Catalogue No. 95-19

COMBINING ADMINISTRATIVE AND SURVEY DATA TO REDUCE RESPONDENT BURDEN IN LONGITUDINAL SURVEYS

Product Registration Number 75F0002M

August 1995

Sylvie Michaud, Social Survey Methods Division

David Dolson, Social Survey Methods Division

Donna Adams, Household Surveys Division

Martin Renaud, Social Survey Methods Division

The SLID Research Paper Series is intended to document detailed studies and important decisions for the Survey of Labour and Income Dynamics. These research papers are available in English and French. To obtain a summary description of available documents or to obtain a copy of any, please contact Philip Giles, Manager, SLID Research Paper Series, by mail at 11-D8 Jean Talon Building, Statistics Canada, Ottawa, Ontario, CANADA K1A 0T6, by INTERNET (GILES@STATCAN.CA), by telephone (613) 951-2891, or by fax (613) 951-3253.

EXECUTIVE SUMMARY

This paper was presented at the 1995 Annual Meetings of the American Statistical Association in Orlando, Florida.

For many years increasing use has been made of administrative records for statistical purposes. Usually they are used because they provide almost the only source for important statistics. Often they are used to enrich, but not replace, data obtained through survey taking. This paper examines the situation where response burden is reduced by using administrative data to replace survey data for some but not all respondents.

A goal of Statistics Canada's Survey of Labour and Income Dynamics is to measure the impact of changes in labour market and family circumstances on the income of individuals. Each panel of this longitudinal survey lasts six years and has two interviews each year: one in January to collect labour information for the previous year and one in May to collect income data. Income questions are perceived as burdensome and are subject to data quality problems. To address these issues respondents are given the choice of answering the income questions or authorizing Statistics Canada to access their tax file information directly. Results from both collection methods are merged into a single output file. The paper presents the predicted impact of this mixed collection method on response and data quality, and discusses related measurement issues.

TABLE OF CONTENTS

1. INTRODUCTION

Demand for data and information from national statistical agencies has increased continuously for many, many years. Over the most recent several years, a common influence has been that of increasingly tight budgets. A sensitivity to response burden has long been important but in the last few years it is becoming even more important as anecdotal evidence suggests increasing respondent resistance to some types of inquiry.

Within the past year, three new longitudinal surveys have been put in place by Statistics Canada to satisfy new demands for information. They are the National Longitudinal Survey of Children, the National Population Health Survey and the Survey of Labour and Income Dynamics (SLID). For policy makers, program managers, analysts, and other users, the payoff will be substantial. However, the cost of these surveys in terms of both dollars and response burden is not small -- all three have moderately large sample sizes, lengthy questionnaires and respondents will be contacted a number of times over several years. Consequently, anything that can be done to reduce the cost or response burden is beneficial. This paper discusses the case of one of these surveys, the Survey of Labour and Income Dynamics, where it has been possible to achieve these goals by using administrative data in a mixed collection methodology.

2. ADMINISTRATIVE DATA

This section provides a very brief overview of administrative data, its uses, advantages and disadvantages. The material presented in this section is based to a large extent upon a paper by G. Brackstone (1987).

Brackstone identifies six broad categories of administrative records. They are records:

1) maintained to regulate the flow of goods and people across borders. Important examples are records of exports and of immigration.
2) resulting from legal requirements to register particular events like births, deaths, and business incorporations.
3) needed to administer benefits or obligations like taxation and unemployment insurance.
4) needed to administer public institutions like those related to schools, hospitals and courts.
5) arising from government regulation of industry. Records from banking and transportation are examples.
6) arising from the provision of utilities like electricity and telephone.

Records from all six categories are used at Statistics Canada; in the case of SLID, income tax records are used.

A first category of administrative record use is evaluation of survey or census data. For example, taxation and immigration records have often been used to evaluate census of population data on income and immigration.

Secondly, administrative records are used for creation and maintenance of survey frames. An example at Statistics Canada is the use of records from telephone companies for the maintenance of a frame for surveys conducted by random digit dialling. A second major example is the use of payroll deduction information submitted to Revenue Canada by employers. The business register is maintained, in part, by using this data source to identify opening of new businesses or changes to existing ones.

Some major statistical programs are conducted by using direct tabulation of data from administrative records. Statistics on external trade and vital events are produced by this means.

Possibly the most common use of administrative records is for indirect estimation where they comprise one of two or more inputs into the estimation process. Partial estimates of migration are produced by linking individual tax returns across years. Each year, Statistics Canada produces demographic and economic estimates for individuals and families using income tax records. The files used for this purpose over the years have been combined to produce a Longitudinal Administrative Data file which has facilitated longitudinal analysis of tax derived data on individuals and families. Some industry statistics are produced by combining survey data for large businesses with taxation data, adjusted if necessary, for small businesses. Recently, Canada's Survey of Employment, Payrolls and Hours has started using data from the payroll deduction data source to replace survey data for smaller businesses. The survey has realized substantial reductions in its cost and respondent burden while improving the quality of its statistics. SLID's use of income tax records falls into this category.

Administrative data sources can provide substantial advantages in terms of cost, respondent burden, and quality. As well they can facilitate estimates for small areas at low or moderate cost. However, several factors affect the suitability of an administrative data source for use in a statistical program. Three major ones are the following. How well do the definitions and concepts used in the administrative system correspond to those used in the statistical program? What is the intended coverage of the administrative system and how well does it match that needed by the statistical program? How good is the quality of the data available via the administrative source? Two other concerns are its frequency and timeliness. A last and important issue is its stability. The administrative systems are typically not

under the control of the statisticians and consequently changes may occur to which the statisticians must be able to adjust whenever they occur. These are all concerns which have had to be addressed for SLID's use of income tax records.

3. THE SURVEY OF LABOUR AND INCOME DYNAMICS

The Survey of Labour and Income Dynamics is a new panel survey that has been implemented in Canada in 1994. It is designed to measure changes over time in economic well-being, and to provide information on determinants of such changes, particularly with reference to demographic, family and labour market events. The survey focuses on medium-term dynamics; the intention is to follow individuals for six years. Respondents to SLID will be interviewed twice a year. A first interview done in January collects labour information for the previous full calendar year. A second contact in May collects income data, also for the previous year. April or May is the optimal time for income data collection because it increases the likelihood that respondents will have tax records available. If they consult these records, better quality income data are obtained.

Instead of collecting income information through an interview, a micro-match could be done linking the labour data collected by interview directly to the income tax data file. This approach has the potential of enhancing the data quality as well as to reduce the response burden (persons would be subjected to six interviews instead of twelve). To implement such an approach, respondent consent must be obtained first. (Canada's Statistics Act, under which Statistics Canada operates, requires that respondents give informed consent to such record linkage activities). Initial results strongly indicate the May income interview could not be completely dropped because too few respondents would give the necessary consent to access their tax records. Therefore, SLID has decided to adopt a "mixed" data collection

mode, where respondents are offered the choice between authorizing SLID to access their tax data or answering the income survey.

This mixed collection is not without drawbacks. This paper will focus on the issues that have to be addressed, studies that have been done to assess the impacts of such an approach and will conclude with future research plans.

4. WHY USE TAX DATA FOR SLID

4.1 TRADITIONAL WAY

At Statistics Canada, the traditional approach for collection of personal income data has been to ask respondents to recall their income received from all sources in the previous year. The annual Survey of Consumer Finances (SCF) is the main vehicle for collection of such data. Recently this survey has started use of Computer Assisted Interviewing. Previously it used a paper and pencil approach in which prior to the interviewer's call, respondents were sent a survey questionnaire and a guide explaining each item. For some items the guide referred to line numbers on the income tax form.

Income data collected by this means have suffered from a number of deficiencies. In particular, some income components are under-reported. Important examples are investment income, social assistance income, unemployment insurance benefits, and self-employment income. Poulin (1993) found that comparisons with other data sources suggest that these items are under-reported by ten to fifty percent at the aggregate level, depending upon the particular item. In the context of a longitudinal survey like SLID, significant measurement errors on individual records become especially important as much of the analysis will be done at the micro level.

4.2 COMPUTER ASSISTED INTERVIEWING

In 1993, SLID began evaluation of an approach to collecting income data that would be more closely linked to income tax records. Poulin (1993) discusses issues, advantages, disadvantages and conceptual issues related to the proposed approach. In May 1993, SLID tested a methodology in which reference to 1992 income tax forms by respondents was facilitated and encouraged -- by referring to specific tax line numbers, respondents were asked to copy numbers from their tax returns submitted to Revenue Canada onto a SLID "notebook".

The primary benefit being sought was a reduction in measurement error. It was also hoped simplifying the data collection process by this direct reference to the tax forms would result in higher response rates.

Some potential disadvantages were also identified. For some income components, the taxation system uses a different definition from that used by the SCF and most household surveys. Because not all income sources are taxable, it is necessary to collect some without reference to the tax forms. Similarly, it must be possible to complete the questionnaire for SLID with or without reference to tax forms since some respondents may not have theirs readily available or may not have even completed tax returns. Last, it is necessary to keep the SLID questionnaire to an acceptable level of complexity even though there are a number of different personal tax forms, most of them somewhat complex in nature.

The May 1993 test was conducted using a sample of 1500 households selected from two Canadian provinces. All had been respondents to the SCF in the previous year and had been interviewed in January 1993 for SLID labour data. The SCF was conducted, at that time, by paper and pencil (P&P) interviewing. The test had two main objectives. The first was to test SLID's proposed method

for collection of income data. The second was to determine the best way to collect income data when using computer assisted interviewing (CAI).

SLID developed three paths for collection of the income data as described in detail by Giles (1993). Prior to the test, each respondent was asked to complete a "notebook" sent to them containing all the questions. References to tax documents were included where possible. If the respondent did complete the notebook, the interview was shorter as it consisted only of reading amounts from the notebook.

For respondents who had not completed the notebook, the complete set of questions was split into groups of logically connected questions, or "blocks". For each block a global question was asked to determine if any of the specific income sources applied. Then, if appropriate, more specific questions were asked.

Respondents whose tax forms were available were prompted for responses using tax line numbers. As well, depending on the tax form used by the respondent, some blocks were skipped automatically.

If the tax form was not available, the "block" approach was used. This approach is similar to the tax approach except that all blocks were asked and there were no references to tax lines.

Apart from the three paths for collecting the income data, there were two other major differences between the CAI and P&P collection methods. With the former, interactive edits were possible. Secondly, dependent interviewing was introduced by deriving a set of flags from the January labour interview. These were used in the income interview to prompt for certain income items if they were not reported.

For example, a respondent who was a paid worker should report a non-zero amount for wages and salaries.

Data quality evaluation was done by linking to tax data. Where possible a direct link was made using the following variables: name, sex, marital status, age, postal code, date of birth, and spouse's name. An exact match was found for about 50% of records. Otherwise Statistics Canada's generalized record linkage system GRLS V1 (also known as CANLINK), was used to do a probabilistic match. An overall match rate of 84% was achieved. Of the non-matches, about half had reported an income of zero to the SCF survey (ie. for the previous year). It thus seemed likely that many of these were not on the tax file used for linking.

Results from this study were reported in a paper by Grondin and Michaud (1994). In general, there was more agreement between survey and tax data for CAI than for P&P. There was also less underreporting for the CAI approach. Between the three paths, the notebook and tax paths both clearly yielded better results than the block path. Use of the dependent interviewing was effective. Comparison to the tax data showed that the presence of a flag was a good predictor for the presence of an amount.

The decision was taken that in production, only one approach similar to the notebook approach would be programmed. The main reason for not also including the tax approach was to simplify the collection instrument as much as possible for interviewers. Interviewers would be trained to encourage respondents to use records whenever possible. Also, the production instrument would continue with the use of dependent interviewing via the flags.

The first wave of SLID data collection took place in January (labour interview) and May (income interview) of 1994. The CAI approach outlined above was used.

The response rates for the January labour interview and the May income interview were 86% and 76%, respectively.

4.3 TAX FEASIBILITY

In mid 1993 SLID began considering the feasibility of collecting income information by matching to Revenue Canada tax records for its survey respondents. As noted before, this has the potential to significantly reduce the respondent burden due to the income questions and to improve the quality of the income data.

As part of the testing, a subsample of the August 1993 Labour Force Survey sample were asked a "permission question". One part of the subsample had previously been included in the SCF while the other had not. The question was designed to determine if respondents would be willing to allow Statistics Canada to use their Revenue Canada income tax records instead of completing an income survey questionnaire.

The wording of the question was as follows: "We would like your opinion about a new way of getting some of the information that Statistics Canada collects. We are looking for ways to reduce cost, as well as your time and effort. Statistics Canada now gets income information by asking up to 25 questions on wages, pensions and other kinds of income. The income tax return has much of the same information. If you were in a Statistics Canada income survey, would you give us permission to get your information directly from Revenue Canada?". This question was asked of two persons in each selected household. Proxy responses were not accepted.

The analysis was conducted on unweighted data because the interest was in the behaviour of sampled persons and not of the population. The sample size was 29,582 persons of which 17% were non-respondents. Of respondents, 59% replied yes and 41% no. These percentages were very nearly the same for the set of persons who had been in the SCF and the set which had not. When examined by geographic or demographic characteristics somewhat greater variation was evident, but in no case was it particularly important. The only notable difference was that among persons who had been non-respondents to the SCF only 42% replied yes.

The response rate to the SCF's P&P interview was over 80%. The results of the test clearly indicated that a sharp drop in response rate would occur if a linkage collection methodology were used exclusively. However, the question was only hypothetical in nature and interviewers were not instructed to make any special efforts to convert a "no" response to a "yes". It is possible that the result would differ if sampled persons were faced with the real dilemma of answering 25 questions on income or granting access to their Revenue Canada tax records.

The 42% "yes" response among SCF non-respondents suggests that the survey's response rate could be improved by combining interviewing and linkage procedures.

In conjunction with the May 1994 collection of SLID income data, the feasibility of collecting income data via linkage to tax records was again evaluated, using a question similar to that used in the August 1993 test -- "As you might have noticed, the income tax return has much of the same information as we are asking you in this interview. With your permission, we could obtain this information from Revenue Canada. Next year, if we offered the choice, would ... give the permission to get his/her information directly from Revenue Canada". In this case,

respondents were asked the permission question immediately after having completed SLID's income related questions. On this occasion it was found that 56% replied yes, similar to the 59% from the first test.

The results of the feasibility studies are discussed in detail by Dibbs et al. (1994). A summary report by Dibbs, Poulin and Webber (1994) is also available.

4.4 MIXED MODE APPROACH

For the second wave, in January and May 1995, it was decided that a mixed mode approach should be implemented. To realize the expected quality benefits and response burden reduction, the tax file approach would be offered to respondents. For those giving permission, income data would be retrieved from tax files by a linkage procedure like that outlined in section 4.2. The Social Insurance Number (SIN) -- the account number used by individuals when filing their income tax returns -- was also retrieved to facilitate retrieval of the income tax data for subsequent waves. This ensures that longitudinal income data will be for the same person, even if the linkage established may sometimes be erroneous. SLID did not directly ask respondents for their SIN in order to maximize the number who would authorize access.

In addition, to help maintain a high response rate, CAI using the notebook approach would be retained as an alternative method of collection for respondents not giving permission to access their income tax records.

In January, at the time of the labour interview, there was no mention of the tax file option. Prior to the May income interview respondents were sent a questionnaire - to be used to facilitate the telephone interview - and an explanation of the tax file option. Those who would prefer the tax file approach were to tick a check box

while those preferring the interview approach were asked to complete the questionnaire and keep it near the telephone. When the interviewer called, she first offered the tax file option. If consent was not given, an interview was conducted. To reduce the impact of tax file under-coverage, all respondents who had not filed a tax return for the reference year were administered the interview. Special care was taken with the preparation of the various materials and the interviewer training so as to maximize the chance of increasing the consent rate above the rate observed in the permission test.

Preliminary results from the May 1995 collection cycle indicate that 63% of the respondents to the interview agreed to give SLID permission to use their tax data. For some of these it will not be possible to establish a link to a tax record -- these will be effectively non-respondents. Close to 4% did not file a tax return and 3% were non-respondents to the permission question. This leaves 30% who said no. Income data were collected via the survey from all of the last three groups -- 37 % of the respondents. (The non-filers were included in order to collect data on any small amounts of income they may have received).

In future waves, no May contact would be made with households in which all respondents gave consent to the tax file approach.

5. IMPLICATIONS AND ISSUES

From the respondents' perspective, a major reduction in response burden can be realized by agreeing to the tax file approach. The total number of interviews for respondents in the current panel would be reduced from 13 to 7. The interviews avoided are on a sensitive and difficult topic for which preparation is required. The simple factor of offering respondents a choice provides the opportunity of

increased response rate while keeping respondents - as well as interviewers - happier.

Based on the test results, data quality is expected to be superior. It is especially important that estimates of changes in income or frequency of changes not be artificially inflated as a result of response errors. Grondin and Michaud (1994) found that for a matched sample, analysis of data on income change coming from survey data (respondents to SCF in 1992 and to SLID's May 1993 test) and income tax showed grossly inflated estimates from the survey source for some variables. This was a key reason the dependent interviewing via the flags, first tested in 1993, was retained for the 1994 income interview.

It is expected that a major area of research using SLID income data will be family economic mobility particularly in terms of income stability and its correlates. Two others are low income dynamics and analysis of interactions between family income and individual labour market behaviour. For all of these, individual and family income data with minimal response error are priorities.

Use of the tax file approach is expected to reduce the extent of under-reporting of certain categories of income. With interview data, the spiking of income data at round figures can be problematic; this should also be reduced via the tax file approach.

The coverage of the tax file is an issue. In 1994, it covered about 94% of the Canadian population aged 20 and over. In offering the tax file approach, respondents were asked if they had completed a tax return the previous year. If not, then an interview is completed. This should reduce the impact of the tax file undercoverage.

The response rate for income at the first wave in May 1994 was 77%. This should be improved using the mixed mode approach.

In a longitudinal survey like SLID attrition of the sample is naturally a concern. Non-response or refusal to the income interview in one wave could lead a respondent to be more prone to do the same at the labour interview in the following January. So, the reduction in response burden via the tax file approach may result in reduced attrition in subsequent waves.

Another major issue is the merging of the data collected via the interview with that from the tax file approach. There are both conceptual and quality differences between the two sources. Mixing the two as in SLID's mixed mode approach is new. SLID has developed an approach to address this issue that is hoped will facilitate use of these data.

First the income data as collected (with some editing) for each respondent, whether tax file or survey data, is included. In addition "merged" income data variables will be created and included for every respondent. The merged data will be somewhat less detailed than either of the two collection sources. Whenever possible the standard for the merged income data will be the concepts used in the interview approach for collection. The merged data will incorporate some adjustments needed to improve comparability of the data arising from the two sources.

As well as difficulties related to concepts, section 2 also noted that stability is a concern with data from administrative systems. Taxation rules change with time. Consequently, the definition of income components can change. For example, components may become grouped or ungrouped. Items may be added or dropped from the tax forms. In the context of longitudinal data analysis, if such changes

become significant over time the analysis can become very complicated. In recent years, the trend has been to include more and more items on the tax form, including amounts which are not taxable.

Cost is an issue which has not previously been discussed in this paper. In 1995 the potential for cost savings is small mainly because all households still have to be contacted for the income interview. As well, some new costs are incurred for processing related to accessing the tax records. In 1995 about 63% of households gave consent to the tax linkage. Starting in 1996, only non-consenting households will be contacted for the May income interview and about $160,000 in collection costs can be avoided.

Confidentiality is also an issue. The mixed collection approach means that tax data will be substituted directly for the income survey data and merged with the rest of the survey data. Since tax data will be directly on the microdata file, there is an increased risk of identification of a person, even if SLID income variables are rounded. .

6. EVALUATION

A series of evaluation projects were undertaken to resolve some of the issues. The evaluation studies will be related to the different issues presented in the previous section.

6.1 POTENTIAL IMPACT ON RESPONSE RATE

The results from the May 1994 income interview were examined to assess the impact of the mixed collection on the response rate. There are two main influences to be considered, one positive and one negative. First, a percentage of non-

respondents to the income interview may be willing to give authorization to use their tax data from Revenue Canada; this would tend to increase the response rate. This would be particularly useful among the people who responded to the labour interview but refused the income interview. On the other hand, in some cases where respondents give permission to access their tax data it may not be possible to locate their tax record (these could be non-filers, late filers or people that could not be matched through the statistical linkage); this would have the impact of decreasing the response rate.

In 1994, 31,927 persons were eligible for the income interview. Out of those, 76% (24,261) responded. The full sample was matched to the 1993 tax file to determine the linkage rates. Again, statistical linkage was done, using the person's name, date of birth, sex, marital status, spouse's name, and postal code. A link to a tax record was found for 85.1 % (20,637) of the respondents and 76.2 % of non-respondents. (Linkage for non-respondents was feasible because demographic data had been collected for most of them in an earlier SLID interview). The May 1994 interview also included the hypothetical permission question. Among respondents, 56.7% (11,702) of those linked through tax replied yes, while 53.3% (1,932) of those not linked replied yes.

Using the mixed collection strategy, the negative effect is due to respondents who replied yes to the permission question but for whom no tax record was found. This results in a reduction in the number of respondents by 1,932 -- a decrease in the response rate of 6%. We were able to determine that 727 of these were not tax filers for that year. The reduction in response rate due to these people could have been avoided by first determining if respondents had filed a tax return or not and then routing all those who had not through the CAI collection of income data.

Tables 1 and 2 give distributions by income and by age group of the 1,932 respondents who granted access to their tax data but could not be matched to a tax record. Table 1 indicates that disproportionately many have low incomes; for example, 74.8% have incomes less than $10,000. Table 2 shows that the youngest age group is significantly overrepresented. The oldest age group is also overrepresented.

Table 1

Income distribution (as reported on the May 1994 SLID interview) of respondents who were not linked to a 1993 tax record but who had agreed to allow access.

Income range on survey	Respondents(%)	Population(%)
$0- $4999	55.4	11.8
$5000- $9999	19.4	13.4
$10000- $14999	8.5	15.2
$15000- $19999	4.5	11.4
$20000- $29999	4.6	17.9
$30000- $39999	3.6	12.5
$40000- $49999	1.6	7.8
$50000 +	2.5	10.0

Table 2

Age distribution of respondents who were not linked to a 1993 tax record but who had agreed to allow access.

Age group	Respondents(%)	Population(%)
16-19	28.9	6.9
20-24	6.6	9.0
25-34	10.1	21.6
35-44	7.5	21.1
45-54	12.4	15.6
55-64	13.0	11.2
65+	21.4	14.5

In Canada, the Social Insurance Number (SIN) is the account number used by individuals when filing their income tax returns. It follows that not having a SIN is a predictor of having a low income -- sufficiently low as to not be taxable. Of the 1,932 respondents who were not linked to a 1993 tax record, there were 727 for which SLID had the SIN and 1205 for which it did not.

Of the latter group, not having SINs, 62% are single, 54% are female, and 45% are less 20 years old. Overall 68% made less than $10,000, while of those less than 20 years old 98% made less than $10,000.

Those for which SLID did have the SIN are a very different group. Married women make up 79% of this group. They are older; 58% are over 54 years old. Even more of them have incomes of less than $10,000 -- 86%.

Clearly, a large proportion of the persons who might become non-respondents with the mixed collection methodology have low incomes, regardless of whether they had a SIN. Many of those for which SLID did not have the SIN are young and/or single people who may not yet have had significant experience in the labour force. On the other hand, those with a SIN tend to be older, and the majority are married women. For the 1995 cycle of SLID it is helpful to note that since these persons did respond to the survey in 1994, imputation is facilitated if they do become non-respondents in 1995.

Counteracting this is the positive influence on the response rate of persons who refuse the income interview but might give permission to access their income tax records. Based on the results from the August 1993 test, about 42% of the people who had refused to respond to the income survey said that they would be willing to let SLID access their tax information. Applying that rate to the number of non-

respondents that were matched to tax records, this effect could improve the SLID response rate by 7%.

Although these two effects appear to roughly balance off, the overall impact on the response rate (at least for complete interviews) is not clear. There does not seem to be a major problem with having a large number of people for whom no tax record can be found, even though they granted access to their tax records. The extent to which it is a problem should be alleviated in the 1995 survey where SLID has been more explicit about asking if people had filled in an income tax return or not. Income data for all persons replying no are collected via interview.

6.2 DATA QUALITY AT THE ITEM LEVEL

To evaluate the data quality, the SLID sample in 1994 was matched to the income tax file, by the same procedure as was done with the 1993 test. A series of comparisons were made with a focus on three main dimensions:

1) comparisons of respondents and non-respondents to evaluate the impact of having some of the non-respondents giving access to the tax data (stated differently: are there differences between respondents and non-respondents and will data quality be improved by adding some non-respondents back into the sample by using their tax data).

2) comparisons of tax data and survey data for "good" respondents to assess the differences in concepts between survey and tax data. Good respondents were defined as respondents that neither replied "don't know" nor refused any of the income questions, and that failed no consistency edits and for which a link to a tax record was found.

3) comparisons of tax data and survey data for the complete sample after processing, edit and imputation to assess the overall impact for users of having these data coming from a mixed mode collection.

This paper will report only on the first two of these.

The comparisons presented here are limited to six categories of income: Wages and Salaries (WS) , Farm and Non-Farm Self-Employment Income (FE, NFE), Interest and Dividends(I), Unemployment Insurance (UI), and Social Assistance (SA).

Wages and Salaries, Farm Self-Employment Income and Non-Farm Self-Employment Income were selected because there are differences in rules as to where self-employment income should be reported, based on the type of self-employment (eg. incorporated vs. non-incorporated).

Interest and Dividends was selected because this item is subject to undercoverage in surveys that collect income.

Unemployment Insurance was selected because it is also subject to some underestimation on the survey side. However, for SLID, receipt of UI benefits is asked twice, once in the labour interview where a general question asks if the respondent received any UI benefits and if yes when, while amounts of UI benefits received are asked in the income survey. It is hoped that asking the information on both occasions and using dependent interviewing will help the response to this item on the income interview.

Finally Social Assistance was studied because it poses a special problem. The question is asked at the person level. However, SA payments are calculated based

on the household composition and, although paid to only one person, are for the benefit of the entire household.

6.2.1 Comparisons of CAI respondents and non-respondents

Table 3 compares tax data of respondents and non-respondents that were successfully linked to their tax data for the six income sources noted above.

In section 6.1 it was shown that many of the persons for whom no link to tax data could be established had fairly low incomes and so may differ somewhat from those for whom a link could be established. Since their tax data was not available (or non-existent) these persons could not be included in Table 3. Consequently this only provides a partial comparison of respondents to non-respondents. Nonetheless, some general conclusions can be drawn.

There is usually little difference between respondents and non-respondents in the percentage reporting each income type. However, there do seem to be higher percentages of non-respondents who had received social assistance payments or income from self-employment (especially non-farm income).

Even though the reporting of different income sources in the tax data is similar for respondents and non-respondents, Table 4 shows that the means and medians of these amounts are usually higher for the latter.

These two tables indicate both that there is a small difference between respondents and non-respondents in the sources of income reported and that when an amount is reported the amount tends to be larger, sometimes by a great deal, among the non-respondents. Some of this bias due to non-response can be corrected if some of the non-respondents do give access to their tax information.

Table 3

Sources of income (per income tax) reported by respondents and non-respondents who were linked to their tax data.

	WS		FE		NFE	
	%	n	%	n	%	n
Resp	67.9	14021	3.6	737	8.4	1734
NR	67.8	3960	4.0	231	9.6	561

	I		UI		SA	
	%	n	%	n	%	n
Resp	39.5	8143	19.6	4053	7.1	1472
NR	39.6	2311	20.2	1177	9.1	533

Table 4

Mean and median income reported (per income tax) by income type and response status

		WS	FE	NFE	I	UI	SA
Mean	Resp	$24061	$4532	$10781	$2497	$5293	$5845
	NR	$24175	$6143	$13282	$3388	$5277	$6103
Median	Resp	$19724	$1249	$3690	$467	$4364	$5662
	NR	$18142	$2726	$4495	$634	$4344	$5382

6.2.2 Comparisons of survey data and tax data for "good" respondents

Further studies were done, comparing survey to tax data for "good" respondents. The purpose of these comparisons was to determine to what extent the definitions of the income sources were comparable between SLID and income tax. To do this study, some categories had to be redefined to create "comparable categories".

Some non-taxable income amounts, like social assistance, are reported on the income tax return. However, it must also be noted that there are some non-taxable income sources which are not reported on the income tax return. Some examples include veterans pensions, inheritances, and lottery gains. For these items, unless some specific questions are asked, the tax route would not provide this information. Of 15,862 "good" respondents in 1994, only 340 reported any non-taxable amount that would not be on the income tax return. Over all the "good" respondents the average amount reported was very low -- about 0.3% of total income. Over the 340 who did report an amount, the mean and median were 18.0% and 9.5% of total income, respectively. Only about 9% of the 340 had 50% or more of their income coming from this kind of source.

In general, there are some conceptual differences between the two sources of data. The purpose of the exercise was to see to what extent some of the conceptual differences were reported in practice. Examples of conceptual differences could be "under the table" income, or small amounts of interest income (income tax receipts are not issued for amounts less than $100), where in theory one could report it in a survey but might not report it on an income tax return.

From Table 5 it can be seen that there is a large difference in reporting of interest and dividend income, even among "good" respondents. There is also more reporting of self employment income on the tax file.

The means and medians shown in table 6 are higher from the survey data, especially for self-employment and interest income. This suggests that the amounts not reported on the survey are usually small ones.

Since all the "good" respondents had a link to a tax record, it was decided to do micro-comparisons of agreement rates of the data reported to the two sources. In

this context agreement refers only to whether an amount was reported or not on each source. Table 7 shows these comparisons. The first four lines show the frequency with which each income type was not reported in either source(=0 both), reported on both sources (>0 both), reported on income tax only (>0 tax), or reported on the survey only (>0 survey)**.** The last line shows among those reporting an amount on either source or both, the percentage reporting an amount on both.

Table 5

Sources of income reported to income tax and SLID (1994) by "good" respondents

	WS		FE		NFE	
	%	n	%	n	%	n
Tax	65.2	10338	2.8	440	7.3	1158
Survey	65.2	10336	2.3	360	5.1	805

	I		UI		SA	
	%	n	%	n	%	n
Tax	38.4	6084	16.5	2614	6.6	1043
Survey	29.7	4391	15.4	2449	6.7	1066

Table 6

Mean and median income by income type and data source for "good" respondents reporting non-zero amounts

		WS	FE	NFE	I	UI	SA
Mean	Tax	$25092	$6446	$12339	$2542	$5386	$6091
	Survey	$25060	$9763	$14635	$2912	$5197	$5882
Median	Tax	$21261	$2292	$3522	$487	$4505	$5905
	Survey	$21353	$5202	$5482	$597	$4247	$5710

Table 7

Micro-comparisons of survey data and tax data for "good" respondents

	WS(%)	FE(%)	NFE(%)	I(%)	UI(%)	SA(%)
=0 both	31.1	96.1	90.3	57.5	82.5	91.8
>0 both	64.1	2.5	4.3	27.9	15.5	5.7
>0 tax	2.4	1.0	3.8	12.6	1.5	1.2
>0 survey	2.4	0.4	1.6	2.0	0.5	1.3
% agree (>0)	92.9	63.8	44.2	65.6	88.8	69.5

A number of interesting conclusions can be drawn from Table 7. There may be some instances where there is some underground economy that is reported in the survey but not in tax, but in most categories, there are amounts reported in tax but not in the survey. There is a very low agreement rate between survey and tax data for self-employment income and more research is required to see if the amounts reported in the self-employment category are reported somewhere else in the survey questionnaire. Interest and dividends are also subject to underreporting in the survey area. Even if the agreement rate is better for UI, it is interesting to note that even if UI is asked on two occasions in the survey, and dependent interviewing is done to try to help the recall problems, assuming the tax was the "truth" there would still be an underestimation of the reporting of UI of 8.8 % in the survey. Finally another interesting point to note is that social assistance has a fairly low agreement rate, and there seems to a mismatch between the reporting on the survey or tax. This could be due to the fact that as mentioned earlier, Social Assistance is allocated to only one member of the household, when more than one person is eligible.

7. CONCLUSIONS

The approach of using a mixed collection of income data via income tax administrative records and via interview data appears very promising -- sufficiently so that this is the means by which income data for 1994 was collected in SLID's May 1995 interview. It is felt that in general, the combined approach should help SLID not just in terms of response burden but also in terms of data quality. However, more work is needed to fully assess the overall impact.

There is a certain fraction of persons who agree to give access to their tax records but for whom no link to a tax record can be established. Most seem to have lower incomes. In some cases, even with a Social Insurance Number no tax record can be found; this indicates persons who have not filed a tax return. Since SLID is a longitudinal survey, it will always be possible to add in their tax data if such persons ever do become tax filers.

Non-respondents do not differ substantially from respondents in terms of reporting amounts in different categories of income. However, the mean and median amounts reported by non-respondents are usually higher than those of respondents. Consequently then, the quality of SLID's income data will be improved if the response rate can be improved by converting some non-respondents to respondents by obtaining their permission to access their tax records.

At the macro level the comparability of tax and survey data is reasonably good. However, at the micro level the agreement rate is not especially high even when the comparisons are restricted to what should be "good respondents". In most cases of non-agreement an amount is reported in the tax data but not in the survey data. It could be that tax has included things that are not in the survey (or it is reported elsewhere in the survey), or that there is under-reporting in the survey.

There are variables such as social assistance that may have to be derived at the household level, or reprocessed at the person level (by averaging out amounts) to ensure consistency, especially from a longitudinal point of view.

It is notable that in the 1995 collection the permission question was asked in May; by agreeing respondents could shorten but not eliminate the interview. For the 1996 collection cycle the 37% of respondents who had not authorized access in 1995 will be asked the permission question again. This time it will be asked in the January interview, thus providing respondents the opportunity of avoiding the May interview entirely. Conversely, briefing material to be sent out in January will indicate that persons who had said yes are free to change their minds.

ACKNOWLEDGEMENTS

The contributions of Ruth Dibbs, Elaine Fournier and Maryanne Webber in defining the income sources for comparability purposes are gratefully acknowledged. Thanks also to Philip Giles who provided many helpful comments on a draft of this paper.

REFERENCES

Brackstone, G. (1987). Issues in the Use of Administrative Records for Statistical Purposes. Survey Methodology, 13, 29-43.

Dibbs, R., Fournier, E., Grondin, C., Leesti, T., Poulin, S., Saint-Pierre, Y., Webber, M. (1994). The Use of Tax File Data in the Survey of Labour and Income Dynamics: A Feasibility Study. Statistics Canada, SLID Research Paper 94-11.

Dibbs, R., Poulin, S., Webber M. (1994). The Use of Tax File Data in the Survey of Labour and Income Dynamics: Summary Report. Statistics Canada, SLID Research Paper 94-11.

Giles, P. (1993). SLID Income Interview - May 1993 Questionnaire and Data Collection Procedures. Statistics Canada, SLID Research Paper 93-04.

Grondin, C., Michaud, S. (1994). Data Quality of Income Data Using Computer Assisted Interviewing: SLID Experience. Statistics Canada, SLID Research Paper 94-15.

Poulin, S. (1993). The Use of Income Tax Data for SLID. Statistics Canada, SLID Research Paper 93-01.