Catalogue No. 94-11

THE USE OF TAX FILE DATA IN THE SURVEY OF LABOUR AND INCOME DYNAMICS: SUMMARY REPORT

September 1994

Ruth Dibbs, Household Surveys Division

Susan Poulin, Household Surveys Division

Maryanne Webber, Household Surveys 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

The SLID income interview is modelled after that of the Survey of Consumer Finances (SCF). Respondents are asked for income received from each of about 20 to 25 income sources during the previous calendar year. Much of this information is also reported by individuals when filing their annual income tax returns. In fact, both SLID and SCF collect their data in the Spring to take advantage of the respondents' heightened knowledge of their incomes at that time of the year.

A study was conducted to evaluate the feasibility of accessing income tax returns instead of collecting income information in a traditional survey. This report examines the issues. It recommends that SLID proceed with this approach for data collection in 1995, although future evaluations will be necessary to indicate whether the approach should be continued.

TABLE OF CONTENTS

Page

1. INTRODUCTION 1

2. OPERATIONAL SCENARIO 2

3. IMPLICATIONS FROM RESPONDENT'S PERSPECTIVE 4

4. DATA QUALITY IMPLICATIONS 5
   - 4.1 MAIN ANALYTICAL USES OF SLID INCOME DATA 5
   - 4.2 NON-RESPONSE AND RESPONSE ERROR IN SURVEY DATA 7
   - 4.3 COVERAGE 8
   - 4.4 RESPONSE ERRORS IN TAX FILE DATA 9
   - 4.5 ATTRITION 9
   - 4.6 MERGING DATA COLLECTED VIA INTERVIEW WITH TAX FILE DATA 10
   - 4.7 OVERALL ASSESSMENT OF IMPACT ON QUALITY 12

5. COST IMPLICATIONS 13

6. IMPLEMENTATION STRATEGY 14
   - 6.1 TIMING CONSIDERATIONS 14
   - 6.2 DEVELOPMENT AND TESTING OF RESPONDENT MATERIALS 15
   - 6.3 SIMULATION OF "MIXED MODE" RESULTS USING FIRST WAVE DATA 15
   - 6.4 APPROVAL PROCESS 16

7. CONCLUSIONS AND RECOMMENDATIONS 16

1. Introduction

The Survey of Labour and Income Dynamics (SLID) contacts respondents twice a year. The purpose of the second annual contact, which occurs in May, is to collect income data 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 is obtained.

Instead of collecting this income information through an interview, we could perform a micromatch, linking the labour data collected by interview directly to tax file data. This approach has the potential of both enhancing data quality and reducing respondent burden.

If we implement this approach, respondent consent would be sought. There are strong indications that we would not obtain consent from enough respondents to allow us to simply drop (or impute income for) those who do not consent. Therefore, this feasibility study is really evaluating a "mixed mode" approach. In other words, the proposal is to perform a link to the tax file for consenting respondents, and interview non-consenting respondents. The survey responses and tax file data would then be merged into a common dataset.

Our reason for examining the tax file option is to address explicitly concerns that have been expressed about income survey non-response and the precision of income information collected via interview. These problems are potentially more serious in a longitudinal survey than in a cross-sectional survey. There was not enough development time to evaluate the tax file approach prior to the first wave of SLID. We are therefore looking at the feasibility of this option with a view to its possible implementation for the second wave.

2. Operational scenario

How and when would we ask respondents to allow us access to their tax file data? After examining various options, we settled on the following scenario as optimal. In this scenario, we have kept open the option of asking wealth questions in May 1995, although it will be apparent that some trade-off may be required.

- In January 1995, the second wave labour interview will be conducted. At this contact, there is no mention of the tax file option. The interviewer simply indicates that she will be calling back in May.

- Before the May interview, respondents are sent a questionnaire (which is a tool to facilitate the telephone interview) and an explanation of the tax file option. On the questionnaire, there is a check box that respondents can tick if they consent to the tax file link. Those who would rather complete the interview are asked to fill in the questionnaire prior to the interviewer's call. We ask that the questionnaires be kept by the phone until the interviewer calls.

- When the interviewer calls, she offers the tax file option. If the respondent does not consent, the interviewer completes the income interview. Wealth questions could in principle be asked in both consenting and non-consenting households.

- In future waves, there is no contact in May with households where all members agree to the tax file link. If wealth information is to be collected -- the current proposal is to include wealth every three years, on the second and fifth wave of each panel -- it is part of the January interview.

- Joiners are asked at the time of their first labour interview if they consent to the tax file link. In future panels, longitudinal respondents would be asked at the time of the first labour interview if they consent. For example, respondents in the second panel would be asked in January 1997.

The inclusion of wealth questions in May 1995 would probably undermine our ability to obtain respondent consent. There is a danger of causing confusion and irritation among respondents who consent, because income and wealth both deal with money. One way around this problem would be to postpone wealth, perhaps until January 1998 (that is, skip the first wealth observation on the first panel). A separate evaluation of options with respect to wealth is currently underway, and the issue of wealth data in SLID will be reviewed more fully in that context.

Can we explain the tax file option to respondents in a way that ensures they understand it? We have sketched out some respondent and interviewer texts and believe that appropriate wording can be found (clear but not offputting) to explain the option to respondents. These materials would include:

- a brochure sent to respondents prior to the May 1995 interview, mailed out with the income questionnaire;

- a statement and "consent check box" on each individual income questionnaire;

- question wordings built into the CAI (computer-assisted interviewing) application;

- training materials and Qs & As (questions and answers) for interviewers and Advisory Services staff.

Testing to date indicates that 50% to 60% of respondents would consent to a link with the tax file. However, all the testing has been done using hypothetical questions and without prior explanations to respondents about the option. To maximize the consent rate, we need to develop these respondent materials very carefully, and test them. It is expected that help from outside consultants will be needed for the design of effective materials.

3. Implications from respondent's perspective

A major advantage of a link to the tax file is the reduction of respondent burden. The tax route reduces burden in three ways:

- The number of interviews for the first panel would be reduced from 13 to 8. For subsequent panels, it would fall to 7 (that is, the preliminary interview and six labour interviews).

- We would no longer need to ask respondents to *prepare* for an interview, by consulting their records and completing a form beforehand.

- We avoid an interview on a topic many respondents find sensitive and difficult. The evidence of sensitivity is the lower response rates associated with income surveys. There is further evidence in the National Census Test respondent debriefing results that respondents find income questions difficult to answer.

From a respondent's perspective, the choice element is a positive feature. Different respondents may have different concerns about providing information on income. If we can offer alternatives, we are likely to have happier respondents. It could

also be beneficial for the interviewing staff who would have a choice to offer to respondents.

4. Data quality implications

In addition to the respondent burden benefits, we would expect substantial benefits in terms of data quality if the tax link is implemented. Given SLID's emphasis on longitudinal data uses, it is important to minimize response errors that could artificially inflate the number of families experiencing changes in income. This concern about overstating change led to the use of dependent interviewing in the labour interview. However, dependent interviewing (in the sense of feeding back specific values collected a year ago) would not be appropriate for the income interview.

There are some disadvantages to using tax data. The pros and cons are discussed below. To provide some context for the discussion on data quality, the following section summarizes the main expected uses of the data.

4.1 Main analytical uses of SLID income data

We cannot predict precisely how the data will be used but, based on discussions with researchers and on uses that have been made of other longitudinal income data, the following are likely to be important areas of research.

There are likely to be studies of *family economic mobility*, similar to work done in the past using data from PSID (Panel Study of Income Dynamics) and SIPP (Survey of Income and Program Participation). These studies look at micro-level changes in family income, categorize families as stable or unstable with respect to

income and look for correlates of increases or decreases in income. The correlates of interest include labour market activity and events (eg, job loss, changes in the amount of work done by the family unit as a whole) and demographic characteristics and events (eg, family dissolution, creation of a blended family, geographic mobility).

A related area of research is *low income dynamics*. Some researchers will want to quantify flows into and out of low income. For example, how much turnover is there in the low-income population? Is there a substratum of "long-term poor"? This issue may be of particular interest from a social policy perspective, so attention may focus on isolating the distinguishing features of families that remain below the relevant low income cutoff for several years. There will be also be interest in assessing the impact of government transfers (particularly UI and SA) on preventing or ending low income spells.

Finally, we expect to see some research on the interactions between family income and individual labour market behaviour. For example, does family income have an effect on flows into and out of self-employment? What is the impact on transitions from work to retirement? Some researchers may also look at patterns of UI and SA receipt in relation to labour market behaviour.

From the foregoing, it can be seen that family income is a priority, as is information on change in income through time.

4.2 Non-response and response errors in survey data

SLID's first income interview in 1994 had a response rate of 83%. An additional 6% did not refuse outright but did not give any amounts to any income questions. If these are counted as non-response, the response rate for income drops to 77%. The use of tax file data could help us to boost response and thereby improve data quality.

In addition to survey non-response, income data collected by interview are susceptible to item non-response and under-reporting of amounts (relative to tax data). For example, SCF captures approximately 80% of UI benefits compared with 94% in the tax system. Investment income is also prone to under-reporting and bias.

A further problem in survey data is spiking in the distribution of income, attributable to respondents providing "ballpark" estimates. Studies of both Canadian and American data indicate that spiking can be quite severe. Among other things, this type of error can cause problems in the creation of income categories.

Most importantly, studies of low income dynamics and family economic mobility require sound micro-level data on changes in family income. Random reporting errors in any of the interviews with a household can result in over-estimation of changes in income. SLID has some empirical evidence of this, from a sample of about 1400 households contacted in two successive years. For this sample (interviewed for SCF and selected a year later for SLID's Test 3), a micromatch to tax file data for the same two years, 1991 and 1992, was performed. The following results pertain to individuals who had responded to both surveys and who were successfully matched to the tax file in both years:

- According to the survey data, 14% received income in one year but not the other. Based on tax data for these same respondents, this change in status occurred for only 3%.

- Large decreases in total income are more common in the survey data than in the tax data. The surveys showed that over 15% of respondents experienced a decrease in income in excess of 25% between 1991 and 1992. In the tax data, 10.5% had a drop of this magnitude. In absolute terms, survey data indicate that 21% had a drop of $2500 or more, compared with 15% in the tax file data.

- The survey data show that 16% were wage and salary earners in one year but not the other; in the tax data, this occurred in only 9% of the cases.

These few highlights support the concern that some have expressed about the impact of reporting errors on longitudinal data applications.[1]

4.3 Coverage

Is the coverage of tax file data adequate? In the past, undercoverage of certain population groups -- in particular low-income persons, older persons and non-working spouses -- would have been a major impediment to the feasibility of the proposal we are considering. Although coverage is still not complete, it is much improved, perhaps because of the introduction of tax credits. In 1994, the tax

---

[1] The findings should be interpreted with caution, because the sample is small and SLID (but not SCF) was done using computer-assisted interviewing. There were differences between SCF and the income questions used in the SLID test; the results deal with sources of income that were conceptually comparable in the two surveys. As the objective was to look at response, unedited data were used.

system covered approximately 94% of the population aged 20 and over, compared with 85% in 1984.

When the tax file option is offered to respondents, we would ask whether the respondent did in fact file a return for the previous year. If not, the interviewer would complete an income interview. This measure would substantially reduce the impact of tax file undercoverage.

4.4 Response errors in tax file data

The main issue here is income not declared to Revenue Canada Taxation in order to avoid paying taxes. Some researchers have advised us to not use tax data and not even ask survey respondents to consult their tax return, because it discourages the reporting of income earned in the "underground economy". Our position on this issue is that, whatever the collection methodology, we cannot adequately capture clandestine labour market activity. In general, people will not tell us about work (or income from other sources) that they are planning to conceal from Revenue Canada Taxation. This view is supported by the results of focus group studies done for SLID in 1992.

4.5 Attrition

Attrition is a major concern in panel studies. In SLID, the higher non-response rate for income may result in higher attrition -- in other words, having refused to complete an income interview, respondents may be more prone to refusing the labour interview in the following wave. The reduction of the number of contacts

and the removal of a sensitive topic could translate into lower attrition and therefore higher quality.

4.6 Merging data collected via interview with tax file data

There are definitely some drawbacks to the "mixed mode" of data collection. These can be summarized as follows:

- Some respondents may consent and subsequently change their mind so that, at the micro level, we have a mixture of data from two sources.

- The income categories used in an interview are not fully comparable to the tax return categories and this lack of conceptual consistency is a quality concern. The tax system does not capture certain forms of non-taxable income that may be especially important to low-income families.

- Researchers have become accustomed to the quality shortcomings in one type of data or the other but a dataset that involves a mixture of the two collection modes is a new and unknown quantity.

However, it is important to note that the data collected by interview are not themselves of uniform quality. Respondents are typically encouraged to consult records and prepare for the interview. Some actually do prepare carefully and the information they provide during an interview has the same level of precision as their tax return. At the other extreme, the interviewer may only obtain rough approximations, and that *by proxy*. In SLID's Test 3, of the respondents who completed the income interview, 37% completed a form beforehand, 17% consulted their tax return during the interview and 46% did the interview "cold". As might be expected, these three groups differ in terms of average income level

and number of income sources reported. Thus, quality will be far from uniform even if all data are collected by interview.

Another important point is that the impact of "mixed mode" depends on how well we can integrate the data coming from the two sources. How would we merge the two sets of concepts? The rest of this section summarizes the work on this topic done to date.

The proposed approach for presenting microdata can be summarized as follows:

- the income data as collected (with some editing) will be provided -- a respondent may thus have tax file data or survey data but not both

- In addition, "merged" income data will be presented for every respondent (that is, income data that can be comparably defined from either source)

- the merged data will be somewhat less detailed than the two feeder sources in terms of the number of income categories

- wherever possible, the standard for the merged data will be the survey -- in other words, we will try to adapt tax data to survey concepts, rather than vice versa[2]

- the merged data will reflect adjustments to improve comparability

Based on a preliminary examination, the "merged" data might consist of the following categories:

- wages and salaries

---

[2] However, we would still use the tax file for imputation, except for income sources which are not available from the tax file.

- farm self-employment
- non farm self-employment
- investment income
- government transfers
    - Child Tax Benefit
    - OAS/GIS/SPA
    - CPP/QPP
    - Unemployment Insurance
    - Social Assistance and Provincial Income Supplements
    - Workers' Compensation benefits
    - Goods and Services Tax Credits
    - Provincial Tax Credits
    - Other government income
- pension income
- alimony, separation allowance and child support
- other income
- total income
- income taxes paid

If we proceed, the merging will have be specified in detail and documented for the benefit of researchers. It is our intention to simulate the merging process using data from the first wave, which is being linked to tax file data for data quality evaluation purposes.

4.7 Overall assessment of impact on quality

The mixed mode approach adds complexity and introduces an "unknown" by mixing the strengths and weaknesses of survey and tax file data. Despite these

drawbacks, the net effect of a mixed mode approach would probably be quality improvement, through the reduction of attrition, the potential for greater cooperation among respondents, a more positive profile among the interviewing staff and greater precision in the data collected from the tax system.

5. Cost implications

The potential for cost saving is limited because of the need to use a "mixed mode" of data collection. Virtually all of the overhead costs of conducting an income interview would still be borne but there would be a reduction in interviewing time. This saving is offset by increased data processing costs (including mainframe) to match SLID records to tax files.

Assuming conservatively that 50% of households would consent to the link, we would save about $20,000 in collection costs in 1995. (The total field costs for the 1994 income interview -- that is, the first wave of the first panel -- were about $375,000.) The savings in 1995 are low because every household would have to be contacted.

In 1996, only non-consenting households would be contacted. The estimated savings are about $130,000. In 1997, assuming a second panel of equal size is introduced, savings would rise again and stabilize at about $235,000 per year.

The additional costs of the mixed mode approach are difficult to estimate. The micromatch to tax files would become more efficient over time. If our costing assumptions are correct, the bottom line is that cost reduction alone would not be enough to justify this approach. However, these savings have been calculated

using conservative assumptions and, if we are successful in obtaining consent from a substantial majority of households, savings will be correspondingly higher.

6. Implementation strategy

The following outlines some of the issues we need to address if we pursue the mixed mode option any further.

6.1 Timing considerations

There is an experimental element to the mixed mode option. Until we try it, we will not know how many households actually consent to the tax file link. We will not know what the characteristics of consenting and non-consenting households are, nor what impact mixed mode has on attrition. However, once we offer the tax route option to a particular respondent, it cannot be retracted. In other words, if we make the offer in 1995, but only 30% agree, we are basically "locked in" to the tax file route for that 30%, until the end of the first panel.

Implementation in 1995 would allow us to gauge respondent reaction and work through the processing implications prior to the introduction of the second panel. This would give us an empirical basis for deciding among the following options:

- retain mixed mode -- if it does indeed maximize response and minimize attrition;

- use a pure interview approach -- if the gains in data quality are not sufficient to warrant the extra complexity;

- consider more radical measures, such as using the tax file as an alternative or additional frame for the second panel -- if neither the pure interview nor the mixed mode approach yields results of sufficient quality.

6.2 Development and testing of respondent materials

If we proceed, the development of appropriate respondent materials is a critical issue. The information package sent to respondents must be attractive enough to entice them to read it, and clear enough to ensure that we are obtaining *informed* consent. The development of these materials would thus become a top priority this fall.

6.3 Simulation of "mixed mode" results using first wave data

Data from SLID's first wave are being linked to the 1993 tax file for purposes of data quality evaluation. The linked file can be used to simulate a "mixed mode" approach by:

- substituting tax data for some survey data;

- applying the necessary algorithms to estimate the missing components;

- deriving output variables from the "merged" tax and survey data.

This would provide practical experience in processing mixed mode data, and a basis for examining the impact on the data. However, the timing is too tight to

allow us to complete this work before deciding whether or not to implement (assuming 1995 implementation).

6.4 Approval process

Assuming that we decide to proceed, we will need to obtain approval from the Record Linkage Committee. Otherwise, we assume that there are no impediments.

7. Conclusion and recommendations

The main expected benefits of seeking respondent consent to a link to the tax file are:

- a reduction in respondent burden;
- improved respondent relations and interviewer morale in that a choice for income reporting is available;
- the possibility of lower non-response to income and lower attrition;
- greater precision in income reporting for that portion of the sample that consents to a link;
- some reduction in costs;

- an opportunity for accurately assessing impacts before the introduction of the second panel.

The main risks are:

- added complexity in the survey planning and processing functions, increasing the risk of error;

- added complexity in the dataset, leading to possible problems in data interpretation and loss of confidence among data users because "mixed mode" is not an established collection methodology;

- impossibility of reverting completely to current approach in future waves of the first panel, because offer to respondents cannot be retracted once it has been made;

- the probable loss of an opportunity to collect wealth twice on the first panel.

The bottom line is that there are so many unknowns with respect to the tax route and "mixed mode" that it is difficult to predict with any precision how it will turn out. To know definitively, we have to try it. In a longitudinal survey like SLID, where attrition and response inconsistency over time are concerns, the payoff in terms of quality improvement could be substantial.

However, there are clearly risks and, if we implement this approach, considerable effort will be needed to increase the likelihood of success. The respondent materials and the follow-through by the interviewers will require substantial front-end planning. It is equally important to merge the data from the two sources well,

and this is likely to be a time-consuming exercise. The approach also needs to be discussed with the data user community. We would need to share with them results from the dataset created to simulate a mixed mode approach. Good documentation on the methods used to do the merging would be equally important.

Despite the risks and the extra work, a major change in our data collection approach could become increasingly difficult to make over time, as the survey ages and our investment in a particular approach grows. There are advantages in addressing this important issue early in the life of the survey, particularly before the introduction of the second panel.