household sizes or coverage of persons within households. It is known that non-respondent households tend to have smaller sizes as compared to respondent households. The difference in non-response rates between rotation groups may result in differences in average household sizes.

If $\hat{Y}(i)$ is estimates total of i_{th} rotation group and $Y(i)$, true value of i_{th} group total, then the estimate of relative bias of estimated total of i_{th} rotation group is given by

$$B_y(i) = \frac{\hat{Y}(i) - Y(i)}{Y(i)} \; ; \; i = 1,2,\ldots 6. \tag{6}$$

Since $Y(i)$'s are not known and can be assumed to be approximately equal (since rotation groups have equal expected sizes at large area level) $\hat{\bar{Y}}(.)$, the mean of six rotation group total estimates can be used in place of $Y(i)$. The rotation group bias index for i_{th} rotation group is given by

$$I_y(i) = \frac{\hat{Y}(i)}{\hat{\bar{Y}}(.)} \cdot 100 = 1 + \beta_y(i) \cdot 100 \tag{7}$$

It may be noted that, since the mean of estimates of six rotation group totals is used instead of true values, $I_y(i)$ may be biased but is useful as a measure for evaluation of difference in relative biases between rotation groups for various sub-groups of the population and adjustment of weight based on household and person counts. Similarly, $P_y(i)$, the rotation group bias of population estimate can be defined for individual rotation groups. The values of the index $I_y(i)$ above 100 indicate positive relative bias and the values below 100 indicate negative relative bias. Similarly, the index $I_p(i)$ can be interpreted.

5. ANALYSIS OF DATA ON ROTATION GROUP BIAS INDEX

In the following tables data on rotation group bias index for population and labour force status categories by type of area and age-sex groups are presented and analyzed. The index values are obtained by using final weights and the same adjusted for non-response by rotation groups using each of the two factors based on household and person counts. A comparison of index values based on adjusted and unadjusted weights is used in evaluation of impact of weight adjustment on estimates of rotation group bias. The adjustment of weight by rotation groups, using household counts, was done at province level. Thus the final weights for households in the six rotation groups in each province were multiplied by adjustment factors $R_H(i)$; $i = 1,2,...6$. Similarly, the adjustment based on count of persons was done at province level by factors $R_P(i)$; $i = 1,2,...6$. In order to evaluate the impact of these adjustments on estimates of population we present Table 3 showing rotation group bias index for population estimates by type of area and months in the survey for twelve surveys in 1981. The index values based on unadjusted weight indicate that there is relative underestimation of persons in the first and the sixth month in both SR and NSR areas. The index values based on weight adjustment using household counts show some improvement in bias; however, this adjustment assumes that household size is the same in six rotation groups. The index values based on weight adjustment using counts of respondents are closer to 100.0 in both the areas, as compared to those based on household adjustment. Thus, the adjustment based on count of persons seems to correct the estimates for differential bias better than the adjustment based on household counts. The higher index values in earlier months and lower in later months could be due to changes in size of non-responding households by month in the survey.

Tables 4 and 5 present data on average index values by type of area and age-sex groups for twelve surveys in 1981. Index values by type of area based on unadjusted weight indicate that relative bias of estimates of unemployed tends to be positive in the first two months and shows a decreasing trend in the later months. Those for employed and in labour force tend to be negative in the first month and positive in the following months. Data on index values by age-sex groups show similar trends as those by type of area.

The adjustment of weight for non-response based on household counts tends to increase the index values in the first month and also fifth and sixth months. The index values in other months tend to decrease. This is true for index values for labour force status by type of area and age-sex groups. The increase in index values in the first month can be attributed to lower than average response rates and the decrease in index values in the following months to higher than average response rates. The decrease in the last two months can not be explained on the basis of higher than average response rates if $(\overline{Y}_{1i} + \overline{\beta}_i)$ is assumed constant.

The adjustment of weight for non-response based on count of persons tends to increase the index values in the first month and decrease the index values in the third to sixth month. The index values for the first month based on adjustment using count of persons tend to be greater than those based on household adjustment. The adjustment based on count of persons seems to correct the estimates for differential response between rotation groups. The response rates are low in the first month resulting in increase in relative bias after adjustment. The decrease in the relative bias in the third to sixth month seems to be due to lower than average response rates at household level, corrected for differential household size between rotation groups.

6. SUMMARY AND CONCLUDING REMARKS

This paper considers a model which decomposes overall bias into three components, showing the contribution due to differences in response rates, response biases and characteristics of respondents and non-respondents between groups of a sample. Rotation groups can be considered as a particular case of these groups in which adjustment of weight for non-response can be done separately. The model also shows contribution of various factors to rotation group bias.

If response rates differ between rotation groups, and the proportion of a characteristic for respondents and the associated response bias is equal for all rotation groups, non-response adjustment by rotation groups does not change the bias of estimates. However, rotation group bias can increase or decrease, according as response rate is lesser or greater than the mean response rate. This is corroborated by data on index values before and after adjustment of weight, based on count of persons.

It is proposed to analyze index values for labour force status and other characteristics for larger data sets and to study the impact of differences in average household sizes between rotation groups and respondent and non-respondent households on estimates of rotation group bias. The contribution of differential response rates and response biases to rotation group bias, after adjustment for non-response by rotation groups, will also be analyzed.

ACKNOWLEDGEMENTS

The author would like to thank R. Vettore and R. Barnes for the development of computer programs and the referees for helpful comments.

REFERENCES

[1] Bailar, B.A., (1975), The Effect of Rotation Group Bias on Estimates from Panel Surveys, JASA, Vol. 70, pp 23-30.

[2] Bailar, B.A., Bailey, L. and Corby, C., (1978), A Comparison of Some Adjustment and Weighting Procedures for Survey Data, Survey Sampling and Measurement, Academic Press, pp 175-198.

[3] Platek, R., Singh, M.P. and Tremblay, V., (1978), Adjustment for Non-Response in Surveys, Survey Methodology, Vol. 3, No. 1, pp 1-24.

[4] Statistics Canada (1976), Methodology of the Canadian Labour Force Survey, Catalogue 71-526, Occasional.

[5] Tessier, R. and Tremblay, V., (1976), Findings on Rotation Group Biases, Internal Technical Memorandum, Statistics Canada.

[6] Thomsen, I., (1973), A Note on the efficiency of Sub-class Means to Reduce the Effects of Non-response when Analyzing Survey Data, Statistical Review, Published by the National Central Bureau of Statistics, Stockholm, Sweden, Vol. 11, No. 4.

[7] Williams, W.H. and Mallows, C.L., (1970), Systematic Biases in Panel Surveys Due to Differential Non-Response, JASA, Vol. 65, No. 331.

[8] Woltman, H. and Bushery, J., (1975), A Panel Bias Study in the National Crime Survey, presented at Annual ASA Meeting.

TABLE 1. % Non-Response Rates for Households by Months in Survey and Type of Area (1981)

Months	Type of Area NSR†	SR	Canada†
1	6.6	7.9	7.3
2	4.0	4.6	4.4
3	3.5	4.4	3.9
4	3.5	4.1	3.8
5	3.2	3.8	3.6
6	3.1	3.6	3.4
Average No. of Households	26,707	28,645	55,352

† excluding special areas

TABLE 2. Estimated Proportions of Employed and Unemployed Heads in Respondent and Non-Respondent Households

Months	Respondents $\hat{\bar{Y}}_{1i}$		Non-Respondent $\hat{\bar{Y}}_{2i}$		$\hat{\bar{Y}}_{1i} - \hat{\bar{Y}}_{2i}$	
	Employed	Unemployed	Employed	Unemployed	Employed	Unemployed
1	0.6893	0.0383	0.7839	0.0335	-0.0946	0.0048
2	0.6962	0.0344	0.7841	0.0321	-0.0879	0.0023
3	0.7006	0.0311	0.7851	0.0300	-0.0845	0.0011
4	0.7006	0.0364	0.7877	0.0281	-0.0871	0.0083
5	0.6972	0.0317	0.7821	0.0317	-0.0849	0.0000
6	0.6927	0.0331	0.7767	0.0320	-0.0840	0.0011
Average	0.6961	0.0342	0.7833	0.0311	-0.0872	0.0031

TABLE 3. Rotation Group Bias Index for Population by Type of Area

Weight	Type of Area	Month in the Survey 1	2	3	4	5	6
Unadjusted	SR	97.0	101.1	101.2	100.6	100.2	99.7
	NSR	97.7	101.0	100.8	100.9	100.2	99.4
Household adjusted	SR	98.7	98.7	99.4	100.0	101.1	102.1
	NSR	99.3	98.6	99.0	100.3	101.1	101.8
Population adjusted	SR	100.4	100.5	100.2	99.7	99.6	99.5
	NSR	100.9	100.3	99.8	99.9	99.6	99.4

TABLE 4. Rotation Group Bias Index by Type of Area (1981)

Weight	Character-istics	Type of Area	Month in the Survey 1	2	3	4	5	6
Unadjusted	Employed	SR	99.9	101.0	101.3	100.7	100.4	99.8
		NSR	96.8	100.9	100.6	101.2	100.7	99.9
	Unemployed	SR	99.1	102.6	101.3	100.4	97.7	98.9
		NSR	103.3	101.5	101.4	99.8	96.5	97.6
	In LF	SR	97.0	101.1	101.3	100.7	100.2	99.7
		NSR	97.3	100.9	100.7	101.1	100.3	99.7
Household adjusted	Employed	SR	98.6	98.5	99.5	100.1	101.2	102.1
		NSR	98.3	98.4	98.7	100.6	101.6	102.4
	Unemployed	SR	100.8	100.3	99.5	99.8	98.5	101.1
		NSR	104.9	99.2	99.6	99.2	97.3	99.8
	In LF	SR	98.7	98.7	99.5	100.1	101.0	102.0
		NSR	98.9	98.4	98.8	100.5	101.2	102.1
Population adjusted	Employed	SR	100.2	100.3	100.4	99.7	99.7	99.6
		NSR	100.0	100.2	99.6	100.2	100.1	99.9
	Unemployed	SR	102.4	102.1	100.4	99.4	97.1	98.6
		NSR	106.4	100.8	100.5	98.9	96.0	97.4
	In LF	SR	100.4	100.5	100.4	99.7	99.5	99.5
		NSR	100.6	100.2	99.7	100.1	99.8	99.6

TABLE 5. Rotation Group Bias Index by Age-Sex Groups (1981)

Weight	Characteristics	Age-Sex Group	Month in the Survey 1	2	3	4	5	6
Unadjusted	Employed	M 15-24	96.5	99.7	100.7	101.0	101.1	101.1
		F 15-24	96.0	99.7	101.1	100.9	101.2	101.1
		M 25+	97.0	101.4	101.3	100.7	100.1	99.5
		F 25+	96.9	101.2	101.2	101.0	100.5	99.3
	Unemployed	M 15-24	100.9	102.3	101.1	100.7	96.9	98.2
		F 15-24	102.4	102.7	97.7	98.9	100.0	98.2
		M 25+	98.0	102.1	101.6	100.3	98.0	100.1
		F 25+	100.1	102.3	104.5	100.4	95.3	97.4
	In LF	M 15-24	97.2	100.1	100.8	100.9	100.4	100.6
		F 15-24	96.8	100.0	100.7	100.7	101.0	100.8
		M 25+	97.1	101.4	101.3	100.7	100.0	99.5
		F 25+	97.1	101.3	101.4	101.0	100.1	99.1
Household adjusted	Employed	M 15-24	98.2	97.2	98.8	100.3	101.9	103.5
		F 15-24	97.6	97.2	99.3	100.3	102.1	103.5
		M 25+	98.7	98.9	99.4	100.1	101.0	101.8
		F 25+	98.6	98.8	99.3	100.3	101.3	101.6
	Unemployed	M 15-24	102.6	100.0	99.3	100.1	97.7	100.3
		F 15-24	104.2	100.3	96.0	98.3	100.8	100.4
		M 25+	99.6	99.8	99.8	99.7	98.8	102.3
		F 25+	101.8	100.0	102.6	99.8	96.1	99.6
	In LF	M 15-24	98.9	97.6	98.9	100.3	101.3	103.0
		F 15-24	98.5	97.6	98.8	100.0	101.9	103.2
		M 25+	98.8	99.0	99.5	100.1	100.8	101.0
		F 25+	98.8	98.8	99.6	100.3	101.0	101.5
Population adjusted	Employed	M 15-24	99.9	99.0	99.7	100.0	100.5	101.0
		F 15-24	99.3	99.1	100.2	100.0	100.6	101.0
		M 25+	100.4	100.7	100.3	99.7	99.5	99.3
		F 25+	100.3	100.6	100.2	100.0	99.9	99.1
	Unemployed	M 15-24	104.2	101.7	100.1	99.7	96.3	98.0
		F 15-24	105.8	102.0	96.8	98.0	99.4	98.0
		M 25+	101.2	101.6	100.7	99.4	97.4	99.8
		F 25+	103.4	101.7	103.5	99.5	94.8	97.1
	In LF	M 15-24	100.5	99.4	99.8	99.9	99.8	100.5
		F 15-24	100.2	99.4	99.7	99.7	100.4	100.6
		M 25+	100.4	100.8	100.3	99.7	99.4	99.3
		F 25+	100.5	100.6	100.4	99.9	99.5	98.9

COMPUTERIZATION OF COMPLEX SURVEY ESTIMATES [1]

M.A. Hidiroglou [2]

Survey data collected by statistical agencies is most likely to be processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design and variables used. The statistics computed from such surveys typically range from simple descriptive totals and means to these required for analytic studies such as comparison of domains, regression analysis and contingency tables analysis. This paper describes a computer program which computes these statistics and their associated sampling errors for commonly used sampling designs.

1. INTRODUCTION

A variety of statistics are computed for survey data which often arise from large, complex national and regional surveys. The statistics computed from such surveys typically range from simple descriptive totals and means to those required for analytic studies such as comparison of domains, regression analysis, and contingency tables analysis. Domain estimation refers to the estimation of statistics for subgroups of the population of interest which are not explicitly provided for in the design. Yates (1960) contains considerable material on the estimation of domain means and their differences. Hartley (1959) and Rao (1975) provide an excellent account of the methodology used for domain estimation. The variance estimators associated with the domain estimators are easy extensions of variance estimators for simple statistics. This is not, however, the case for more complex statistics. The estimation of regression equations from survey data presents several problems; for example, the definition of the regression equations, the identification of the population for which inferences are desired, and the variance estimation for the regression coefficients (see Konijn (1962), Kish and Frankel (1974) and

---

[1] Presented at the Annual Meetings of the American Statistical Association, Detroit, August 1981.

[2] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada.

Fuller (1975). The testing of hypotheses for contingency tables given survey design considerations have been studied by Nathan (1969, 1972), Rao and Scott (1981), Garza-Hernandez and McCarthy (1962) and Koch, Freeman and Freeman (1975) to name a few.

Survey data collected by statistical agencies is most likely processed through to the tabulation stage by these agencies. The computer programs associated with this processing are also most likely tailored to the particular design used. It is quite possible that computer programs used to produce estimates of totals (say) and their associated variances must be developed from scratch every time that a new survey design is introduced. This is time consuming, expensive, tedious and in some sense repetitive. Use of statistical software packages such as SPSS or SAS may be considered as an alternative. These packages may be readily used to produce weighted'estimates. However, the variances that they compute do not take sample design factors such as stratification and clustering into account unless they are programmed to do so. A user must therefore be fairly familiar with the language used by these packages if he wants to obtain proper variance estimates for survey estimates.

Recently, there have been attempts to develop programs which compute variances for a general class of designs. Some of these programs are STDERR by Shah (1974), SURREGR by Holt (1975), SUPER CARP and MINI CARP by Hidiroglou, Fuller and Hickman (1980). These programs basically require the specification of the estimator to be used and the variables to be analysed. It will be assumed that the data sets that these programs are being applied to have been edited and that missing observations have been imputed. In this paper, SUPER CARP and MINI CARP will be described. SUPER CARP can be used to construct estimated totals, ratio estimates, the difference of ratio estimates and contingency tables tests for multistage stratified samples. It contains a number of regression procedures appropriate for data observed subject to response (Measurement) error. Covariance matrices can be estimated for sub-population means, and totals and for stratum means and totals. MINI CARP is a smaller program which differs from SUPER CARP in that it does not contain and of SUPER CARP's regression procedures. A comparison of the capabilities of the two programs is given in Table 1.

TABLE 1. Capabilities of SUPER CARP (S) and MINI CARP (M)

Multivariate Estimate of	For		
	Entire Population	Individual Strata	Sub-population
Simple Parameters			
. Means	S,M	S,M	S,M
. Totals	S,M	S,M	S,M
. Ratios	S,M	S,M	S,M
. Difference of Ratios	S,M	S,M	S,M
. Proportions	S,M	S,M	S,M

Complex Parameters		Tests	
. Weighted Least Squares	S	. Regression Coefficent	S
. Weighted Errors-in-the Variables (Known & Estimated error covariances)	S	. Goodness-of-fit	S,M
		. Independence for Two-Way Table	S,M

2. GENERAL DESCRIPTION

2.1 Notation

In general, SUPER CARP and MINI CARP can accept data from a multistage stratified design. Assuming that the design has s stages, a g dimensional data vector is read in for each observation. We denote this data vector as

$$(Z_{hi_s1}, Z_{hi_s2}, \ldots, Z_{hi_sg}),$$

where h = 1, 2, ..., L denotes strata; $i_s = (i_1, i_2, \ldots, i_s)$ represent the stages; $i_1 = 1, 2, \ldots, n_h$ represents the first stage identification; $i_2 = 1, 2, \ldots, n_{hi_1}$ represent the second stage identification; ... ; $i_s = 1, 2, \ldots, n_{hi_{s-1}}$ represents the last s-th identification. $Zhi_s k$ is the hi_s-th observation for the k-th variable of interest. Weights associated with the n_{hi_s}-th observation will be referred to as w_{hi_s}. These weights would be inversely proportional to the selection probabilities of each ultimate sampled

unit. The specification of the variables to be used in the analysis (be it total or ratio estimation or regression estimation) is done by using a selection vector $\underset{\sim}{v} = (v_1, v_2, \ldots, v_{p+1})$ where $1 \leq v_k \leq g$ for $k = 1, 2, \ldots, p + 1$. Given that the type of analysis and the identification of the variables has been decided upon, let the chosen vector for the $h\underset{\sim}{i}_s$-th observation be

$$(Y_{h\underset{\sim}{i}_s}, X_{h\underset{\sim}{i}_s 1}, X_{h\underset{\sim}{i}_s 2}, \ldots, X_{h\underset{\sim}{i}_s g}),$$

where Y denotes the dependent variable and X denotes the independent variables if regression analysis is specified. Note that v_1 is always the index for the dependent variable in the case of regression. For other types of analyses, the ordering within the selection vector is not important.

2.2 Types of Computations

The simple statistics and a partial list of the regression options available in the program are outlined. A complete descritpion of all the available options is written up in the SUPER CARP or MINI CARP manuals (1980).

(i) Total Estimator, e.g.

$$\hat{X}_{(k)} = \sum_h \sum_{i_1} \cdots \sum_{i_s} w_{h\underset{\sim}{i}_s} X_{h\underset{\sim}{i}_s(k)}, \; k = 1, 1, 2, \ldots, p.$$

The estimated covariance matrix for

$$\hat{X} = \{\hat{X}_{(1)}, \hat{X}_{(2)}, \ldots, \hat{X}_{(p)}\} \text{ is}$$

$$v_1(\hat{X}) = \sum_{h=1}^{L} (n_h - 1)^{-1} n_h(1-f_h) \sum_{i_1=1}^{n_h} (\hat{d}_{h\underset{\sim}{i}_1 .} - \hat{\bar{d}}_{h..})^T (\hat{d}_{h\underset{\sim}{i}_1 .} - \hat{\bar{d}}_{h..}) \quad (2.2.1)$$

where

$$\hat{\underset{\sim}{d}}_{hi_1.} = \{\hat{d}_{hi_1(1)}, \hat{d}_{hi_1(2)}, \ldots, \hat{d}_{hi_1(p)}\}$$

$$\hat{d}_{hi_1(k)} = \sum_{i_2=1}^{n_{hi_1}} \cdots \sum_{i_s=1}^{n_{hi_{s-1}}} w_{hi_s} X_{hi_s(k)}$$

$$\hat{\bar{\underset{\sim}{d}}}_{h..} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{\underset{\sim}{d}}_{hi_1.} \quad .$$

Note that the above variance formula may be applied to pps schemes with and without replacement. For with replacement schemes, only the first stage variance needs to be computed (Des Baj, 1968, pg. 120) and the correction factors f_h are set to zero. In large scale surveys, it is often assumed that the first stage clusters have been selected without replacement even though the actual selection scheme may have been without replacement. This assumption inconjunction with small sampling fractions implies that resulting variance is fairly close to the one which would have been obtained by taking all stages and selection prodcedure sinto account. If the sampling fractions are not negligible at each stage and that the sampling has been performed using without replacement S.R.S. at each stage, Des Raj's rule (1966) can be used to advantage to compute each stage component of covariance. The covariance matrix accounting for s stages is:

$$v(\hat{\underset{\sim}{X}}) = \sum_{r=1}^{s} v_r(\hat{\underset{\sim}{X}})$$

where for $r \geq 2$

$$v_r(\hat{\underset{\sim}{X}}) = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \cdots \sum_{i_{r-1}=1}^{n_{hi_{r-2}}} \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}} \right] \tag{2.2.2}$$

$$\times\ n_{hi_{r-1}} \ (n_{hi_{r-1}} - 1)^{-1} \ (1-f_{hi_{r-1}})$$

$$\times \sum_{i_r} (\hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.))^T$$

$$\times \ \hat{d}_{hi_r}(.) - \hat{\tilde{d}}_{hi_{r-1},.}(.)$$

where

$$f_{hi_{r-1}} = n_{hi_{r-1}} N^{-1}_{hi_{r-1}}, \ n_{hi_o} = n_h, \ N_{hi_o} = N_h,$$

$$\hat{d}_{hi_r}(.) = \hat{d}_{hi_r}(1), \hat{d}_{hi_r}(2), \ldots, \hat{d}_{hi_r}(p)$$

$$\hat{d}_{hi_r}(k) = \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi_s} X_{hi_s}(k)$$

$$\hat{\tilde{d}}_{hi_{r-1},.}(.) = n^{-1}_{hi_r} \sum_{i_r} \hat{d}_{hi_r}(.) \ .$$

The variance estimation for an r-stage design can therefore be done by estimating the components at each stage ($v_r(\hat{X})$) and summing them up. This can be done by passing over the data set r separate times. The first time around, strata and first stage units are read into the program to give $v_1(\hat{X})$. The second time around, the original primary sampling units are read into the program as "strata" and the secondary units are identified as clusters to give $v_2(\hat{X})$. The r-th time around, the original ($r \geqslant 2$) (r-1)-th stage units are read into the program as "strata" and the r-th stage units are identified as clusters to give $v_r(\hat{X})$.

On each pass a sampling rate $g_{hi_{r-1}}$ must be read in for the hi_{r-1} -th unit where

$$g_{hi_{r-1}} = 1 - \left[\prod_{j=0}^{r-2} \frac{n_{hi_j}}{N_{hi_j}}\right]\left[1 - \frac{n_{hi_{r-1}}}{N_{hi_{r-1}}}\right] .$$

Using this procedure, the program will be computing $v_r(\hat{X})$ in the format given by $v_1(\hat{X})$.

If the sampling factors are not negligible at each stage and that sampling has been performed using without replacement p.p.s. schemes at each stage, the variance expression at each stage must take into account joint selection probabilities. SUPER CARP and MINI CARP do not compute joint selection probabilities. For the case where two units per stratum have been selected without replacement and unequal probability, the variance of the estimator for total can be obtained using formula (2.2.1) with a correction factor for each stratum which includes the joint probabilities of selection. This correction factor is given by

$$f_h = \frac{2\pi_{h12} - \pi_{h1}\,\pi_{h2}}{\pi_{h12}} \quad , h=1, 2, \ldots, L$$

where π_{h12} is the joint probability of selection for the selected units 1 and 2. If $n_h \geqslant 2$ and that the joint probabilities of selection are not available, an approximation to the without replacement variance has been given by Gray

(1975). Gray shows that the variances of an unequal without replacement sample may be partitioned into a "with replacement" variance component times a finite population correction factor which depends on the joint probabilities. This correction factor has been found to be roughly equal to one minus the inverse of the sampling fraction for populations which have more than 15 elements within each stage. Using Gray's approximation, variances for multistage unequal without replacement schemes can be computed.

If domain estimation is required for some of the variables, a new variable ${}_dY_{hi_s}(k)$ is defined for all elements in the population, where

$$ {}_dY_{hi_s}(k) = \begin{cases} Y_{hi_s}(k) & \text{if the } hi_s\text{-th element belongs to the domain d (say } D_d) \\ 0 & \text{otherwise} \end{cases} $$

An alternative way of defining ${}_dY_{hi_s}(k)$, is

$$ {}_dY_{hi_s}(k) = {}_da_{hi_s}\, Y_{hi_s}(k) \text{ where} $$

$$ {}_da_{hi_s} = \begin{cases} 1 & \text{if the } hi_s\text{-th element belongs to } D_d \\ 0 & \text{otherwise} \end{cases} $$

Note that if $\hat{Y}$ and $v(\hat{\underset{\sim}{Y}})$ are unbiased for Y and $v(\hat{\underset{\sim}{Y}})$ respectively, then the corresponding domain estimators ${}_d\hat{\underset{\sim}{Y}}$ and $v({}_d\hat{\underset{\sim}{Y}})$ are unbiased for ${}_d\underset{\sim}{Y}$ and $V({}_d\hat{\underset{\sim}{Y}})$. The standard formulae for $\hat{Y}$ and $v(\hat{Y})$ can now be applied to the "synthetic" variables ${}_dY_{hi_s}$. Stratum totals can be computed individually by treating the strata as classification variables.

(ii) <u>Ratio Estimator</u>

The vector $\{Y_{hi_s}(1), X_{hi_s}(1), \ldots Y_{hi_s}(p), X_{hi_s}(p)\}$ is used in the analysis and the estimated ratios are:

$$ \hat{R}(t) = \hat{X}^{-1}_{(t)}\, \hat{Y}_{(t)},\ t = 1, 2, \ldots, p\ ; $$

where $\hat{Y}_{(t)}$ and $\hat{X}_{(t)}$ are of the form given in the previous section. The estimated covariance matrix for $\hat{\underset{\sim}{R}} = \{ \hat{R}(1), \hat{R}(2), \ldots, \hat{R}(p) \}$ is as given in the previous section with

$$\hat{d}_{hi_{\sim r}(t)} = \hat{X}^{-1}_{(t)} \sum_{i_{r+1}} \cdots \sum_{i_s} w_{hi_{\sim s}} \{Y_{hi_{\sim s}(1)} - \hat{R}(t)\, X_{hi_{\sim s}(t)}\};\ t = 1, \ldots, p.$$

The ratio estimator can be used for computing the mean for each variable of interest by setting all X-variables to 1. Domain means can be computed by using ${}_dY_{hi_{\sim s}(t)}$ in the place of $Y_{hi_{\sim s}(t)}$ and ${}_da_{hi_{\sim s}}$ in the place of $X_{hi_{\sim s}(t)}$. If subpopulation proportions of Y for a domain D_d are required, the numerator of the ratio is the sum of weighted ${}_dY_{hi_{\sim s}(t)}$ and the denominator is the sum of weighted $Y_{hi_{\sim s}(t)}$. The estimated ratio for two variables defined over a domain D_d may similarly be obtained. Stratum proportions and ratios may be computed with the strata serving as the classification variables.

(iii) Regression Estimation

Some considerable attention has been paid recently to regression concepts in survey sampling. There are several explanations for this. First, there is an increased emphasis on analytic surveys, with partly unresolved questions of proper weighting of observations. Secondly, modeling in general, especially in the regression context, has attracted widespread interest, as well as criticism, as a tool in making survey estimates. SUPER CARP properly weights the observations and computes the variances of the estimated regression coefficients using a method given by Fuller (1975).

The regression coefficients estimated from a stratified cluster sample are given by

$$\underset{\sim}{b} = (\underset{\sim}{X}'_n\, W\, \underset{\sim}{X}_n)^{-1}\ \underset{\sim}{X}'_n\, W\, \underset{\sim}{Y}_n$$

where the (rs)-th element of $(\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)$ is

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \quad X_{hi_1i_2s} \quad W_{hi_1i_2}$$

and the r-th element of $\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{Y}_n$ is

$$\sum_{h=1}^{L} \sum_{i_1=1}^{n_h} \sum_{i_2=1}^{n_{hi_1}} X_{hi_1i_2r} \quad X_{hi_1i_2} \quad W_{hi_1i_2} \; .$$

The estimated covariance matrix of $\underset{\sim}{b}$ is computed as

$$v(\underset{\sim}{b}) = (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} \hat{\underset{\sim}{G}}_n (\underset{\sim}{X}_n' \underset{\sim}{W} \underset{\sim}{X}_n)^{-1} \; ,$$

where the (rs)-th element of $\hat{\underset{\sim}{G}}_n$ is

$$\hat{g}_n(r,s) = \frac{n-1}{n-p} \sum_{h=1}^{L} \frac{n_h(1-f_h)}{(n_{h-1})} \sum_{i_1=1}^{n_h} (\hat{d}_{hi_1.r} - \bar{d}_{h..r}) \times (\hat{d}_{hi_1.s} - \bar{d}_{h..s})$$

where

$$\hat{d}_{hi_1i_2r} = X_{hi_1i_2r} \; \hat{v}_{hi_1i_2} \; w_{hi_1i_2} \; ,$$

$$\hat{v}_{hi_1i_2} = Y_{hi_1i_2} - \sum_{r=1}^{P} \hat{b}(r) \; X_{hi_1i_2} \; ,$$

$$\hat{d}_{hi_1.r} = \sum_{i_2=1}^{n_{hi_1}} \hat{d}_{hi_1 i_2 r} ,$$

$$\bar{d}_{h..r} = n_h^{-1} \sum_{i_1=1}^{n_h} \hat{d}_{hi_1.r} ,$$

$$n = \sum_{h=1}^{L} \sum_{i_1=1}^{n_h} n_{hi_1} .$$

The variance estimation procedure is based on an asymptotic Taylor expansion of the sample regression coefficient vector. This method has several advantages over the Balanced Repeated Replication and Jack-Knife Replication methods. Firstly, it is relatively easy to program, and it can be adopted to multistage sample designs. Secondly, no restrictions are placed on the sample design (two replicates per stratum, for instance) and the assumptions used require some well-behaved moments in the population of interest. Thirdly, it requires the least number of computations.

Data is quite frequently measured with error. Theory for regression models which takes measurement error into account has been given by Fuller (1980a), Fuller (1980b) and Fuller and Hidiroglou (1978). SUPER CARP also has the flexibility to compute tests of hypothesis for any subsets of the regression parameters.

(iv) Contingency Tables

SUPER CARP and MINI CARP perform the goodness-of-fit test and the independence test for data resulting from complex surveys. These two tests take the stratification and the clustering of the design into account. As pointed out by Rao and Scott (1981), pratictioners using traditional Pearson chi-quare statistics for those two tests, given that there may be serious design effects can be seriously misled.

For the goodness-of-fit test, SUPER CARP and MINI CARP use the modified Wald Statistic given by

$$F_{WG} = [(k-1)d]^{-1} (d-k+2) (\hat{\underset{\sim}{p}}-\underset{\sim}{p}_o)^T \hat{\underset{\sim}{V}}^{-1} (\hat{\underset{\sim}{p}}-\underset{\sim}{p}_o)$$

where

$\hat{\underset{\sim}{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k-1})^T$ is the vector of estimated proportions given in in the stratum and cluster configurations,

$\underset{\sim}{p}_o = (p_{o1}, p_{o2}, \ldots, p_{o,k-1})^T$ is the vector of hypothesized proportions,

$\hat{\underset{\sim}{V}}$ = the covariance matrix of $\hat{\underset{\sim}{p}}$ given the stratum and cluster configuration,

k = number of categories considered,

$$d = \sum_{h=1}^{L} (n_h-1),$$

L is the number of strata in the sample and n_i is the number of clusters in the i-th stratum. The covariance matrix $\hat{\underset{\sim}{V}}$ is computed using the methods given for ratio estimation. In large samples, F is approximately distributed as a central F with k-1 and d-k+2 degrees of freedom when the null hypothesis is true.

For the test of independence, Fuller (SUPER CARP p. 65-69) has developed a test which takes the design into account. Given that the contingency table which splits the population according to two criteria is made up of R rows and C columns, the null hypothesis to be tested is H_o: $p_{ij} = p_{i+} p_{+j}$ or $p_{+j} = p_{i+}^{-1} p_{ij}$. where p_{ij} = ij-th cell proportion in the population,

$p_{+j} = \Sigma_i p_{ij}$ and $p_{i+} = \Sigma_j p_{ij}$.

Given that $p_{ij|i}$ is defined as $p_{i+}^{-1}\ p_{ij}$ and that the corresponding sample estimators are $\hat{p}_{ij|i} = \hat{p}_{i+}^{-1}\ \hat{p}_{ij}$, estimates for $(p_{+1}, p_{+2}, \ldots, p_{+,c-1})$ can be obtained by regressing $\hat{p}_{ij|i}$ $(i = 1, 2, \ldots, R;\ j = 1, 2, \ldots, C-1)$ on (C-1)-dimensional row vectors whose elements are one for the j-th entry correspond ing to $\hat{p}_{ij|i}$ and zero otherwise. The regression is of a generalized least-squares nature because the $\hat{p}_{ij|i}$ do not have the same error structure. An estimator for the covariance matrix of the $p_{ij|k}$'s, incorporating the sample design, is obtained using the ratio estimator formulae. The test statistic for H_o is then based on the residual sums of squares for this regression.

3. INPUT

In a typical survey situation, the data associated with a given selected unit is characterized by stratum, first stage, second stage up to s-th stage identification and a sampling weight. The data must be ordered hierarchically with respect to this identification in order to produce estimates of variance which reflect the stratified and clustered of the data.

SUPER CARP and MINI CARP are run using command language specified in numeric codes in fixed card positions. For both programs, there are six mandatory control cards to be input at all times. A number of optional control cards may also be input if more information is required by options specified in the mandatory cards. The mandatory cards are the parameter card, the variable name card, the format card, the screening card, the analysis card and the variable identification card. The parameter card provides overall preliminary information to the program such as, problem identification, number of observations to be read in, input service identification (tape, disk or cards), data identification structure, data output and stratum collapsing controls. The format card specifies the input format for the data as well at its identification and the associated weight. the variable name card assigns chosen names to input data fields in the order that they are read in. The screening card specifies tolerance limits for given variables provided that screening is required. The analysis card specifies the type of analysis to be performed (see table 1). Finally, the variable identification card identifies the variables to be used in the chosen analyses. The optional cards include such

cards as the sampling rate card (sampling rates by stratum can be read in), the errors-in-the variable cards for supplying the program with covariance matrices for variables measured with error, the hypothesis testing card for specifying coefficients in a regression analysis to be tested equal to zero.

4. COMPUTATIONS

4.1 For Means and Corrected Sums of Squares and Cross-Products

The means, corrected sums of squares and cross-products are statistics routinely computed in a survey package. The choice of algorithms for computing these statistics should take into consideration precision, speed and storage requirements. Beaton, Rubin and Barone (1976) have noted that a "concern about highly accurate computation methods must be tempered with a concern for whether the data are accurate enough to make the results meaningful". Different variations of one-pass and two-pass algorithms have been studied by Ling (1974). Ling's conclusion is that there is no universally best algorithm. The best algorithm for a given data set depends on the numbers in that data set. One of his recommendations is to use double precision arithmetic to be beyond the accuracy attainable in single-precision arithmetic. One-pass recursive algorithms should be chosen over the usual one-pass 'desk-machine' method because they have a higher tendency to produce less computational errors. This is especially the case for subroutines programmed in single precision. In SUPER CARP and MINI CARP one-pass recursive algorithms programmed in double precision have been chosen.

4.2 Inversion of Matrices

Matrix inversion is required for regression and contingency table analysis. the choice for inversion algorithms is quite important in packages. This has been reported by Longley's (1967) paper in which he examined the accuracy of some inversion algorithms and found serious computational inaccuracies. He reported that the most accurate results were obtained by using the orthonormalization procedure. Kopitze, Boardman and Graybill (1975) recommend the use of the Cholesky decomposition as an inversting algorithm. They point out that as compared to the Gaussian elimination schemes, it does not require

pivoting to stabilize symmetric positive definite matrices. This means less time for inverting. The Cholesky decomposition does not need much core storage and is easier to program than the Gaussian elimination scheme. One of its other advantages, as Wilkinson's (1965) analysis shows, is that it is quite accurrate. Another of its advantages is that it can be used to find eigenvalues for systems of equations of the form $\underset{\sim}{A}\,\underset{\sim}{x} = \lambda\,\underset{\sim}{B}\,\underset{\sim}{x}$ where $\underset{\sim}{A}$ is a positive matrix and $\underset{\sim}{B}$ is a positive semi-definite matrix. Computations of eigenvalues are required in SUPER CARP for some of the errors-in-the variables regression analyses. It is for this reason and the precision considerations that the Cholesky decomposition has been adopted for inversion purposes in SUPER CARP.

4.3 Stratum Collapsing

If a sampled population is highly heterogeneous and several criteria are available for stratification, it is quite possible that some strata my contain only one cluster. For such strata, it is not possible to estimate the variability. In such cases, the user may request that the one cluster strata be collapsed with neighbouring strata. If such a request if not made, SUPER CARP or MINI CARP exclude with one cluster from variance computations but include them for estimation purposes. The program lists those strata with only one unit. This information may lead the user to collapse those strata in a subsequent pass. If collapsing is to be done, the strata which are to be collapsed should be similar to neighbouring strata. A suggested method for collapsing which is easily amenable to programming is as follows. If a stratum is encountered that contains only one cluster, that stratum is combined with the following stratum in the file sequence. If the last stratum contains only one element, the last stratum is combined with the next to last stratum. A stratum with a sampling rate of one is not collapsed because such a stratum makes no contribution to the between primary component of the sampling variance. Strata with a sampling rate of one should never appear after a stratum with only one cluster. One way to ensure this condition is to place all observations with a sampling rate of one at the beginning of the file sequence. If two strata are collapsed, the resulting sampling rate for the new stratum is computed as a function of the old sampling rates and and number of elements in the previous strata.

4.4 Clusters of Size One

If clusters of size one within a stratum at the first stage, collapsing of adjoining strata ensures that variance estimates will be computed. For a multi-stage design, some of the stages may contain single element clusters. For those clusters, no within cluster variation can be computed. There are several ways for handling this situation. One is to assume a zero-variance contribution from those single-element clusters. Another is to collapse them with neighbouring clusters. An alternative is to assume that they contribute a variance equal to the overall within variation of the clusters for which the within variation can be computed. The variance contribution for those stages where some of the clusters are of size one would incorporate this approximation.

5. SOME DESIRABLE FEATURES OF A VARIANCE ESTIMATION PROGRAM

Francis, Heiberger and Velleman (1975) listed criteria useful in evaluating programs in general. In this section, some of the desirable features of a computer program for estimating variance from complex surveys will be listed. These include user's documentation, input controls, printed output and statistical effectiveness. These desirable features will be related to those provided by SUPER CARP and MINI CARP.

User's documentation should consist of a manual which basically tells the user how to use the program. SUPER CARP and MINI CARP both have manuals which explain to the user how to use them. These manuals are structured as follows. They contain an introduction which summarizes the various available statistical options. Data input and command statements used to specify procedures, variables and options are explained and examples are provided to illustrate their use. Since data input and command statements are to be entered in a specific sequence, a flow diagram is provided. The program procedures are described in terms of the formulea used, the numberical techniques employed and some references to the literature.

As stated earlier, the command language used for SUPER CARP and MINI CARP is in the form of code number or alphanumeric codes in fixed card columns. As pointed out by Francis, Heiberger and Velleman (1975), the most computationally efficient command languages employ code number in fixed card columns. The disadvantage of this method is that users may make excessive references to the manual to identify the commands. Procedures and options could have been specified with the addition of a control statement translator which the addition of a control statement translator which would have allowed English like commands. The advantage of this input method is that it is relatively easy to learn. The disadvantage is that the time and effort required for programming this translator can be prohibitive.

The printed output in SUPER CARP and MINI CARP identifies the statistical procedure used and labels the variables used in the analysis. Part of the output refers to the program's version number, name and date it was last updated. This identification can be used to trace and fix bugs in the stated program version. Some informative diagnostic messages are also printed out. These include messages referring to input controls such as attempting to read in more variables than the program has been dimensioned to handle, trying to read too many cluster, improper input format. If some strata contain one cluster, the program will print out list of such strata. If the user requests collapsing of single cluster strata, the resulting strata will be printed out.

SUPER CARP and MINI CARP are written in FORTRAN and in double precision. They can be run on installations that have a FORTRAN compiler with minor modifications to the job control language. They can both be extented to accommodate new statistical procedures. These can be placed in the program in the form of new subroutines which can be connected to existing software in the program.

REFERENCES

[1] Beaton, A.E., Rubin, D.B., and Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy, Journal of the American Statistical Association, 71, 158-168.

[2] Raj, Des (1968), Sampling Theory. McGraw-Hill Inc.

[3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 393-397.

[4] Francis, I., Heiberger, R.M., and Velleman, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages, The American Statistician, 29, 52-56.

[5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.

[6] Fuller, W.A. and Hidiroglou, M.A. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.

[7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.

[8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples. Paper presented at the meetings of the American Educational Research Association, Boston, Massachussetts.

9] Garza-Hernandez, T. and McCarthy, P.J. (1962), A Test of Homogeneity for a Stratified Sample, Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.

[10] Gray, C.B. (1975), Components of Variance Model in Multistage Stratified Samples, Survey Methodology, 1, 27-43.

[11] Hartley, H.O. (1959), Analytic Studies of Survey Data, Instituto de Statistica, Rome, Volume in onore di Corrado Gini.

[12] Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.

[13] Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.

[14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, unpublished report, Research Triangle Institute, Research Triangle Park, North Carolina.

[15] Kish, L., and Frankel, M.R. (1974), Inference from Complex Samples, Journal of the Royal Statistical Society B, 36, 1-37.

[16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys, Journal of the American Statistical Association 57, 590-606.

[17] Kopitze, R., Boardman, T.J. and Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure, The American Statistician, 29, 64-66.

[18] Ling, R.G. (1974), Comparison of Several Algorithms for Computing Sample Means and Variances, Journal of the American Statistical Association, 69, 859-866.

[19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User, Journal of the American Statistical Association, 62, 819-841.

[20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples, Journal of the American Statistical Association, 67, 917-920.

[21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data, Survey Methodology Vol 1, supplement, Statistics Canada.

[22] Rao. J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.

[23] Rao, J.N.K. and Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.

[24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.

[25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem, Oxford: Clarendon Press, 229-233.

[26] Yates, F. (1960), Sample Methods for Censuses and Surveys, Charles Griffin and Sons, London, Third Edition.

SURVEY METHODOLOGY

December 1981 Vol. 7 No. 2

A Journal produced by Methodology Staff, Statistics Canada

CONTENTS

de l'information préliminaire d'ordre général, comme la définition du problème, le nombre d'observations à stocker, le support sur lequel se trouve les entrées (bande, disque ou cartes), la structure des données, de même que des renseignements sur la sortie des données et des contrôles pour la combinaison des strates. La carte du format indique la composition des données d'entrée ainsi que leur nature et le poids correspondant. La carte des noms des variables attribue des noms choisis aux zones des données d'entrée, suivant l'ordre dans lequel les données sont introduites. La carte de sélection contient les limites valables pour certaines variables lorsqu'une telle opération est nécessaire. La carte d'analyse énumère les analyses à effectuer (voir le tableau 1). Enfin, la carte d'identification des variables désigne les variables qui seront utilisées dans les analyses demandées. Parmi les cartes facultatives, on retrouve les cartes des fractions de sondage (on peut indiquer les pas de sondage de chaque strate), les cartes des erreurs sur les variables, qui donnent au programme la matrice des covariances pour les variables mesurées avec une erreur, et la carte des tests d'hypothèses qui permet de spécifier des coefficients de régression et de vérifier s'ils sont égaux à zéro.

4. CALCULS

4.1 Moyennes, sommes des carrés corrigées et produits vectoriels

Les moyennes, les sommes des carrés corrigées et les produits vectoriels sont des fonctions normalement traitées dans un programme d'enquête. Pour choisir les algorithmes nécessaires au calcul de ces fonctions, il faut prendre en considération le degré d'exactitude visé, la vitesse d'exécution et les contraintes liées au stockage des données. Beaton, Rubin et Barone (1976) ont admis que la recherche de méthodes de calcul très exactes doit être subordon-

née à la préoccupation de s'assurer que les données sont assez précises pour que les résultats soient significatifs. Divers modèles d'algorithmes à un et à deux passages ont été étudiés par Ling (1974), qui est arrivé à la conclusion qu'il n'y a pas d'algorithme supérieur aux autres dans tous les cas. Le meilleur algorithme pour un ensemble de données en particulier dépend des chiffres contenus dans cet ensemble. Une des suggestions formulées par Ling est d'effectuer des calculs en double précision pour obtenir des résultats plus précis que ceux des calculs en simple précision. Les algorithmes récurrents à un passage sont préférable aux méthodes habituelles à un passage du type "machine de bureau", parce qu'ils tendent à produire moins d'erreurs de calcul. Cela se note surtout dans les sous-programmes avec simple précision. Dans SUPER CARP et MINI CARP, on a choisi des algorithmes récurrents à un passage programmés avec double précision.

4.2 Inversion de matrices

L'inversion de matrices est nécessaire à la régression et à l'analyse de tableaux de contingence. Le choix de l'algorithme d'inversion dans une méthode informatique est très important, comme le démontre l'étude de Longley (1967), où l'auteur examine la précision de divers algorithmes d'inversion et découvre de sérieuses imperfections dans les calculs. Longley affirme avoir obtenu les résultats les plus précis en utilisant la technique d'orthonormalisation. Kopitze, Boardman et Graybill (1975) préconisent l'application de la décomposition de Cholesky comme algorithme d'inversion. Ces auteurs mentionnent que, contrairement à la méthode d'élimination de Gauss, la technique de Cholesky ne requiert pas de pivot pour stabiliser les matrices définies positives symétriques. L'inversion prend donc moins de temps. La décomposition ne demande pas beaucoup de place en mémoire et elle est plus facile à programmer que la méthode d'élimination de Gauss. Un autre avantage de la décomposition de Cholesky est qu'elle est assez précise, comme le démontre

l'analyse de Wilkinson (1965). De plus, cette technique permet de trouver les valeurs propres de systèmes d'équations ayant la forme A x = B x, où A est une matrice positive et B, une matrice semi-définie positive. Le calcul de valeurs propres est nécessaire dans SUPER CARP pour quelques-unes des analyses de régression avec erreurs sur les variables. Pour cette raison et compte tenu des critères de l'exactitude établis, on a adopté la décomposition de Cholesky comme méthode d'inversion dans SUPER CARP.

4.3 Combinaison de strates

Lorsqu'une population échantillonnée est très hétérogène et qu'on applique plusieurs critères de stratification, il est fort possible que certaines strates contiennent seulement une grappe. Dans ce cas, il est impossible d'estimer la variabilité, et l'utilisateur peut demander que les strates composées d'une seule grappe soient combinées avec des strates voisines. Si cette demande n'est pas faite, SUPER CARP et MINI CARP excluent ces strates des calculs de la variance, mais les incluent pour les besoins d'estimation. Le programme produit une liste des strates à une seule grappe, ce qui peut aider l'utilisateur à combiner ce genre de strates quand il présente un programme par la suite. Pour cette combinaison, les strates à une grappe doivent avoir les mêmes caractéristiques que des strates voisines. On peut suggérer la méthode suivante qui se prête bien à la programmation. Lorsque le programme découvre une strate qui ne contient qu'une grappe, cette strate est fondue avec la strate suivante classée dans le fichier. Si la dernière strate est composée d'une seule grappe, la dernière strate est combinée avec l'avant-dernière. Une strate dont la fraction d'échantillonnage est égale à 1 n'est pas combinée parce qu'une telle strate n'influe pas sur la variance observée entre les unités primaires de l'échantillon. Les strates dont la fraction d'échantillonnage est égale à 1 ne doivent jamais figurer après une strate

formée d'une seule grappe. Pour s'assurer que cette exigence est respectée, il suffit d'enregistrer tous les groupes d'observations ayant une fraction d'échantillonnage égale à 1 au début du fichier. Quand deux strates sont regroupées, la fraction d'échantillonnage de la nouvelle strate est calculée en fonction de la fraction d'échantillonnage de chacune des deux strates combinées et du nombre d'éléments qu'elles contiennent.

4.4 Grappes formées d'un seul élément

Si des grappes qui ne contiennent qu'un élément se trouvent dans une strate au premier degré d'échantillonnage, la combinaison de strates voisines permet de calculer des estimations de la variance. Dans un plan de sondage à plusieurs degrés, il peut arriver que certains degrés d'échantillonnage produisent des grappes formées d'un seul élément. Pour ce genre de grappes, il est impossible de calculer la variation à l'intérieur de l'ensemble des grappes pour lesquelles cette variation peut être calculée. L'incidence de ces degrés d'échantillonnage sur la variance, lorsque certaines grappes n'ont qu'un élément, peut alors être représentée par cette approximation.

5. QUELQUES CARACTÉRISTIQUES SOUHAITABLES DANS UN PROGRAMME D'ESTIMATION DE LA VARIANCE

Francis, Heiberger et Velleman (1975) ont dressé une liste de critères utiles d'évaluation des programmes statistiques en général. Dans la présente section, nous énumérons les caractéristiques qu'un programme informatique doit posséder pour l'estimation de la variance dans les enquêtes complexes. Parmi ces facteurs nécessaires, mentionnons la documentation des utilisateurs, les contrôles des données à l'entrée, les listes imprimées et l'efficacité statistique. Nous examinons ici dans quelle mesure les caractéristiques de SUPER CARP et MINI CARP répondent à ces besoins.

Essentiellement, la documentation des utilisateurs est un guide qui explique à l'utilisateur la façon de se servir du programme. SUPER CARP et MINI CARP comprennent tous les deux un guide de ce genre qui se présente sous la forme suivante. D'abord, une introduction résume les diverses options statistiques offertes par la méthode. Ensuite, les données d'entrée et les commandes relatives aux analyses, aux variables et aux options sont expliquées, avec exemples à l'appui. Comme les données doivent être introduites et les commandes placées selon un ordre particulier, un organigramme d'analyse est inclus. Les techniques offertes par les programmes sont décrites en fonction des formules et des méthodes numériques utilisées, et avec quelques références à divers ouvrages.

Comme il a été mentionné précédemment, le langage de commande de SUPER CARP et de MINI CARP est constitué de codes numériques ou alphanumériques enregistrés dans des colonnes fixes sur des cartes. Francis, Heiberger et Velleman (1975) signalent que les langages de commande qui peuvent exécuter les calculs les plus efficaces sont fondés sur l'utilisation de codes numériques dans des colonnes fixes. L'inconvénient de cette méthode est que les utilisateurs doivent se référer trop souvent au manuel pour trouver des commandes. Il serait peut-être possible de permettre aux utilisateurs de demander des analyses et des options par des commandes semblables avec des mots anglais, en ajoutant un programme de traduction des commandes. L'avantage de cette méthode est qu'elle serait assez facile à apprendre, par contre le temps et le travail nécessaires pour programmer le traducteur en rendraient le coût prohibitif.

Les listes imprimées par SUPER CARP et MINI CARP indiquent la technique statistique appliquée et les variables utilisées dans l'analyse. Une partie de chaque liste imprimée montre le numéro de la version de SUPER CARP ou de MINI CARP et la date de la dernière mise à jour. Ces renseignements peuvent

servir à repérer des erreurs dans la version indiquée et à les corriger. De plus, divers messages sont également imprimés. Par exemple, certains messages concernent les contrôles des données à l'entrée, comme lorsqu'on tente d'enregistrer plus de variables que le programme ne le permet, quand on tente d'introduire trop de grappes ou quand le format des données d'entrée ne répond pas aux normes établies. Si certaines strates contiennent une seule grappe, une liste de ces strates sera dressée. Dans le cas où l'utilisateur demande la combinaison de strates formées d'une grappe seulement, les strates ainsi produites sont imprimées.

SUPER CARP et MINI CARP sont écrits en FORTRAN avec double précision. Ils peuvent être utilisés sur les ordinateurs dotés d'un compilateur FORTRAN moyennant quelques petites modifications du langage de contrôle des travaux. On peut aussi ajouter de nouvelles techniques statistiques à SUPER CARP ou à MINI CARP. Ces techniques peuvent s'intégrer au programme sous la forme de sous-programmes qui peuvent être reliés aux autres éléments du logiciel.

6. BIBLIOGRAPHIE

[1] Beaton, A.E., Rubin, D.B., et Barone, J.L. (1976), The Acceptability of Regression Solutions: Another Look at Computational Accuracy. Journal of the American Statistical Association, 71, 158-168.

[2] Raj, Des (1968), Sampling Theory, McGraw-Hill Inc.

[3] Raj, D. (1966), Some Remarks on a Simple Procedure of Sampling Without Replacement, Journal of the American Statistical Association, 61, 391-397.

[4] Francis, I., Heiberger, R.M., et Velleman, P.F. (1975), Criteria and Considerations in the Evaluation of Statistical Program Packages. The American Statistician, 29, 52-56.

[5] Fuller, W.A. (1975), Regression Analysis for Sample Surveys, Sankhya, 37, 117-132.

[6] Fuller, W.A. et Hidiroglou, MA. (1968), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.

[7] Fuller, W.A. (1980a), Properties of Some Estimators for the Errors-in-Variables Model, Annals of Statistics, 8, 407-422.

[8] Fuller, W.A. (1980b), Estimation of Measurement Error Models from Cluster Samples, Communication présentée aux réunions de l'American Educational Research Association. Boston, Massachusetts.

[9] Garza-Hernandez, T. et McCarthy, P.J. (1962), A Test of Homegeneity for a Stratified Sample. Proceedings of the Social Statistics Section of the American Statistical Association, 200-202.

[10] Gray, G.B. (1975), Components of Variance Model in Multistage Stratified Samples, Techniques d'enquête, Statistique Canada, 1, 27-43.

[11] Hartley, H.O. (1959), Analytic Studies of Survey Data. Instituto de Statistica, Rome, Volume in onore di Corrado Gini.

[12] Hidiroglou, M.A., Fuller, W.A. et Hickman, R.D., (1980), SUPER CARP, Survey Section, Iowa State University, Ames, Iowa.

[13] Hidiroglou, M.A., Fuller, W.A., et Hickman, R.D. (1980), MINI CARP, Survey Section, Iowa State University, Ames, Iowa.

[14] Holt, Mary M. (1977), SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data, rapport non publié, Research Triangle Institute, Research Triangle Park, North Carolina.

[15] Kish, L., et Frankel, M.R. (1974), Inference from Complex Samples. Journal of the Royal Statistical Society B. 36, 1-37.

[16] Konijn, H.S. (1962), Regression Analysis in Sample Surveys. Journal of the American Statistical Association 5 7, 590-606.

[17] Kopitze, R., Boardman, T.J. et Graybill, F.A. (1975), Least Square Programs - A Look at the Square Root Procedure. The American Statistician, 29, 64-66.

[18] Ling, R.G. (1974), Comparaison of Several Algorithms for Computing Sample Means and Variances. Journal of the American Statistical Association, 69, 859-866.

[19] Longley, James (1967), An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User. Journal of the American Statistical Association, 62, 819-841.

[20] Nathan, G. (1972), On the Asymptotic Power for Tests for Independence in Contingency Tables from Stratified Samples. Journal of the American Statistical Association, 67, 917-920.

[21] Rao, J.N.K. (1975), Analytic Studies of Sample Survey Data. Techniques d'enquête, Statistique Canada.

[22] Rao, J.N.K. (1975), Unbiased Variance Estimation for Multistage Designs, Sankhya C, 37, 133-139.

[23] Rao, J.N.K. et Scott, A.J. (1981), The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness-of-Fit and Independence in Two-Way Tables. Journal of the American Statistical Association, 76, 221-230.

[24] Shah, B.V. (1974), STDERR: Standard Errors Program for Sample Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.

[25] Wilkinson, J.H. (1975), The Algebraic Eigenvalue Problem. Oxford: Clarendon Press, 229-233.

[26] Yates, F. (1960), Sample Methods for Censuses and Surveys. Charles Griffin and Sons, London, Third Edition.

TECHNIQUES D'ENQUÊTE

décembre 1981 Vol. 7 No. 2

Preparé par les méthodologistes de Statistique Canada.

TABLE DES MATIÈRES