Catalogue No. 96-12

SOME DATA QUALITY IMPACTS WHEN MERGING SURVEY DATA ON INCOME WITH TAX DATA

Product Registration Number 75F0002M

November 1996

Sylvie Michaud, Social Survey Methods Division

Michel Latouche, 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 7-C6 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

Generally, measurement error causes some problems. A lot of work has been done in attempting to measure it and compensate for it. It has been shown that measurement error can create more problems in a longitudinal survey, especially if the data are used in regressions. The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey that attempts to measure the impact of changes in labour market activities and family characteristics on income. To try to reduce response burden and improve data quality, the survey has offered a choice to respondents: either respond to the income survey or give permission to SLID to use their administrative records. This paper aims to quantify the impacts of this mixed approach on the response error, especially on the measures of change.

This paper was presented at Statistics Canada's Symposium 96 *Nonsampling errors*, held in Ottawa in November 1996.

TABLE OF CONTENTS

1. INTRODUCTION

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey that measures the impact of changes in labour market activities and/or family circumstances on income. People in a given panel stay in the sample for six years and are interviewed twice a year. In January, labour information is obtained while the May interview collects income data. The income interview is done in May because Canadians file their income tax return by the end of April and it is generally felt that people are in a better position to provide accurate income data around that time. In May 1995, in an effort to reduce response burden, respondents in SLID were offered the choice between responding to the income survey and allowing SLID to get the income information from their tax return. This question is re-asked every year and after three years of collection, more than 75% of the respondents use the second option. However, the integration of survey data and tax data is not without problems. Definitions are not always compatible and there are linkage problems. This is balanced against quality issues usually found with income surveys (such as under-reporting of certain income sources) and the need for imputation. This paper gives an overview of the different sources of errors that occur with this methodology and presents some results on the impact of this mixed approach. The research has focused on micro-comparisons and has attempted to quantify the impact on measures of change.

2. SLID'S SAMPLE DESIGN

The SLID sample is selected using a multi-stage sample design. Respondents selected in the sample have already been part of the Canadian Labour Force Survey (LFS) for six months before being selected to participate in SLID. They are then interviewed twice a year for six years. A first interview in January asks detailed labour information. It also records changes in family composition and dates of the changes. The second interview in May collects detailed income

information, according to 24 categories. The survey also collects income tax paid to determine after-tax income. Income is collected for each individual in the household aged 16 years and older. It is aggregated at the family level to determine low-income measures.

Collection of income information is not a new process within Statistics Canada. The Survey of Consumer Finances (SCF) has been collecting annual income data for the last thirty years, and SLID's income questions are identical to those used by SCF. This experience will greatly aid SLID.

In general, income surveys suffer from lower response rates than many other surveys. While a non-income survey such as the Canadian Labour Force Survey has a usual response rate of 95%, the Survey of Consumer Finances has a response rate of 80% . The SLID response rate for the income interview is 76%. The data have also been linked to other sources for data quality evaluations. Based on these comparisons, there is under-reporting of certain income sources such as unemployment benefits, social assistance and interest and dividends[5].

Tax data has also been used more recently as a source of income information. In particular the Longitudinal Administrative Data (LAD) is a longitudinal file based on tax data[7]. A 10% sample of survey respondents has been randomly selected and families are reconstructed, based on the information provided on the tax return (spouses and children are created based on other fields from the tax form). LAD does not agree perfectly with the other administrative sources. Only census families (father-mother-children) can be constructed and there is a tendency to overestimate families of one person. There is also an under representation of certain age groups (particularly older persons) and small incomers. Recently however, with the implementation of tax credits, the population coverage of the universe by the tax system has improved. The problem of constructing families still

remains. On the other hand, the quality of income data from tax sources is felt to be superior to that from a survey.

SLID uses the mixed approach to try to maximize response rates and data quality. There are however issues with such an approach.

3. THEORETICAL SOURCES OF ERRORS USING SURVEY DATA VS USING TAX DATA

There are a number of issues with the principle of using administrative data for income sources. For example timeliness of the data may have an impact on the survey's target dates. The link to the administrative files with or without unique identifier may also raise some problems, and if one tries to assess the overall impact of a mixed strategy, this should be included. However, in this case the discussion will be restricted to the issues of combining both of the sources from a data quality point of view. A general discussion of the use of the tax file in SLID can be found in [1].

To provide a global measure of quality, surveys should compute mean square errors; that is the sum of the variance of a given variable and the square of the bias. This is usually hard to do because the bias can be very difficult to measure. Table 1 attempts to identify the potential advantages or drawbacks of each of collection methods (survey data and tax data) and tries to specify on which component of error it may have an impact.

Coverage is affected only by the use of tax data. Tax file coverage has improved over the years, covering 94% of the population aged 20 and over. If the population of non filers is different from the population of filers, this could create some bias in the data. Students, for example, are a group that is likely going to be under

represented amongst filers. Since they are usually associated with lower income, and that the filer students may be different from the non filer students, bias can be present.

Table 1. Comparison of the survey collection method vs using the tax data.

	Survey only	**Tax only**
coverage of population (bias)		↘ filers only
response rate (total) (variance) (bias)	↘ sensitivity response burden tracing	↗ all filers are "respondents"
		↘ not linked or wrongly linked
response error (bias)	↘ ↘ under-reporting of certain income sources (UI, interests...)	↘ under-reporting of certain income sources (underground economy)
	↘ rounding proxy reporting	
		↘ non taxable sources
time series consistency (bias)	↘ response error & longitudinal inconsistencies	
		↘ potential inconsistencies in definitions of income categories

↗ suggests an improvement ↘ suggests a disadvantage

Non-response can create problems both in terms of variance and bias. Income is a sensitive topic for some respondents, and it tends to have a "lower" response rate when it is collected by a survey. The fact that SLID is a longitudinal survey also impacts on the response rate; people move through the years and the inability to trace a person also decreases the response rate. The extent to which these non-respondents are different from the respondents will determine the magnitude of bias. The use of tax data should compensate for some of these problems in theory; as long as a person is a filer, it should be possible to locate their record on the tax file and this should increase the response rate. However, SLID does not collect the Social Insurance Number which is the unique link to the tax file. Other fields are used in a statistical matching procedure to link people in the SLID sample to the tax file of individuals. Some data quality control measures are done to improve the quality of the linkage but there is always the possibility of having a wrong linkage or that a person is not linked even if he/she is a filer. This also decreases the response rate.

Response error has been studied for income variables because of the availability of an external source to validate the results and assess potential biases. Some studies have suggested that there is an under reporting of certain income sources when data are collected from a survey. SCF captures approximately 80% of UI benefits compared to 94% in the tax system. Investment income is also prone to under-reporting. This creates bias in the results. In addition, there is a general feeling that tax data also suffer from some under-reporting of certain income sources that are related to the underground economy. However, because SLID asks respondents to consult their tax form to provide their income information, and because it is not clear that respondents will actually declare those kinds of income source through a survey, one could conclude that tax data may also be prone to bias, but not as severely as survey data.

A second source of response error is due to the rounding of income amounts reported in a survey. Rounding of a reported value for total income is problematic, but it is worse if rounding is done on the different sources of income since total income is derived as the sum of these components.

A third source of response error, affecting tax data, is the non availability of certain income sources on the tax form. Even though some non taxable income are reported on the tax file, it is limited to those sources that need to be reported for calculation of tax credits. Items such as lottery gains and inheritances are not reported, but are collected by the survey.

A fourth source of response error may arise because of changes in definitions and concepts in the tax environment. Time series may be affected because of changes in income tax regulations.

As can be seen, there are issues with both sources of data. The study wanted to see the impact of SLID's mixed collection on data quality. Because of the longitudinal nature of SLID, measures of change are important. Response errors create more problems in a longitudinal survey compared to a cross-sectional survey, since it is usually expected that correlation between the repeated measures will be larger than the correlation between the response errors. Because of the potential rounding and under reporting of error, it is expected that income from administrative sources will be less prone to response error than survey data.

In particular, assume a person wanted to measure a variable X (income), but what is really measured is $x = X + u$, where u is the response error. In a regression, where one would like to predict: $Y = X\beta + \epsilon$,
what is really measured is: $y = x\beta' + \epsilon$

where β' is biased towards zero, under the regular assumptions of the independence and the normality of the errors. If one was interested in doing a regression on the measures of change: $\Delta Y = Y_{t+1} - Y_t$, it has been shown [1] that the measure of bias in the equation of change is bigger than that on the measure of level. Mathematically,

$$\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} < \frac{\sigma_{\Delta x}^2}{\sigma_{\Delta x}^2+\sigma_{\Delta u}^2}$$

4. EMPIRICAL IDENTIFICATION OF THE SOURCES OF ERRORS

For its first year of data collection, SLID did not make use of the tax route and collected income data directly from the respondents. As a quality assurance activity, a comparison was performed between survey and tax data. This allowed an identification and quantification of error sources.

Tax data are obtained though matching but since the Social Insurance Number (SIN) is not asked of respondents, matching is done through a statistical procedure. Records were first linked though a direct match on name, postal code, data of birth, sex and marital status. This procedure linked 50% of the records. Records not matched were then run through a statistical matching process (allowing for missing values or discrepancy for one or more of the matching fields). This lead to an overall linkage of 85%. The study concentrated on response rate, coverage, linkage, and response errors. Special attention was devoted to the impact on yearly trends.

4.1 Response rates and potential biases

The SLID sample file was linked to the 1993 tax file using direct and statistical match approaches. Table 2 presents the distribution of the sample, by response status to the income interview and by the outcome of the linkage to tax data.

Table 2. Response status by linkage to tax data.

	Not linked to tax	Linked to tax	Total
Respondents	3,605	20,651	24,256 (76%)
Non-respondents	1,774	5,709	7,483 (24%)
Total	5,379 (17%)	26,360 (83%)	31,739

If everybody that was linked to the tax file had agreed to do so, there would actually be an increase in response rate. However, only 75% of respondents actually gave permission to use their administrative data. SLID also attempts to collect income for people who say they are non tax filers. Overall, there were two groups of people who could affect response rates: non-respondents to the income survey who gave permission to use their tax data and were linked would have a positive effect on the response rate while persons who gave permission to use their tax data but were not linked would have a negative effect. Approximately 1,700 persons were in each of the two groups. This means that in the end, the response rate remained the same with the mixed strategy.

However, because of potential biases due to a difference between linked and not linked respondents, these two groups were compared, using their income data from the first year survey collection. There were three subgroups with significant differences: single persons aged 15-19, single persons aged 20-24 and married women aged 45 years and more. Among these groups, there was a high percentage

of records not linked and usually the incomes for the people linked and not linked were different (the non linked persons having a lower income). These groups of not linked persons were then compared based on whether permission was given or not. Five large categories were used in those comparisons : employment earnings (wages and salary plus self-employment income), investment income (taxable investment income including interests and dividends), government transfers (Unemployment Insurance, Social Assistance, Child Tax Benefits, Old Age Security, Canada Pension Plan, Workers' Compensation and Goods and Services Tax (GST) credits), and total income. The comparisons were done on a subset of records labelled "good" respondents. This was done to exclude potential effects due to imputation. Table 3 shows the results. A similar pattern was found for all income categories. This suggests that if a proper pool of recipients was defined, a fairly valid imputation model could be done for the unlinked persons to tax, since there does not seem to be a difference between those who gave permission and those who did not.

Table 3: Comparisons of total income of "good" respondents (using survey data) for respondents who gave permission to use their administrative records vs the ones who did not.

	"good respondents" who gave permission			"good respondents" who did not give permission		
	n	mean	median	n	mean	median
single 15-19	865	$ 2,624	$ 1,500	727	$ 2,458	$ 1,000
single 20-24	606	$ 10,987	$ 8,800	470	$ 9,623	$ 7,188
women/married 45+	1599	$ 12,771	$ 7,677	1183	$ 13,573	$ 8,160
others	7857	$ 25,657	$ 20,000	5509	$ 26,667	$ 21,567

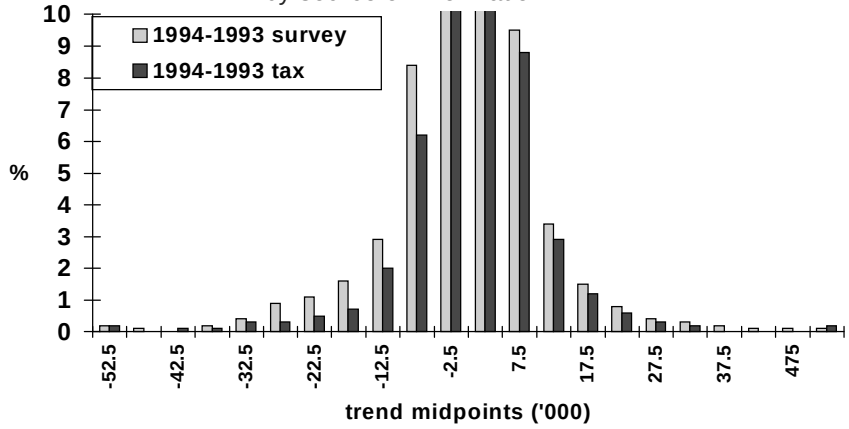
4.2 Response error

Income comparisons between survey and tax data have been done for a number of categories. Michaud et al. (1995) compared survey data with tax data for the "good" respondents, to see potential differences in income category definitions. These findings suggested that there were differences for self-employed income and social assistance. There were no differences in averages and medians for Wages and salaries and Unemployment Insurance benefits (UI). However there was still some under reporting of UI. The difficulty with this approach was to determine what was right. In particular, when there were differences in reporting of self-employment income, it was hard to see if this was representing income from the underground economy which would not be reported in tax, if it represented an income amount reported in some other income source or if it was wrongly reported. Since SLID is also interested mainly in longitudinal analysis, it was decided to study the differences in measures of change and to try to reconcile the micro-differences on records with two years of data. This also allows the study of response error with the longitudinal aspect in mind.

Only a subset of SLID respondents have two years of data from both the survey and tax. This limited the study to a sub-sample of 4274 respondents. This subsample is not quite representative of the whole sample; it has a slightly higher percentage of people in the 65 years of age and older group and smaller representation in the very young group 16-19. The differences were however not found to be important enough to invalidate the study. Of this subset, 86% of the records had been obtained through the direct match and so only that subset was kept, again to remove potential effects of incorrectly linked records. The comparisons were restricted to that subset of 3670 records. However, another 600 records were removed; 400 of these records had partial non-response in the second year, and most of the remaining 200 had no income in one or both years.

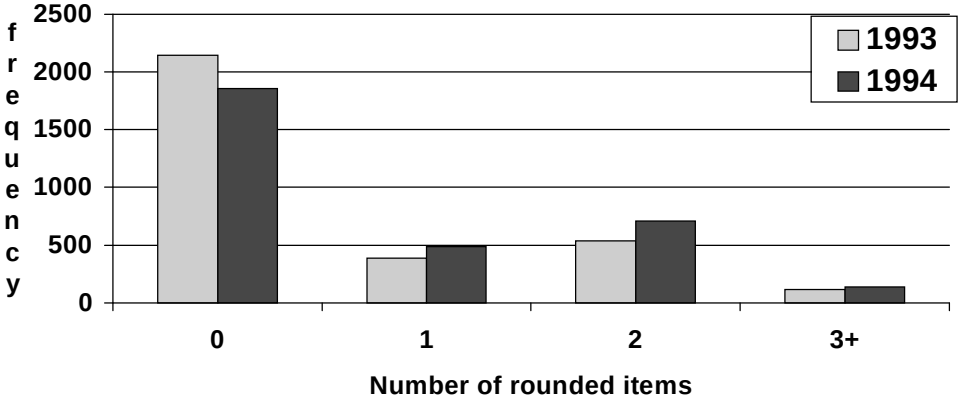
Attention was paid to the measures of change in total income between the two years from both survey and tax data. For each person, a change of total income -- or micro-trend -- was calculated for the survey data and the tax data. Histogram 1 shows the distribution of the changes. The vertical axis scale has been cut to a maximum of 10% to allow a better view of the tails of the distribution (the scale should have gone up to around 30%) . The average change from the tax data was an increase of $498 while survey data suggested a decrease of $3 (the difference was significant at the 1% level). There also seemed to be more variability in the measures of change from the survey.

HISTOGRAM 1.
Distribution of yearly micro trends for total income
by source of information



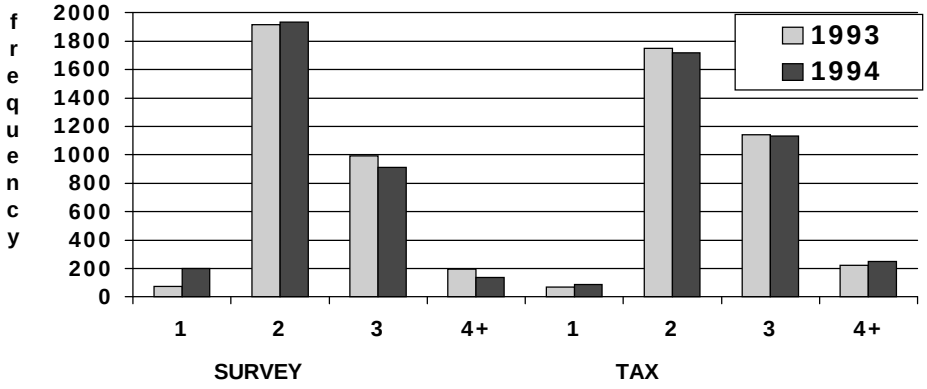
The studies then compared the pattern of reporting. An initial look at the data suggested that there were two different behaviours depending on whether a person was giving approximate amounts (that was detected by looking at the income sources that were rounded) or exact amounts. So the study focused on the amount of rounding done. The study was restricted to the rounding of ten income categories only because other income categories were not reported in a similar way on both the survey data and the tax data. Rounding was defined to be when the two last digits were zero on survey but not on tax. Histogram 2 shows the distribution of the respondents according to the number of items rounded in their survey data in 1993 and 1994 reference years. Rounding of income amounts happens frequently; only 1530 records, that is 47.5 % of the people do not round any of their income amounts in the two years. It also seems that the amount of rounding increases in the second year of collection.

HISTOGRAM 2.
Distribution of the number of rounded items per individual
based on 10 items
(by year)



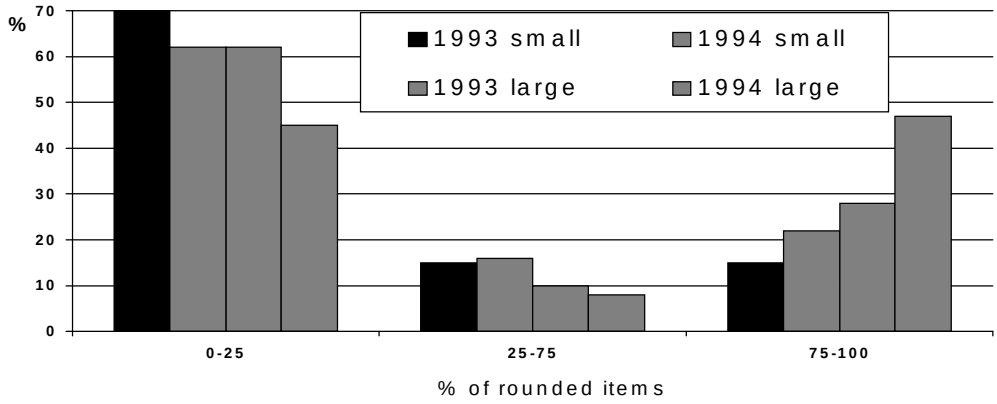
Histogram 3 shows the distribution of the respondents according to the number of non-zero items reported to the survey. It is interesting to note that the average number of items reported to the survey slightly decreased in the second year of survey compared to the first year but this is not observed in the reporting on the tax for those same items.

HISTOGRAM 3.
Distribution of non zero items per individual
(by year and source)



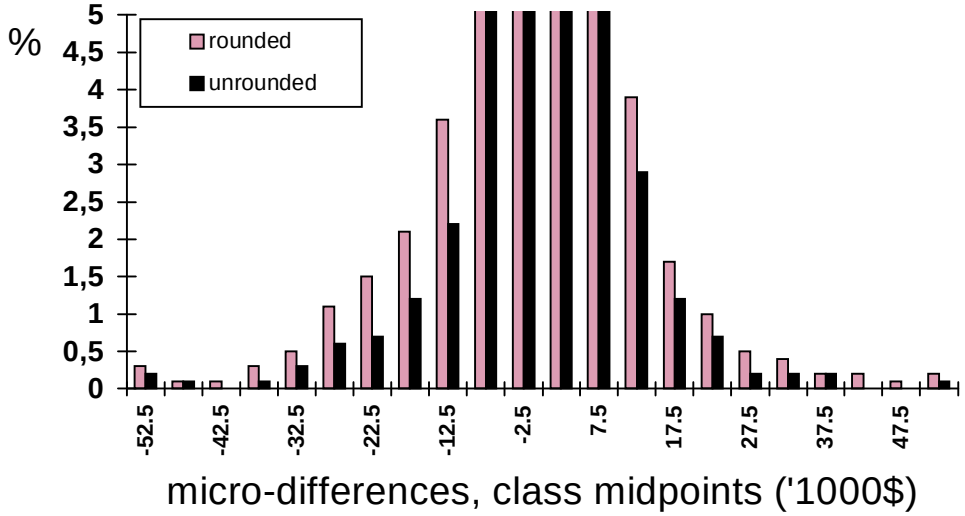
It was of interest to know if rounding was different for different groups of respondents. Rounding behaviour over time was compared for various income groups. As indicated in Histogram 4, income seems to have an impact on rounding; those with large income tend to round more than those with small income. There also seemed to be less rounding for older people.

HISTOGRAM 4.
Distribution of individuals by proportion of rounded items
for small and large total income respondents
1993 and 1994 reference years



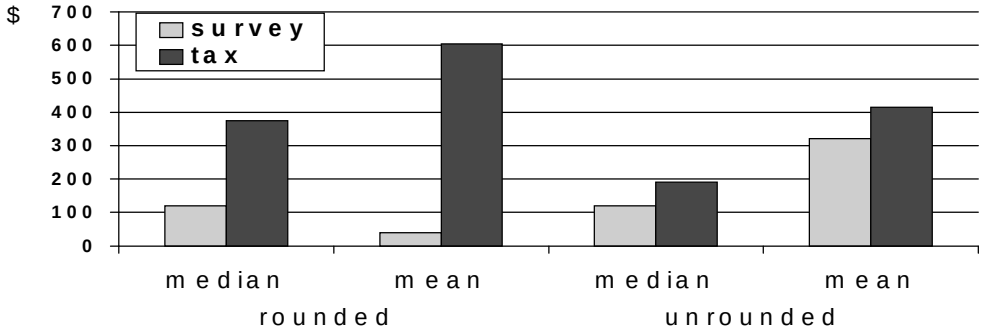
Once again total income was compared. The measures of change -- or micro-trend -- were compared using survey data, dividing respondents in two groups; respondents who rounded their reported income vs the ones who did not. This is shown in Histogram 5. A respondent was assigned to the rounded group if at least one of the items was rounded. It is interesting to note that much of the variability of the measures of change observed in histogram 1 is observed in the group of respondents who rounded at least one source.

HISTOGRAM 5.
Distribution of micro-differences
(yearly trends between 1994 and 1993)
for "rounded" and "unrounded" groups



To confirm this hypothesis, the measures of change were compared between survey data and tax data for the group of respondents who did not round the income amounts that were reported. Histogram 6 shows mean and median micro-trend separately for the “rounded” group and the “unrounded” group. For the group of people who did not round, the differences were not that large. As can be expected, the biggest discrepancies happen in the group of people who round their income.

HISTOGRAM 6.
micro-trends central tendencies
for rounded and unrounded groups



5. CONCLUSIONS

This was just a first look at the issue of response error. When some of the largest differences were examined by subject matter specialists, the differences were attributed to a response error in the survey in approximately 80% of the cases. Approximately 10% of the cases were attributed to an "error" in the tax data (a non taxable item was missing in one of the two years or there seemed to be an error in the tax field). Finally, the remaining 10% was not explainable.

There were some other interesting findings; approximately 30% of people provide exactly the same amounts (to the dollar) on both the survey and tax files, for at least one year of data. The rest have response error either from the survey or from tax files, or possibly both. Both sources of data have their limitations; the tax information has the problem of non tax filers and the underreporting of non taxable amounts while survey data seems to be prone to response error. The response error on the survey data also seems to increase with time. Based on the observed

results, even if tax data is prone to error, the use of tax information in this mixed approach will probably improve the quality of the income data, especially because of the longitudinal nature of the survey. There are still things to investigate; the overall findings do not seem to hold quite as nicely for the self-employed. This group should be analysed further. In a similar fashion, the study should be refined to study reporting by income source. The impact of non taxable income amounts on the measure of change should be evaluated in more detail. Finally, a number of techniques have been suggested to correct for response error [3], [6]. These techniques should be applied and tested to see if they can be incorporated to improve the quality of the measures of income.

The authors would like to acknowledge the contribution of Chantal Grondin, Martin Renaud, Carole Janelle and Elaine Fournier in preparing the study.

REFERENCES

[1] Bound, J., Brown, C. Duncan, G. And Willard, R. (1991), "Measurement Error in Cross-Sectional and Longitudinal Labor Market Surveys: Validation Study Evidence", Panel Data and Labor Market Studies, J. Hartog, G. Ridder and J. Theeuwes (eds.), Elsevier Science Publishers

[2] Dibbs, R., Poulin, S., Webber, M. (1994), "The use of tax file data in the Survey of Labour and Income Dynamics: Summary report", SLID Research Paper Series, Cat. No. 94-11.

[3] Fuller, W. (1987), Measurement error models, Wiley

[4] Groves, R. (1989), Survey Errors and Survey Costs, Wiley

[5] Michaud, S., Dolson, D., Renaud, M., (1995), "Combining survey data and administrative data", SLID Research Paper Series, Cat. No.95-19.

[6] Plewis, I. (1985), Analysing change, Measurement and Explanation using longitudinal data, Wiley.

[7] Statistics Canada (1995), An overview of LAD, Longitudinal Administrative data. LAD report #94-20-01E, may 1995. Prepared by the Small Area and Administrative Data division and Social Surveys Methods Division of Statistics Canada.