Catalogue No. 98-12

IMPACT OF EDIT AND IMPUTATION ON INCOME ESTIMATES: A CASE STUDY

Product registration Number 75F0002M

May 1998

Maryanne Webber, Income Statistics Division, Statistics Canada
Cathy Cotton, Income Statistics Division, Statistics Canada

The Income and Labour Dynamics Working Paper Series is intended to document detailed studies and important decisions for the Income and Labour Dynamics program. It is a continuation of the SLID Research Paper Series. These working papers are available in English and French. To obtain a summary description of available documents or to obtain a copy of any, please contact the Client Services Unit, at 7-B5 Jean Talon Building, Statistics Canada, Ottawa, Ontario, CANADA K1A 0T6, by INTERNET (DYNAMICS@STATCAN.CA), by telephone (613) 951-7355 or toll-free 1-888-297-7355, or by fax (613) 951-3012.

EXECUTIVE SUMMARY

Statistics Canada has recently concluded a review of its household and family income statistics programs. This review was inspired by a need to harmonize income data emanating from various surveys and administrative sources. This paper looks at the work of the task force, and at one of associated program changes, namely, the integration of two major sources of annual income data in Canada, the Survey of Consumer Finances (SCF) and the Survey of Labour and Income Dynamics (SLID).

TABLE OF CONTENTS

1. INTRODUCTION

Statistics Canada has recently concluded a review of its household and family income statistics programs. This review was inspired by a need to harmonize income data emanating from various surveys and administrative sources. Apart from issues of coverage, data collection techniques, response rates and classification issues, edit and imputation procedures appear to have a considerable impact on the final results.

This paper looks at the work of the task force, and at one of associated program changes, namely, the integration of two major sources of annual income data in Canada, the Survey of Consumer Finances (SCF) and the Survey of Labour and Income Dynamics (SLID). This integration has triggered a number of evaluation studies that shed some light on the impact of edit and imputation on the final survey results.

The following section reviews the context, mandate and recommendations of the Income Statistics Task Force. The major data issues that need to be addressed in integrating SLID and SCF are outlined in Section 3. The subsequent three sections discuss some of the quantitative differences between the two surveys, including differences in low income rates.

2. INCOME STATISTICS TASK FORCE

2.1 Background

The Survey of Consumer Finances began in the 1950s, and has been a high-profile source of information on income distributions in Canada. The Census of Population and Housing began collecting income data in 1961 and, for many years, the SCF and the Census were the only major sources of family income data.

In the 1970s, Statistics Canada began exploiting income tax file data as a statistical source but for many years, the coverage of the tax system (both in terms of the population covered and the income sources covered) was not sufficiently high to serve as a comprehensive source of information. This changed in the 1980s, when various tax credits were introduced in Canada. The availability of these credits had the effect of increasing the number of tax filers. Moreover, several non-taxable sources of income (formerly unreported) are now reported because they affect entitlements to the tax credits. The result has been a great improvement in the coverage of the tax file data. Although the SCF and tax file data have different strengths and weaknesses, the results are increasingly compared to each other. Differences can be difficult to explain and risk causing some confusion.

In 1993, the Survey of Labour and Income Dynamics started up. Although primarily intended to provide longitudinal labour, income and family data, SLID is also capable of producing cross-sectional income estimates. The income questionnaire is very similar to SCF's, although survey respondents are offered the option of allowing us to access their tax file information directly (assuming they have filed a tax return) rather than completing an income interview. Currently, tax file data are used for 75% of all respondents. SLID has become a third potential major source of annual data on income distributions. While there are benefits in having a range of sources, the scope for conflicting results and confusion among users has also expanded.

Finally, SLID is not the sole source of data on income dynamics, A sample drawn from the tax file has been linked longitudinally, and provides a second source of information on for the analysis of income transitions, spells of low income, and so on. The Longitudinal Administrative Database (LAD) has the advantage of covering a longer time period than the SLID file.

Income data are very high on the agenda of governments, policy analysts, anti-poverty groups and academics. There was increasing concern in Statistics Canada about the possibility of deriving inconsistent messages about trends in average income, income inequality or income inadequacy from the various data sources. Accordingly, an Income Statistics Task Force was created in 1996, with a mandate to recommend ways of harmonizing the Agency's income statistics; of producing them at lower cost; of improving the range and quality of income estimates; and to develop a conceptual framework.

2.2 Task force recommendations

The work of the task force ended early in 1998. The main recommendations were, first, to complete the conceptual framework. This framework will cover income, expenditures, assets and debts. In addition to articulating the concepts and their interrelationships, this framework will develop "ideal" operational definitions that are thought to be achievable, and then evaluate the current data sources against these operational definitions. The framework will also be linked to the conceptual framework of the National Accounts. Differences will be noted explicitly, along with their rationale. This is a major undertaking and will take some time to complete but the document is pivotal to our efforts to align the various income data sources.

Second, throughout its deliberations, the task force members were repeatedly struck by the variation in edit and imputation procedures across the various data sources, and their potential impact on the final results. Because income is considered a sensitive topic by the public, income surveys (or income questions included in other surveys) are prone to higher refusal rates than, for example, labour market surveys. Moreover, some respondents can recall that a certain source of income was received but not the amount. In short, income data may require substantial editing and imputation, and discrepancies between data sources can arise because of differences in these procedures. Therefore, the task force

recommended that a set of data processing guidelines be developed, taking the best practices from the various sources of income data.

Third, income estimates from surveys should be post-stratified using administrative data. This would not only help to harmonize estimates from various sources but would also stabilize the estimates and compensate for some of the weaknesses inherent in survey results, such as poor representation of high income earners, while retaining their subject matter richness.

Fourth, a new Income Statistics Division should be created, responsible for producing information on income, expenditures, assets (including pensions) and debts. This recommendation, which will facilitate the achievement of the other recommendations, has already been implemented.

The work of the task force was actually the continuation of a process that began with the decision to integrate the Survey of Consumer Finances and the Survey of Labour and Income Dynamics, for reasons of harmonization and efficiency. The rest of this paper looks at what we have learned so far about the differences between the two surveys.

3. INTEGRATION OF SLID AND SCF

The decision to integrate the two surveys was made in 1995. Over the 1996-1998 period, the surveys are running in parallel. This is a period where we can evaluate the differences, eliminate the unnecessary ones, understand and document the remaining ones. In 1999, SCF will be discontinued. The cross-sectional income estimates for the 1998 reference year will come from SLID.

Over the past two years, the staff of SLID and SCF have been collaborating closely to ensure the smoothest possible transition to the new survey. This is matter of concern in the user community, as the SCF data feed into

microsimulation models used to develop and monitor social and fiscal policies. The results are also very much in the public eye, particularly from the perspective of poverty measurement and poverty reduction.

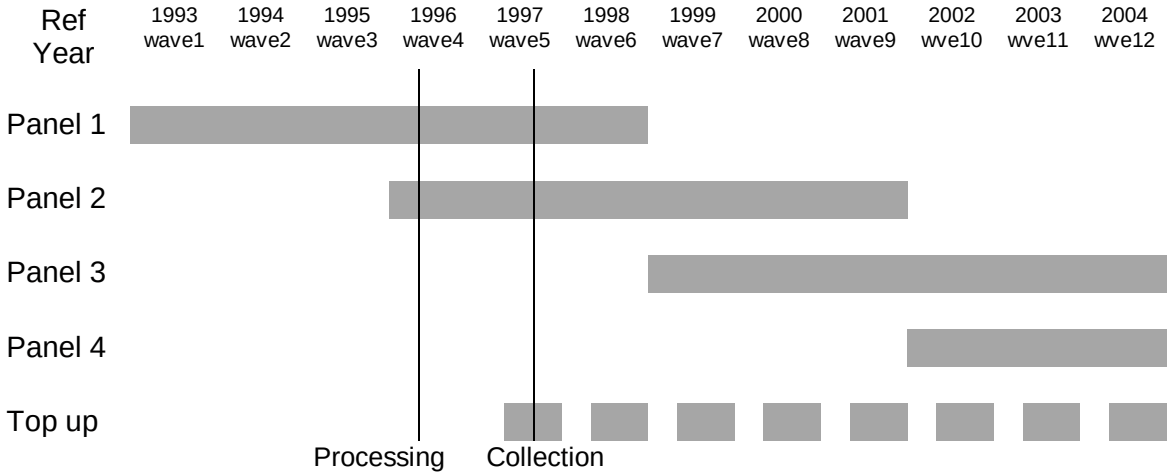
The main concerns expressed by data users are, first, what will be the impact of a shift to tax data? As noted earlier, SLID is a household survey but it uses respondents' tax information where possible rather than completing an income interview. The reasons for doing so are to offset respondent burden and increase precision. Second, what impact will the shift to the SLID sample have? The sample selection procedures in the two surveys are virtually identical but the SLID sample is subject to attrition and under-representation of immigrant households. Third, what is the impact of the differences that exist between the two processing systems? The SLID and SCF staff have begun a series of studies to document the global quantitative differences between the two surveys and to examine more closely the specific concerns expressed by data users.

Before addressing these issues, a few words on SLID's stage of development would be useful. Unlike SCF, SLID is a new survey. Currently, two years of data are available, for 1993 and 1994. When the 1993 data became available, a comparison with SCF was undertaken. There were several differences, including a gap of about 4% ($20 billion) in the estimate of aggregate income derived from the two sources. Some fine-tuning of the SLID system ensued and the 1993 gap declined to $14 billion. The point here is that SLID system is still young, although the necessity of producing cross-sectional income estimates is ensuring a very rapid road to maturity.

Currently, the SLID/SCF comparisons are based on 1994 data. The 1995 data have been recently finalized and the 1996 data, currently being processed, will be available in December 1998. By that time, it will be possible to compare results over a four-year time period.

For reference purposes, the SLID sample is made up of six-year panels that overlap (Chart 1). Each panel has 15,000 households and information on labour market activities and income is collected from all household members aged 16 and over (about 30,000 persons per panel). The first panel started in 1993, and the second in 1996. In 1997, an annual top-up sample was added to improve cross-sectional quality. In all, this yields a target sample of about 37,500 households or 75,000 persons aged 16 and over. The sample is slightly larger than the SCF sample. Both are selected from the same area frame.

Chart 1: SLID Panels and Top-up Sample



4. SLID/SCF COMPARISONS

Estimates of aggregate and average income from the two surveys have been compared, as well as income distributions and low income rates. In general, these comparisons are reassuring, but there are features that will require further analysis.

4.1 Aggregate income

Aggregate income refers to the sum of the income reported for the full population covered by the survey. It can be useful as a broad measure of over- or under-estimation for various sources of income through comparisons to external data sources, such as taxation data and National Accounts.

Table 1: Aggregate Income, Selected Income Sources, SLID, SCF and Taxation Statistics, 1994

	SCF $B	SLID $B	Tax $B	SCF/ Tax	SLID/ Tax
Total Income	527.5	533.1	546.8	.96	.97
Wages & Salaries	373.0	375.9	332.5	1.12	1.13
Old Age Security	19.3	18.5	17.6	1.07	1.05
Employment Insurance	13.0	14.6	14.4	.90	1.01
Social Assistance	11.2	12.0	12.0	.93	.84
Worker's Compensation	2.7	3.5	3.5	.75	.97
Private Pensions	20.9	25.9	24.7	.85	1.05

As Table 1 shows, both surveys yield estimates of aggregate income that are quite close to the taxation estimates. It is not surprising that the SLID results are slightly closer, because three-quarters of the data come from the tax file.

However, there are some interesting anomalies when the results for particular income sources are examined. For example, SLID underestimates social assistance (income assistance) to a greater extent than SCF. Moreover, the taxation data themselves are thought to be an underestimate of social assistance payments. Some possible explanations are:

- attrition in the SLID sample is higher at the lower end of the income scale because the population in question is more mobile and difficult to trace;

- differences in imputation procedures, which result in greater boosting of social assistance in income in SCF than they do in SLID.

Wages and salaries earned by employees, as estimated by the two surveys, exceed the level indicated by the taxation data. This could be due in part to unreported earnings in the taxation data but this is certainly not the whole story. As will be seen later, average incomes are very close. Another possible reason lies in the demographic estimates (population by age group, sex and province) that are used to weight the sample data. Both surveys use the same set of independent estimates to benchmark their data, which will tend to align them with each other. But the population estimates are themselves subject to error so that the survey data may differ from other sources (such as the taxation statistics) that are conceptually comparable but not dependent on the same demographic estimates.

Table 2: Aggregate Income Before and After Tax, SCF, SLID and Taxation Statistics, 1994

	SCF $B	SLID $B	Tax $B	SCF/ Tax	SLID/ Tax
Income before tax	527.5	533.1	546.8	.96	.97
Taxes Paid	102.5	101.9	100.5	1.02	1.01
Income after tax	425.0	431.2	446.4	.95	.97
After tax/Before tax ratio	.81	.81	.82		

In SLID, income tax information generally comes directly from the tax file, unless authorization to use tax data was not obtained. In that event, the amount is generally imputed based on a regression model of taxes payable. The model reflects among other things income level, family circumstances, allowable income tax deductions, and province. The model has been tested against actual tax file data and found to be quite reasonable. In the SCF, the methodology for imputing income taxes payable is different. In both cases, the surveys are estimating taxes owing, which may be different from taxes actually paid. This may be a factor contributing to the overestimation.

The taxation statistics are a useful benchmark, but by no means the only one. The National Accounts provide alternative estimates of aggregate income. In effect, the National Accounts gain strength from the use of several administrative and survey sources. In the table below, the estimates from the National Accounts have been adjusted to the extent possible to correspond conceptually to the content and population covered by the two surveys.

Table 3: Aggregate Income, Selected Sources, SCF, SLID and National Accounts, 1994

	SCF $B	SLID $B	NA $B	SCF/ NA	SLID/ NA
Total Income	527.4	533.1	497.9	1.06	1.07
Wages & salaries	373.0	375.9	349.9	1.07	1.08
Investment Income	18.7	20.6	35.3	.53	.58
Old Age Security	19.3	18.5	18.8	1.02	.98
Employment insurance	13.0	14.6	14.9	.87	.98
Workers' Compensation	2.7	3.4	3.9	.68	.88

In this comparison, both surveys overstate aggregate income. Wages and salaries (the overwhelmingly largest component) is overstated to roughly the same extent. Among the other income sources shown in Table 3, investment income stands out. Household surveys typically have great difficulty capturing investment income, and these two are no exception. This phenomenon, which is probably due to a combination of factors (survey non-response of high income earners and under-reporting of investment income by survey respondents) is one instance where benchmarking to external estimates would have a very beneficial effect.

4.2 Average income

The following two charts show the proportion of the population reporting income, and average income estimates. The estimates of the population reporting some income are somewhat higher in SLID, particularly in the 24 to 44 age band. This is important because the average income estimates are calculated by dividing aggregate income by the number of income earners, so a higher number of income earners will drive down the average income level, other things being equal.

The reason the estimates of income earners are higher in SLID appears to be largely due to differences in the way the processing systems handle government transfers that are essentially made to the family unit. SLID assigns them to whatever family member reports them. SCF used a more complex approach, but has since aligned its procedures with SLID.

Chart 2: Proportion of persons with income by age group, 1994

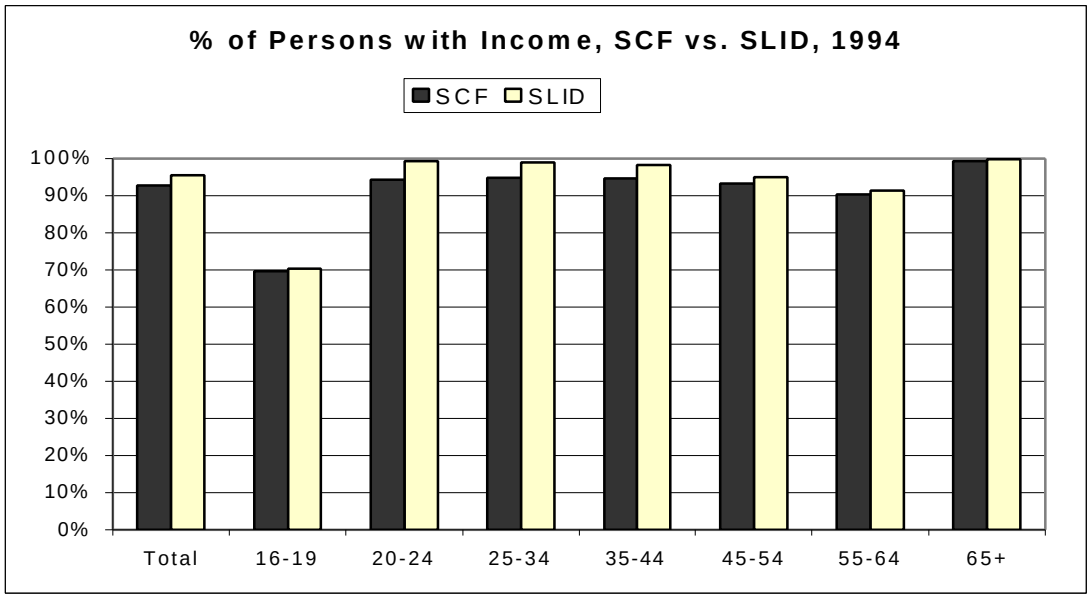
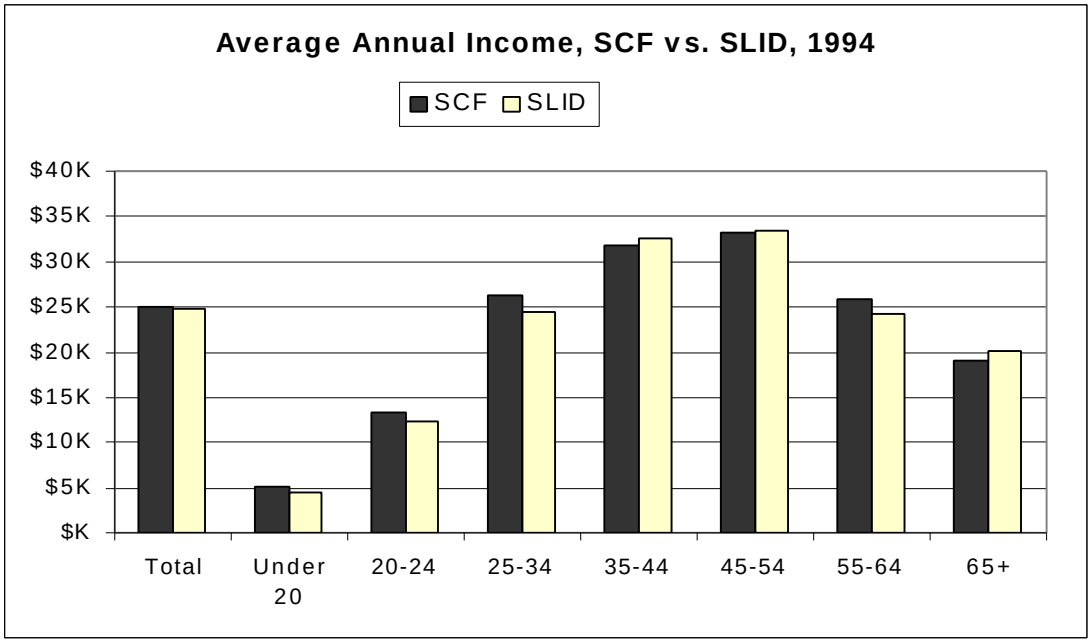


Chart 3: Average annual income, individuals with income by age group, 1994

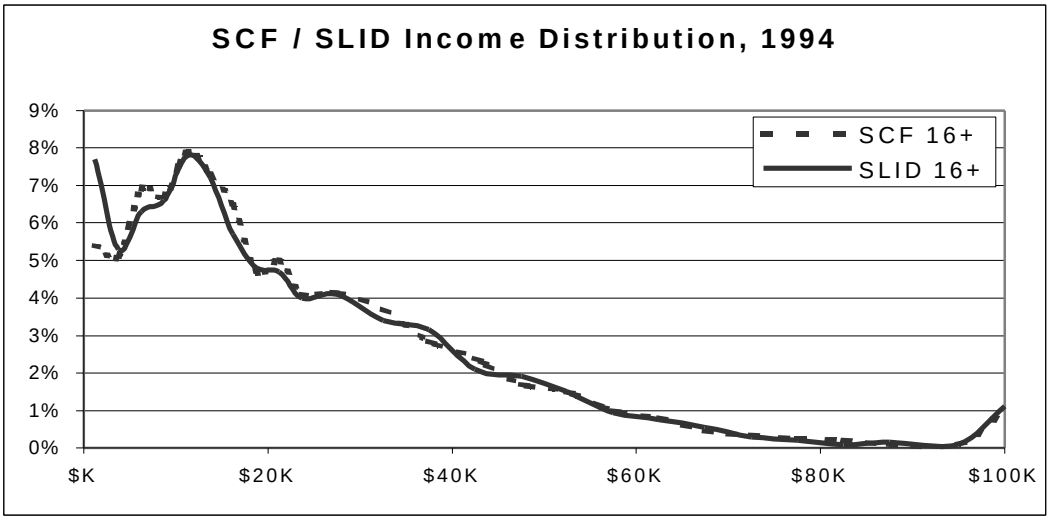


The average income for all ages is similar in SCF and SLID. However, SCF shows higher salaries in the three younger age groups, while SLID gives higher salaries in three of the four older age groups.

4.3 Income distribution

Income inequality is a major concern and so it is important to examine the "story" that the two surveys tell with respect to income distribution: are they similar? The fact that SLID uses tax data might lead one to suspect that it would capture more in the high end of the income distribution curve than SCF, which relies entirely on reported income. However, the opposite is the case. SLID actually captures a higher proportion of low income earners (Chart 4) than SCF.

Chart 4: Distribution of population by personal income (before tax), 1994



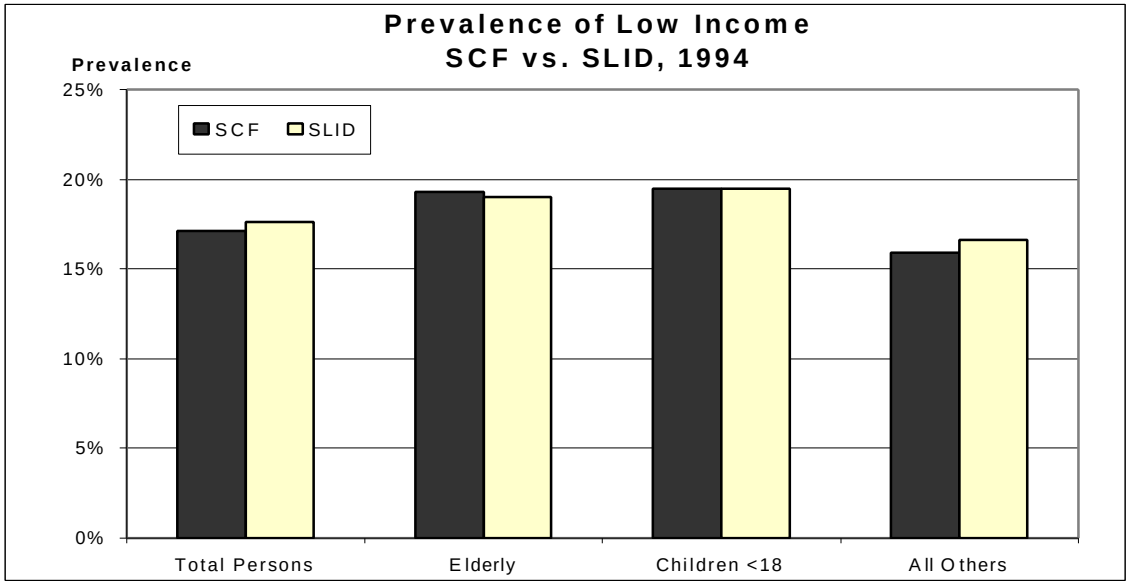
This is certainly a phenomenon that will require further study.

4.4 Low income rates

Statistics Canada produces a measure of low income called the Low Income Cutoff (LICO). This measure is based on household income and expenditure data. It is the line where 55% of a family's before-tax income is likely to be spent on food, shelter and clothing. LICOs are calculated for different family sizes and different sizes of community.

Although low income rates are contentious, movements receive a great deal of public scrutiny. Therefore, consistency in the rates between SLID and SCF is an important issue. Chart 5 compares the rates for 1994.

Chart 5: Low income rates, 1994



Despite the higher aggregate and average income levels, SLID produces higher estimates of low income. However, the results are close: the prevalence of low income is 17.1% based on SCF and 17.6% based on SLID. The rate for SLID among seniors (aged 65 and over) is lower than SCF's, and there are indications that this occurs because of SLID's use of tax data.

5. IMPACT OF USING TAX DATA: STUDY

Since data users have expressed particular concern about the fact that SLID uses tax data, a special study was devised to assess the impact of this factor. The SCF sample for 1995 was matched to the tax file, using such matching characteristics as name, address, date of birth and marital status. A match rate of about 80% was achieved. For those records where a match was found, we substituted the tax data for the income data collected via interview. This gave us two data sets to compare – the original SCF data and the SCF records with tax data substituted for 80% of the records – to allow some analysis of the impact of moving to tax data holding other factors constant.

Table 4 shows that the use of tax data increases the aggregate income estimate by $11 billion or 1.9%, based on this set of respondents.

Table 4: Aggregate income, SCF and combined Tax/SCF, unattached individuals and families*, 1995

	SCF $B	Combined Tax/SCF $B	Difference
Total	549.8	560.2	1.9
Unattached individuals	93.6	93.4	-0.2
Elderly	20.9	21.3	2.2
Non-elderly	72.7	72.0	-0.9
Families	456.2	466.8	2.3

* Based on economic families, defined as all persons related by blood, marriage or adoption and living in the same dwelling.

Table 4 highlights the situation of persons aged 65 and over living alone, as this group has historically had high rates of low income. Both elderly persons living alone and families show higher levels of income based on their tax data than what was reported in the SCF.

The low income rate (which is based on each person's family income) is nevertheless somewhat higher when tax data are used: 18.5% compared with 17.8% in SCF, indicating that the underlying income distribution is different.

6. SIMULATION STUDY

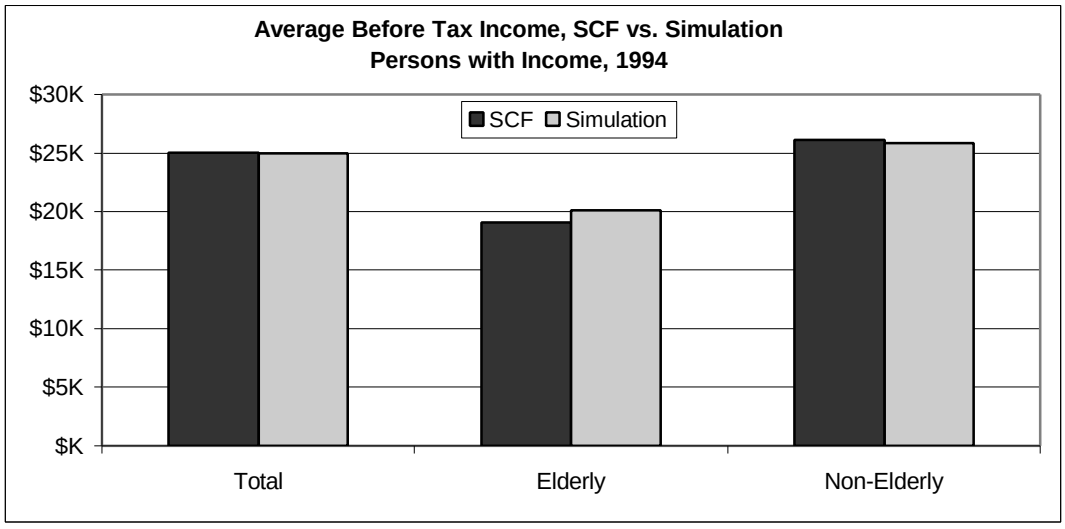
A second study was undertaken to assess the impact of moving to the SLID sample. There is a risk that sample attrition will bias estimates of family income since geographical mobility and our ability to successfully trace movers is thought to vary according to income level. Also, the SLID panel design leads to possible under-representation of immigrant households. At the beginning of a panel,

immigrant households are represented but, as the panel ages, new households made up solely of immigrants are not captured. (Immigrants who move in with existing households are represented in the sample because SLID captures information on persons who move in with anyone originally selected for the panel).

To test the impact of sample differences, data for the SLID respondents were processed through the SCF processing system. This required some preparatory work to make the SLID input variables "look like" the SCF variables. However, by comparing the results of this simulation with the SCF results, we gain some understanding of the sample differences, since the processing system is a constant.

Chart 6 shows the impact on average income estimates. There is very little difference for the population at large, although the seniors in SLID's sample appear to be "better off". This is consistent with other results indicating that the use of tax data boosts the income levels of seniors somewhat.

Chart 6: Average before tax income, SCF vs. Simulation



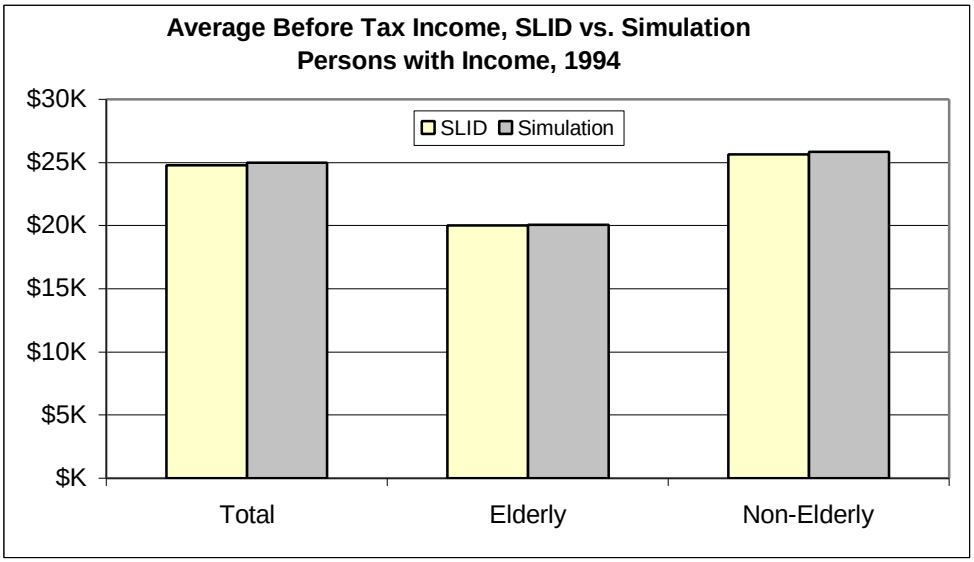
The impact on low income rates was also slight. The 1994 rate for all persons from SCF was 17.1%; the SLID sample processed through the SCF system

yielded a low income rate of 17.2%. The rates for seniors were identical. For children aged 18 and under, the simulation produced a slightly lower rate -– 18.7% compared with 19.5% for SCF.

This same simulation can also be compared to the SLID results. In this case, because the same sample records have been processed through two distinct systems, it helps to isolate the impact of processing differences.

Chart 7 compares average income estimates for SLID and for the SLID sample processed through the SCF system. At this high level, the two sets of estimates are very close indeed. However, the SCF system yielded a low income rate of 17.2% against SLID's 17.6% and, for children under 18, the rates were 18.9% and 19.5% respectively. Thus, the two processing systems do generate differences in the distribution of income.

Chart 7: Average income before tax, SLID vs. Simulation



7. CONCLUSION

These studies have shown that the differences observed between SLID and SCF in 1993 were greatly moderated through adjustments to the processing systems. On the whole, the cross-sectional income estimates for 1994 are reasonably close. These results have reassured data users somewhat, although they remain concerned about how well the similarities hold up when the data are disaggregated.

The evaluation studies have also shown us how difficult it can be be to disentangle and pinpoint exact sources of differences between surveys so that these can be documented and shared with data users. The work we have done so far is helping to point out the areas where further work is needed. Beyond the SLID/SCF merger, it is a first step towards the development of processing guidelines that will help us to align income estimates from various sources.

8. ACKNOWLEDGEMENT

The authors would like to thank Peter Hewer for the production of the tabulations upon which this paper is based.