Catalogue No. 97-06

SURVEY OF LABOUR AND INCOME DYNAMICS:
PROCESSING STRATEGY FOR WAVE 1 INCOME DATA

Product Registration Number 75F0002M

March 1997

Élaine Fournier, Household Surveys Division

Tracey Leesti, Household Surveys Division

Mylène Lavigne, Social Survey Methods Division

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 Dissemination Unit, at 7-C6 Jean Talon Building, Statistics Canada, Ottawa, Ontario, CANADA K1A 0T6, by INTERNET (DYNAMICS@STATCAN.CA), by telephone (613) 951-4633, or by fax (613) 951-3012.

EXECUTIVE SUMMARY

This report documents the edit and imputation approach taken in processing Wave 1 income data from the Survey of Labour and Income Dynamics. The intention is to inform data users who are interested in these procedures and also to ensure that the rationale for taking certain measures is documented for future reference. As this report is being prepared, the survey staff is processing Wave 2 data, using this report as a reference tool.

TABLE OF CONTENTS

Page

1. BACKGROUND

From a processing perspective, the SLID income data are unusual in two ways. First, respondents had the option of allowing access to tax data rather than completing an income interview. Since the data for some respondents come from the tax system, while for other respondents they were collected by interview, the processing system must merge data obtained two different ways. The differences in content between tax and collected data are not accidental. The explicit income sources used in the interview are essentially the same as those used in the Survey of Consumer Finances (SCF). Content decisions rest on conceptual reasoning, tempered by expectations of what can realistically be achieved in a telephone interview with a respondent who may or may not have access to financial documents. In contrast, the content of the tax file is essentially the set of inputs needed to calculate taxable income, deductions and eligibility for tax credits. A series of decisions were needed to bring these income sources together.

The second way in which SLID income data are unusual is that they are longitudinal. Beginning with Wave 2 processing, there will be issues to deal with regarding the use of previous years' data in editing or imputing the current year's information.

2. CONCEPTUAL ISSUES

2.1 The Processing System

The income portion of SLID's processing system is divided into four main parts, or "entities" (Figure 1). The first entity, INCCOL, contains income information obtained by interviewing a person. The full entity, which contains survey data as collected, is intended for internal use only, and none of the variables are included on the public use file. A record exists for all persons for whom income data were collected (partial or complete).

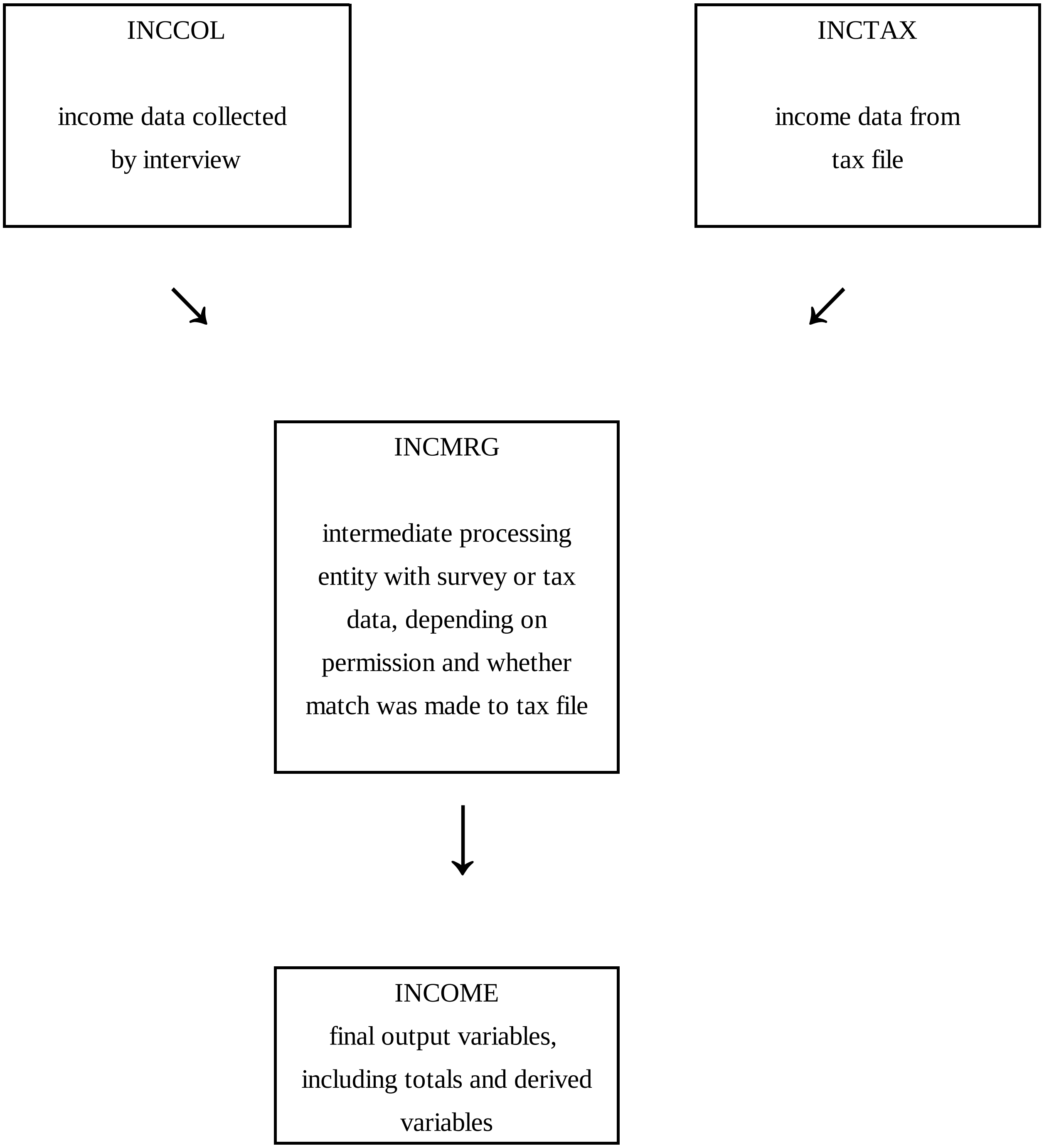
The second entity is INCTAX. This entity contains tax data obtained by linking to Revenue Canada tax files. As with the INCCOL entity, the full entity is intended for internal use only, and none of the variables will be included on the public use file.

The third income entity in the data model is INCMRG. This entity "merges" the survey and tax data into a single set of variables. If a respondent gave permission to link to Revenue Canada,
and the link was successful, their tax return information is stored in this entity. If an individual did not give permission, or did not link, then their survey data is stored in this entity. One record exists for each person aged 16 and over at the end of the reference year.

The INCOME entity is the final entity used for data dissemination. As with INCMRG, it is a blend of tax and survey data. This entity also includes imputed values, various total incomes and derived variables.

For 1993 data, as the tax permission question had not yet been asked, all information was obtain through the survey. All respondents (partial or complete), therefore, had a record on the INCCOL entity. Permission to use tax data was first sought during May 1995 collection. Permission is retroactive, covering the six years of the panel. For those who gave permission to link, 1993 tax data were used, while survey data were used for the remainder.

Figure 1. Four Entities of Income Portion of SLID Processing System


INCCOL
income data collected by interview
INCTAX
income data from tax file
INCMRG
intermediate processing entity with survey or tax data, depending on permission and whether match was made to tax file
INCOME
final output variables, including totals and derived variables

2.2 Merging of Tax and Survey Data: Definition of Income Categories and Total Income

As noted above, the income interview items and the tax return lines do not line up perfectly. Table 1, below, shows how a unique set of income categories was created from the two data sources.

Table 1: Definitions Used to Merge SLID Survey Income Data and Tax Sources for 1993

Variable	Survey source	Tax source	Comments on comparability
1. Employment income			
Wages and salaries	Wages and salaries *Q1*	Employment (wage and salary) income *L101*	Fringe benefits: In 1993, not reported in survey data (at least in principle), but they are in tax data. From 1994 onward, fringe benefits included with wages and salaries in both sources.
Farm self-employment net income	Farm self-employment net income *Q2*	Farming income *L141*	Checked for consistency with class of worker and industry as reported in labour interview. Persons reporting income here should have an unincorporated farm.
Non-farm self-employment net income	Non-farm self-employment net income *Q3*	Business, professional, commission, fishing income *L135, L137, L139, L143*	Consistency check with class of worker and industry.

2. Investment income			
Investment income (incl. net dividends)	Interest *Q4* + net dividends *Q5* + other investment income *Q7*	Dividends *(L120* (x.8))+ interest & other investment income *(L121)* + Net partnership income *(L122)* + Net rental income *(L126)*	- Only broad category of investment income is available in merged income data – many differences between the two sources *within* investment category. - Interest from loans and mortgages and regular income from a trust fund or estate included with other investment income in the survey. No direct reference to these items on tax form. - Survey collected net dividends. Taxable dividends obtained by multiplying net (i.e., real) dividends by 125%. From 1994 onwards, survey asks for taxable dividends. - Only difference between the 2 investment variables is treatment of dividends. Investment income used for total money income. Taxable investment income used for calculating taxes payable.
Taxable investment income (incl. taxable dividends)	Interest *Q4* + taxable dividends *Q5* (x 125%) +other investment income *Q7*	Dividends *(L120)* +interest & other invest. income *(L121)* + Net partnership income *(L122)* + Net rental income *(L126)*	
3. Government transfers			
Child Tax Benefits	Child tax benefit *Q8*	Not on T1 but available in separate file from Revenue Canada	For survey route, if amount not reported, amount calculated based on number of children and net family income.
Old Age Security, Guaranteed Income Supplement, Spouse's Allowance	OAS/GIS/ SPA *Q9*	OAS *(L113)* + Federal supplement *(L146)*	OAS and GIS/SPA available as separate variables, because OAS is taxable and GIS/SPA is not. On survey side, amount is Q9 can be split because OAS is generally a standard amount.
Canada or Quebec Pension Plan benefits	CPP/QPP *Q10*	CPP/QPP *(L114)*	

Unemployment Insurance benefits	UI benefits *Q11*	UI benefits *(L119)*	
Social assistance	SA and PIS *Q12*	Social assistance *(L145)*	
Worker's Comp benefits	WC *Q13*	WC *(L144)*	
Goods and Services Tax Credit	GST credit *Q14* – imputed if eligible and GST not reported	Not on T1. Imputed based on family income	For 1994, it will be possible to use GST file rather than imputation, as for CTB. Note that GST credit is also imputed to eligible survey respondents.
Taxable government transfers			Roll-up of OAS, CPP/QPP and UI. Excludes taxable income from government sources other than those listed above, which is counted in other taxable income.
Non-taxable government transfers			Roll-up of CTB, GIS/SPA, GST credits and Worker's Compensation.
4. Pension income			
Employer pension, RRSP withdrawals and RRIF annuities	Employer pensions *Q18* + RRSP annuities & RRIF withdrawals *Q19*	Other pensions and superannuation *(L115)* + RRSP income (annuity or withdrawal) if age 65+ *(L129)*	In tax data, cannot distinguish between RRSP annuities and withdrawals from unmatured RRSPs (which we are not counting as money income). Assumption is made that, for persons 65+, amount in L129 is an annuity; for persons under that age it is a withdrawal.

5. Other taxable income			
Alimony & child support	Alimony, separation allowance, child support *Q21*	Alimony or maintenance income (taxable) *(L128)*	On tax side, generally refers to amounts paid under agreement or court order. Survey refers to same thing and references L128.
Other taxable money income	Other taxable money income – taxable subset of *Q17, Q23, Q24 & Q25*	Other income *(L130 + L104*)	Ex: severance pay, income maintenance, employer or union supplementary unemployment benefits, Children's Aid payments, scholarships, death benefits (survey). Also other taxable government income from sources not explicitly identified above.
6. Items not included in money income but needed to determine taxable income			
RRSP withdrawals	RRSP withdrawals *Q20*	RRSP income (annuity or withdrawal) if <age 65 *(L129)*	Withdrawals from unmatured RRSPs not counted in total income, but included in taxable income. See note under pension income.
Taxable capital gains	Capital gains *Q6*	Taxable capital gains *(L127)*	Taxable is 75% of net. Taxable capital gains included in total taxable income but excluded from total money income.
7. Other excluded items			
Provincial tax credits	Provincial tax credits *Q15*	Prov tax credits *(L479)*	Excluded because data on tax side not available for some provinces. Also, on survey side, appeared to be substantially under-reported.
Veterans' pensions	Veteran's Pensions *Q16*		Collected in survey but not included in merged dataset (or any income totals) because not reported in tax data.

2.3 Labour-Income Dependencies

There are many links between the demographic and labour market information collected in SLID. Dependent interviewing - the feeding back of previously collected information to respondents- helps to reduce inconsistencies between labour information reported in January and income information reported in May. However, not all inconsistencies can be resolved during the interview itself; even after the interview, inconsistencies remain. Furthermore, about two-thirds of respondents report income by the tax route. A decision was made early in processing not to adjust data obtained from tax records. In some cases, however, income data collected by interview are modified. The reverse may also happen: labour data may be adjusted if they are inconsistent with income data that appear to be correct.

In January, respondents were asked if they had received, at some time during the year, Workers' Compensation, Unemployment Insurance benefits or Social Assistance. The January interview also determines if a person worked as an employee, or if they were self-employed during the year. Based on their responses, income of various types should be reported in the May interview. For example, wages and salaries would be expected in May if the respondent reported at least one paid worker job in January. There was, in fact, a "flag" set based on the labour interview and, if the expected income was not reported in May, the flag triggered a probe question in an attempt to resolve the inconsistency[1].

As mentioned above, the survey asked about the months in which social assistance was received in the context of the labour interview. The difficulty with this variable was not only the potential for conflicting responses between what was reported in the labour

[1] This is an "interactive edit" made possible by the use of computer-assisted interviewing.

interview and the presence/absence of an amount in the income interview, but in addition, the fact that social assistance is a family source of income. It may be reported by one person in the labour interview and by a different person (usually a spouse) in the income interview (or on the tax return). Some families may split the amount in reporting it. In short, there are many possible reporting patterns and the editing strategy can be very light-handed or very heavy-handed. A light-handed approach was adopted, because greater intervention would entail too many arbitrary decisions. The following principles were implemented:

- Do not "force" one single member to report all social assistance income for the family. Do not distribute social assistance amount evenly among all members of the family. In other words, let the data reflect how families choose to report social assistance.

- If an individual reported receiving social assistance in the labour interview, but some other family members reported social assistance in the income interview, change the labour interview reporting to make it consistent (this change has no effect on family level data).

- If an individual reported social assistance in the income interview, but no one in the family reported it in the labour interview, changed the individual's status in the labour interview to "received social assistance" (months received are not imputed, though).

- If an individual reported receiving social assistance in the labour interview, but no amount is reported in the income interview by any family member, impute an amount for social assistance. This procedure was only applied to cases where income data was collected by interview.

The labour interview also determined the respondent's class of worker, e.g., did they work during the year as an employee, on a self-employed basis, etc. The next table shows some of post-collection processing checks that are made between the class of worker reported in the labour interview and type of income expected for each respondent during the income interview.

TABLE 2. Consistency Between Class of Worker and Income Information

If known from labour interview...	***...then value expected under:***
Workers' compensation was received at some time during the year	*Workers' compensation (if missing, and income data were obtained by interview, an amount will be imputed)*
Unemployment insurance benefits were received at some time during the year	*Unemployment insurance benefits*
Age 65+, with some work experience and Canadian citizen for at least 10 years	*Canada Pension Plan or Quebec Pension Plan*
For at least one job, class of worker is an "employee"	*Wages and salaries*
For all jobs, class of worker is "unpaid family worker"	*Earnings are zero*
For at least one job, class of worker is "self-employed, incorporated"	*Wages and salaries OR (and) investment income*
For at least one job, class of worker is "self-employed, unincorporated"	*Farm self-employment income or non-farm self-employment income, depending on the industry*

When inconsistencies are apparent from the above comparisons, the approach for resolving them depends on the source of income data. For income data collected by interview, if an amount appears missing or inconsistent, it may be imputed or edited. For income data obtained from the tax source, inconsistencies are resolved if possible by changing the class of worker, provided the nature of the error is quite apparent. Otherwise, the inconsistency remains.

Total annual earnings and the composite hourly wage rate (labour interview data) may be imputed, if missing, using annual wages and salaries (income data), provided the person reported only one paid worker job and the value collected on wages and salaries is good. The total annual earnings are set to the amount for annual wages and salaries, and the composite hourly wage rate is derived from total annual earnings and information on hours worked. Users should also be aware that the wage rate and annual earnings derived from labour interview data may be different from the annual wages and salaries.

When other labour information appears to be missing, for example if annual wages and salaries are positive but no job was reported in the labour interview, no imputation is done because of the complexity of imputing jobs and job characteristics. These inconsistencies remain in the data and can be identified by comparing job information with income information.

2.4 Some Basic Edit and Imputation Principles

Tax Data Not Altered

For roughly half of the respondents in Wave 1, the data come from the tax file. A decision was taken to not alter tax data if inconsistencies with labour or with demographic data were detected. There are three reasons for this decision:

- It is difficult to edit income data “lightly” because of the relationships that exist between the various income items. For example, if a specific source is altered, it affects the total income and taxes payable. A decision to intervene has a snowballing effect.

- Editing of the income tax data would tend to introduce longitudinal inconsistencies.

- It would add complexity to an already complex processing system.

In hindsight, it appears a reasonable precaution to perform some basic checks and a few will probably be introduced for Wave 2.

Discrepancies Between Wages and Salaries and Total Earnings from Paid Employment Are Allowed

In the context of the labour interview, information is collected on all paid worker (“employee”) jobs to allow the calculation of an hourly wage rate and total annual earnings for each job. The sum of total annual earnings from all paid worker jobs is more or less conceptually equivalent to the wages and salaries collected during the income interview. It was decided that discrepancies between the two would not be resolved through an edit and imputation process. This is the first time we can conduct a micro-level analysis of the differences in reporting patterns that arise between these two well-established approaches for collecting earnings data. It is possible that editing will be done in future years, once the patterns are well understood.

3. EDITING OF COLLECTED SURVEY DATA

Once survey data are collected, a number of edit steps are performed. The following describes the steps taken to clean up the survey data before it was merged with tax data.

3.1 Classify and React to Interviewer Notes

Two steps were involved in handling notes: coding them and reacting to them. Notes that were related directly to income were coded as such using a custom-made coding program.

In order to react to the income notes, a list of actions for various different types of income notes was created and referred to when making corrections on the income file.

3.2 Compare Total Income Entered (Q26) With the Sum of the Components (Q01 to Q25)

The next step was to compare total income as entered (Q26) with the sum of the components (Q01 to Q25) in order to check the quality of the data. It should be kept in mind that the total income **entered** in Q26 is not necessarily the total income **reported** by the respondent but may instead represent the income **calculated** by the computer.

As an interviewer enters amounts during the income interview, a running total of those amounts is calculated by the computer. When all income has been entered for the respondent, the interviewer asks the respondent if the running total appears to be correct. If the respondent disagrees with the total, the total can be overridden by the interviewer.

In general, the only cases where total income and the sum of the components should **not** be equal is where the respondent gave "don't knows" or refusals for certain items or gave no amount except for total income.

It should also be noted that when the total was calculated by the computer, the amount was rounded to the nearest dollar. Thus, to make our comparisons between the total income entered and the sum of the components, it was assumed that the two were equal when the difference was less than $2. The results obtained, shown in Table 3, indicate that in 91.3% of cases (115+233+12+20,116), the data appear to be correct. Thus, for 89.7% (20,116) of cases, the total income entered -- or calculated by the computer -- is exactly equal to the sum of the components (see table below).

Table 3. Comparison of Total Income Recorded in Q26 With the Sum of the Components (Q01 to Q25), for Cases Where Total Income >0

Types of response for questions Q01-Q25 as a group	Results of comparison			
	Sum < Q26	Sum = Q26	Sum > Q26	Total
No income reported for any	115	12	0	127
Yes for some questions**/no	352	20 116	491	20 959
Yes / no / don't know or refusal	99	997	18	1 114
No / don't know or refusal	233	0	0	233
Total	**799**	**21 125**	**509**	**22 433**

* $0 or blank ** amount

Manual checking of the various types of errors was carried out in an attempt to determine how they should be corrected. In some cases either the total income or a component was changed depending on the logic of the situation. In other cases, total income was left as reported, because it is impossible to know which amount is correct (the sum of the components or total income). In any event, total income was recalculated by adding up the components after imputation has been performed.

3.3 Edits

3.3.1 Child Tax Benefit

To receive Child Tax Benefit, a person has to be responsible for a child who is less than 18 years old. Since we thought there might be a line reporting error for older people who said they received CTB, we looked at all persons 60 and over who had reported an amount in Q08 (child tax benefit). We located three persons, and all three did in fact have a child.

3.3.2 Old Age Security, Guaranteed Income Supplement, Spouse's Allowance

In general, all persons aged 65 and over should receive Old Age Security income (unless they immigrated less than 10 years ago or have an income in excess of $80,000 on retirement). Because the eligibility conditions are known, it was possible to identify individuals who had not reported OAS, but in all likelihood received it. We found that some persons aged 65 and over had not reported any Old Age Security or Guaranteed Income Supplement (Q09). Why?

For those persons aged 65 and over for whom Q09=0, we first compared their survey data with their income tax data. In general, we found that the amount that should have been reported in Q09 was reported elsewhere in the survey or was included with the amount in Q10 (Canada or Quebec Pension Plan benefits). If the exact amount for Q09 (for example, $4,586.16 in 1993), or an amount very close to it, appeared elsewhere than in Q09, it was moved. It was also decided to include in the interviewer's manual somewhat more detailed information concerning the pensions of elderly persons. It is hoped that as a result there will be fewer errors in future years.

3.3.3 Employer Pension

Table 4 shows the respondents aged 59 and over who did not have an amount in Q18, as compared to what appeared on the tax return.

Table 4. Tax/Survey Comparison of Employer Pension (Q18) for Persons Aged 59 and Over

Survey (Q18)	**income tax**		
	amount	**zero**	**total**
don’t know	54	5	59
refusal	27	8	35
zero	437	1345	1782
total	**518**	**1358**	**1876**

Of the 1,876 individuals in this situation, 28% reported no employer pension in the survey, but had an amount on their tax return. The same table by more detailed age groups indicates that:

- 16.8% of persons aged 59 to 64 should have reported an amount in Q18 but did not;
- 26.7% of persons aged 65 to 69 should have reported an amount in Q18 but did not;
- 24.8% of persons aged 70 and over should have reported an amount in Q18 but did not.

3.3.4 Canada Pension Plan/Quebec Pension Plan

To evaluate the reporting quality of pension income, amounts for this question reported in the income interview were compared to tax file data. It was discovered that Q10 was often missing on the survey. Tables 5 to 8 outline the problem.

Table 5. Tax/Survey Comparison of Canada Pension Plan and Quebec Pension Plan (Q10), All Ages

Survey	**Income Tax**		
	amount	**zero**	**total**
don't know/refusal	244	68	312
amount	3,057	107	3,164
zero	682	17,080	17,762
total	**3,983**	**17,255**	**21,238**

Table 5 shows that 4.5% of the study population (244+682) had no amount in the survey although a CPP/QPP was reported on their tax return. We took a closer look at tax results for older persons according to whether or not they had work experience.

Table 6. Presence or Absence of a CPP/QPP Amount on the Tax Return, According to Work Experience, Persons Aged 60-64 Years

CPP/QPP (tax)	**work experience (survey information)**		
	don’t know	**experience**	**no experience**
amount	97 32.9%	405 79.3%	11 44.0%
zero	198 67.1%	106 20.7%	14 56.0%
total	295 100.0%	511 100.0%	25 100.0%

Table 7. Presence or Absence of a CPP/QPP Amount on the Tax Return, According to Work Experience, Persons Aged 65-69 Years

CPP/QPP (tax)	**work experience**		
	don’t know	**experience**	**no experience**
amount	140 91.5%	786 91.5%	28 47.5%
zero	13 8.5%	73 8.5%	31 52.5%
total	153 100.0%	859 100.0%	59 100.0%

Table 8. Presence or Absence of a CPP/QPP Amount on the Tax Return, According to Work Experience, Persons Aged 70 Years and Over

CPP/QPP (tax)	work experience		
	don't know	**experience**	**no experience**
amount	92 93.9%	1577 86.0%	156 59.3%
zero	6 6.1%	256 14.0%	107 40.7%
total	98 100.0%	1833 100.0%	263 100.0%

The three tables above show that, according to tax data, 79% of respondents aged 60 to 64 with work experience received CPP/QPP. For those aged 65 to 69, the proportion rises to 92%. It then drops to 86% for those aged 70 and over, presumably because some of these older people had not worked since the CPP legislation was introduced in 1966.

Rules for Assigning CPP/QPP Flag

Given the under-reporting of CPP/QPP, a decision was taken to impute when this source of income had a high probability of being received. An outline of the approach, and its rationale follow. For the age groups 65-69 and 70 and over, the established rules are simple and the same, whereas for person 60-64 years of age, they are more complex.

Table 9. Survey/Tax Comparison of CPP/QPP Reporting, Persons Aged 70 and Over

SURVEY	**INCOME TAX**		
	CPP=0	CPP=amount	**Total**
CPP=0	303 14.8%	**219** **10.7%**	522 25.4%
CPP=dk or refusal	15 0.7%	**134** **6.5%**	149 7.3%
CPP=amount	33 1.6%	1349 65.7%	1382 67.3%
Total	351 17.1%	1702 82.9%	2053 100.0%

Table 9 shows that, without edits, 353 respondents aged 70 and over would erroneously be shown as not having received CPP/QPP. This represents 17.2% of the study population, or 52.6% of all respondents in this age group reporting no CPP/QPP on the survey.

To correct this, any person in this age group who had worked since 1966 was treated as non-response if no CPP/QPP was reported. An amount was imputed from a similar record.

This edit was validated. Table 10 shows that this approach results in the correct action being taken in 492 cases and the incorrect action in 179 cases (the total of 671 represents all persons 70 and over who had no amount for CPP/QPP reported during the interview, whether it was right or wrong).

Table 10. Errors Caused by Imputation Algorithm, Persons Aged 70 and Over

IMPUTATION	**ERROR**		
	no	yes	**Total**
no	214 31.9%	**75** **11.2%**	289 43.1%
yes	278 41.4%	**104** **15.5%**	382 56.9%
Total	492 73.3%	**179** **26.7%**	671 100.0%

Table 11. Survey/Tax Comparison of CPP/QPP Reporting, Persons Aged 65 to 69 Years

SURVEY	**INCOME TAX**		
	CPP=0	CPP=amount	**Total**
CPP=0	102 10.5%	**99** **11.5%**	201 20.6%
CPP=dk or refusal	3 2.7%	**63** **7.3%**	66 6.8%
CPP=amount	6 0.6%	702 72.0%	1708 72.6%
Total	111 11.4%	864 88.6%	975 100.0%

From Table 11, it can be seen that 162 respondents in the 65 to 69 age group would be erroneously shown as not receiving CPP/QPP, representing 60.6% of all those in the survey with no amount for this source.

The editing approach was similar to that applied to persons 70 years and over. The impact is shown in Table 12. Essentially, the incorrect action was taken only 18.7% of the time.

Table 12. Errors Caused by Imputation Algorithm, Persons Aged 65 to 69 Years

IMPUTATION	**ERROR**		
	no	yes	**Total**
no	69 25.8%	**14** **5.2%**	83 31.1%
yes	148 55.4%	**36** **13.5%**	184 68.9%
Total	217 81.3%	**50** **18.7%**	267 100.0%

Table 13. Survey/Tax Comparison of CPP/QPP Reporting, Persons Aged 60 to 64 Years

SURVEY	**INCOME TAX**		
	CPP=0	CPP=amount	**Total**
CPP=0	270 34.4	**99** **12.6**	369 47.0
CPP=dk or refusal	4 0.5	**21** **2.7**	25 3.2
CPP=amount	7 0.9	385 49.0	392 49.9
Total	281 35.8	505 64.3	786 100.0

For persons aged 60 to 64, the number incorrectly identified as not receiving CPP/QPP would have been 120 without editing, that is 30.5% of all those with no amount reported in the survey. As mentioned earlier, the algorithm for determining whether or not to impute CPP/QPP for someone in this age group was somewhat more complex. Briefly, based on work experience, if an individual was still working at a job, CPP/QPP was not imputed. It was also not imputed for people who stopped working before 1967 (the introduction of the program), or had stopped working after 1966 because they had lost their job or were laid off then CPP/QPP was not imputed. If an individual had stopped working after 1966 for other reasons, then CPP/QPP was imputed.

Table 14. Errors Caused by Imputation Algorithm, Persons Aged 60 to 64 Years

IMPUTATION	**ERROR**		
	no	yes	**Total**
no	231 56.6	**41** **10.4**	272 69.0
yes	79 20.1	**43** **10.9**	122 31.0
Total	310 78.7	**84** **21.3**	394 100.0

Table 14 indicates that this approach misclassified 21% of the cases with no response in the survey.

3.3.5 "Specifies"

The first step of editing the specifies is to code them. Using SLID's automated coding program, the specifies are first coded automatically based on a dictionary file from the

Survey of Consumer Finances. If a code is not automatically assigned, the remainder are manually assigned a code. Once all the specifies are coded, the amounts are transferred from the "specifies" to the appropriate income category using a program based on SCF specifications. To carry out this step, the class of worker (COW) variable was needed. Specifications to derive the class of worker variable were created, so that this step could be carried out in advance of labour processing being completed (this was a short-term measure to allow labour and income processing to proceed in parallel for Wave 1. For Wave 2, this step of income processing will await completion of class of worker coding).

3.4 Study of Extreme Values

A definition of “extreme value” was developed based on the tax data of the survey’s respondents. Table 15 shows the ranges (after rounding) obtained for each income item.

Table 15. Income Ranges of Survey Population by Income Source (Tax Data)

Income source	**Observed range**	
	Minimum	Maximum
Wages and salaries	$0	$1,500,000
Net self-employment income -- farm	-$65,000	$90,000
Net self-employment income -- non-farm	-$1,050,000	$740,000
Investment (interest, dividends, other inv. income)	-$190,000	$300,000
Net capital gains	$0	$630,000
Child Tax Benefit	$0	$9, 600
OAS/GIS/SPA	$0	$19,000
CPP/QPP	$0	$23,000
Unemployment Insurance	$0	$23,000
Social assistance	$0	$25,000
Worker’s Compensation	$0	$59,000
Provincial tax credit	$0	$7,500
Empl. pensions, RRSP annuities, RRIF withdrawals	$0	$110,000
RRSP withdrawals	$0	$78,000
Other money income (taxable)	$0	$200,000

Income amounts collected by interview and falling outside the above ranges were examined. Sixteen such cases were found, affecting seven income categories. These outliers, which appear to be errors, are described in Table 16.

Table 16. Observations Collected in Income Interview that Fell Outside Ranges Determined by Tax Data

Income source	Nature/amount
Wages & salaries	- 1 negative amount (less than -$2,000) - 1 extreme positive amount (more than $3M)
Farm self-employment	- 4 amounts greater than $100,000
Net capital gains	- 4 negative amounts, ranging from -$900 to -$15,000
CTB	- 2 observations exceeding max on tax file, both above $10,000
CPP/QPP	- 2 observations exceeding maximum
UI	- 1 observation exceeding maximum
Provincial tax credit	- 1 observation exceeding maximum

For the Wave 1 public use file, the top three income amounts in each source were suppressed, thus removing most of these outliers. For Wave 2, the outlier detection approach will distinguish between extreme values that are clearly erroneous and ones that are extreme but valid. Edits will be performed on the first category.

3.5 Assigning Income Response Codes

The procedure for assigning person-level response codes depends in part on whether tax file data are used or not. If the income data come from the tax file, then the person-level response code is "complete" (code 000), unless any of the fields on the person's tax information contain a "1", which is a processing code rather than a true dollar amount. For these cases, the person-level response code is set to "partial" (code 001). There were 139 cases where "1" was recorded on at least one field. These cases were individually examined. In 85 cases, the code referred to wages and salaries. In 28 cases, it referred to investment income and in 14 to OAS. The rest were distributed over various categories. The "1" observations were deleted in processing.

In contrast, where the information was obtained by interview, the person-level response code was determined by the types of responses obtained in the interview. To this end, two variables were created: **RESP25** describes the response obtained to the questions on income components (Q1 to Q25) and **Q26** indicates the response obtained to the total income question (Q26).

RESP25 takes 1 of the following 5 values:

1	No	$0 or blank recorded for all income sources
2	Yes/No	amount recorded in at least 1 item + some items with $0 or blank
3	Yes/No/Don't know or Refusal	at least 1 amount + some items with $0 or blank + at least one item marked don't know or refusal
4	No/Don't know or Refusal	some items $0 or blank, the rest are marked don't know or refusal
5	Don't know or Refusal	all items marked don't know or refusal

Q26 takes one of the following 3 values:

1	No	$0 or blank for total income
2	Yes	an amount entered for total income
3	Don’t know or Refusal	don't know or refusal entered for total income

With the help of these two variables, a response code was defined for the person.

Complete response is defined as any of the following combinations:

	Q26	**RESP25 and response code from collection**
1	Yes	Yes/No
2	No	No & Response code=000 (complete)*
3	No	Yes/No
4	Don’t Know/Refusal	Yes/No

* Corresponds to a person with no income.

Total non-response is used for those who did not report any amount and are not people with “no income”. Total non-response can be any of the following combinations:

	Q26	**RESP25 and response code from collection**
1	No	No & Response code=non-response
2	Refusal	No
	Q26	**RESP25 and response code from collection**
3	Don’t know	No
4	No	Don’t know/refusal
5	Refusal	Don’t know/refusal
6	Don’t know	Don’t know/refusal
7	No	No/don’t know/refusal
8	Refusal	No/don’t know/refusal
9	Don’t know	No/don’t know/refusal
10	No	Yes/No & Resp. code=non-response
11	Refusal	Yes/No & Resp. code=non-response
12	Don’t know	Yes/No & Resp. code=non-response
13	No	Yes/No/R/DK & Resp. code=non-resp.
14	Refusal	Yes/No/R/DK & Resp. code=non-resp.
15	Don’t know	Yes/No/R/DK & Resp. code=non-resp.

All other possible combinations are considered as **partial response**.

3.6 Consistency Between the Class of Worker and Employment Income

A person may have reported up to six jobs during the labour interview. Depending on class of worker, there should be wages and salaries reported in the income interview and/or self-employment income. More specifically:

- If the class of worker is *paid worker* (COW=1) in at least one job: employment income should be reported in Q01 (wages and salaries).

- If the class of worker is *self-employed worker, incorporated* (COW=3 or 4) in at least one job: employment income should be reported in Q01 (wages and salaries) or in Q05 (dividends).

- If the only class of worker is *unpaid family worker* (COW=2), no employment income should be reported.

- If the class of worker is *self-employed worker, unincorporated* (COW=5 or 6), employment income should be reported in Q02 (net income from farm self-employment) or Q03 (net income from non-farm self-employment), depending on the industry code.

To carry out this step, we used the class of worker variable derived for transferring the "specifies" in Step 3.5. A number of rules were established for comparing class of worker from the labour interview and the income reported in the income interview (Table 17) . If there is an inconsistency between labour and income, depending on whether a person reported through survey or tax route, the inconsistency could be corrected by moving the income amount to another category or by changing the class of worker in the labour interview, or no changes may be made and the inconsistency flagged.

Table 17: Counts for Actions Taken to Resolve Inconsistencies Between Labour and Income Information

	Counts		
	Total	Survey	Tax
Total records	23,030	9,517	13,513
No inconsistency between labour and income	20,229	8,334	11,895
Some inconsistency, but it will not* be resolved	1,785	602	1,183
Inconsistency, to be resolved by changing at least 1 class of worker	602	167	435
Amount moved from farm to non-farm self-employment income or vice versa	116	116	0
Amount moved from non-farm self-employment income to wages and salaries	57	57	0
Amount to be imputed to wages and salaries, self-employment income or investment income during GEIS	241	241	0

* these records will be excluded from donor pool during GEIS imputation step.

** constitutes difference between edited and collected class of worker.

4. TAX FILE MATCHING OPERATION

When the Wave 1 income data were collected in May 1994, the idea of offering a “tax route” alternative to respondents had not yet been assessed for its feasibility. That assessment was completed early in 1994. The decision was taken to offer the tax route option to respondents at the time of the May 1995 interview (in later years, it would be offered in January, so that no May contact would be required in the case of full households electing to provide tax data).

In May 1995, permission was sought to use tax data for the full life of the panel. For respondents who agreed, and who were successfully matched to the tax file, the decision was taken to use tax file data for 1993 as well as for subsequent years. The reason was mainly to ensure that micro-level changes from 1993 to 1994 would not be affected by switching the data source.

In May 1995, about 62% of respondents agreed to the tax route. These were the people for whom tax data would be used, unless it was impossible to find a match on the tax file. In the latter case, income interview results could still be used. It should be noted that, when the tax file option was offered again in January 1996 (to new respondents and those who had refused the previous time), the proportion electing to go the tax route rose to nearly 73%.

4.1 Summary Description of Approach -- Direct or Statistical Match

SLID does not ask respondents for SIN because of the difficulties and potential sensitivity of collecting this information through a telephone interview. Instead, matching is done through a two-stage statistical procedure.

The first stage is a direct match to the tax file based on name, date of birth, sex, marital status and postal code. A match was found for 50% of the records using this approach. The second stage is a statistical matching process where a link is attempted, allowing missing values for one or more of the match criteria. An additional 35% of the records were matched this way.

4.2 Some Match Rates and Plans for Future Evaluations

Table 18 shows the total number of SLID respondents for whom tax data will not be used for 1993, for one of the following reasons:

- the person is a non-filer
- he/she did not want to take the “tax route” option
- he/she granted permission for STC to use tax data but we could not find a match on the tax file.

The response status indicates whether household response is complete (i.e., a complete interview was obtained for every eligible household member), partial or non-response. Each row refers to a pair of response codes: the first one is the response code for the January labour interview and the second one refers to the May income interview. For example, “complete/complete” means complete interviews were obtained for all household members in both January and May.

In Table 18, the cases requiring imputation are flagged. Imputation was done through the Generalized Edit and Imputation System (GEIS). Wherever possible, a tax route respondent will be selected as a donor. Table 19 provides comparable information for the respondents whose income data for 1993 will come from the tax file.

Some Highlights

- The number of persons for whom tax data was used for Wave 1 was just under 15,000. This was about one-half of all respondents aged 16 and over.

- In Table 18, the most common pattern, apart from total response, is "complete/non-response". In other words, complete interviews were obtained for labour, but no information at all was obtained for income (3,311 cases).

Table 18. Response Status : Persons for Whom Survey Data Was Used in Wave 1

Response status Jan/May	Total	Longitudinal respondents	Cohabitants
Total	15,055	14,180	875
Complete/complete	9,446	9,072	374
Complete/partial	842	787*	55*
Partial/complete	103	101	2
Partial/partial	1	1*	0
Complete/non-response	3,311	3,092**	219**
Partial/non-response	31	31**	0
Non-response/complete	770	751	19
Non-response/partial	62	54*	8*
Non-response/non-response	489	291**	198**

* Partial imputation

** Total imputation

Table 19. Response Status: Persons for Whom Tax Data Was Used (Access Granted and Match to Tax File Successful)

Response status Jan/May	Total	Longitudinal respondents	Cohabitants
Total	14,627	14,082	545
Complete/complete	13,318	12,868	450
Complete/partial	125	122	3
Partial/complete	14	13	1
Partial/partial	1	1	0
Complete/non-response	0	0	0
Partial/non-response	0	0	0
Non-response/complete	1,161	1,071	90
Non-response/partial	8	7	1
Non-response/non-response	0	0	0

The following tables show where imputation was required for various income sources, according to whether or not the respondents took the "tax route" or not. The first of each pair of tables refers to people for whom survey (interview) data was used rather than tax data. In addition to individuals who refused access to their tax file data, this population includes non-filers, and non-matches.

Some Explanation of Terminology

January flag means that, based on the January labour interview, income should be reported for the source in question. In other words, we were expecting the respondent to report wages and salaries in the May interview because he or she had reported at least one paid worker job in January. There was in fact a "flag" set and, if no income was reported in May, the flag triggered a probe question in an attempt to resolve the inconsistency. The cases below are those where a inconsistency remains. In the tables for respondents who chose the tax route (and where the link to the tax file was successful) the “flag” simply indicates whether a particular income source should be reported, based on the labour interview.

When there is an inconsistency between the labour and income data, there may be no action taken. Where an action is taken, it is either to impute an amount (indicated by * in the tables) or to change the flag (** in the tables).

Table 20. Reporting of Wages and Salaries: Survey Route Respondents

Response status Jan/May	Total	No Jan flag & no May amount	Jan flag but no May amount	Jan non-resp.	No Jan flag but May amount	Jan flag & May amount	Other
Total	15055	5869	2207	765	556	5196	462
Comp/comp	9446	3903	169*	2	475**	4897	0
Comp/part	842	359	160	0	29**	294	0
Part/comp	103	63	0	0	37**	3	0
Part/part	1	1	0	0	0	0	0
Comp/non	3311	1432	1876	1	0	2	0
Part/non	31	28	2	1	0	0	0
Non/comp	770	7	0	302	15**	0	446
Non/part	62	0	0	47	0	0	15
Non/non	489	76	0	412	0	0	1

Table 21. Reporting of Wages and Salaries: Tax Route Respondents

Response status Jan/May	Total	No Jan flag & no May amount	Jan flag but no May amount	Jan non-resp.	No Jan flag but May amount	Jan flag & May amount	Other
Total	14627	4122	263	355	728	8370	789
Comp/comp	13318	4028	253	1	692**	8340	4
Comp/part	125	86	10	0	2**	27	0
Part/comp	14	3	0	0	8**	3	0
Part/part	1	0	0	0	1**	0	0
Comp/non	0	0	0	0	0	0	0
Part/non	0	0	0	0	0	0	0
Non/comp	1161	5	0	347	25	0	784
Non/part	8	0	0	7	0	0	1
Non/non	0	0	0	0	0	0	0

* Impute amount

** Change flag

The top row of Table 21 shows the analogous information for respondents who chose the tax route. Under 2% reported a wage and salary job in January, but had no wages and salaries in their income information. However, 728 (5%) had not mention a wage and salary job in the labour interview, but had this type of information on their tax return.

Thus labour/income discrepancies in the reporting of wages and salaries were less likely to arise where tax data were used (7%, versus 18% for survey data).

The top row of Table 20 indicates that, for respondents taking the survey route, there was 2,207 cases (15%) where wages and salaries were expected (based on jobs identified in the January labour interview), but none were reported. There were 556 cases (4%) where the reverse was true.

Tables 22 and 23 present a similar comparison to wages and salaries, but this one deals with Workers' Compensation. Among survey route respondents, 140 respondents reported receiving Workers' Compensation in both the labour and income interview. An additional 178 reported in one but not the other. Thus, for 2% of the survey route respondents, there was some indication that Workers' Compensation was received.

Among the tax route respondents, this proportion was 4%. This includes 305 cases where the labour interview and tax information are consistent; 212 cases where Workers' Compensation was received based on the tax return but was not reported in the labour interview; and, 43 cases where the reverse was true.

Table 22. Reporting of Workers' Compensation: Survey Route Respondents

Response	Total	No Jan	Jan flag but	Jan	No Jan flag	Jan flag &	Other
Total	15,055	13,509	115	1,216	63	140	12
Comp/comp	9,446	9,226	25*	2	62**	131	0
Comp/part	842	806	26	0	1**	9	0
Part/comp	103	103	0	0	0	0	0
Part/part	1	1	0	0	0	0	0
Comp/non	3,311	3,246	64	1	0	0	0
Part/non	31	30	0	1	0	0	0
Non/comp	770	21	0	737	0	0	12
Non/part	62	0	0	62	0	0	0
Non/non	489	76	0	413	0	0	0

Table 23. Reporting of Worker's Compensation: Tax Route Respondents

Response	Total	No Jan	Jan flag but	Jan	No Jan flag	Jan flag &	Other
Total	14,627	12,923	43	1,106	212	305	38
Comp/comp	13,318	12,755	42	5	211**	305	0
Comp/part	125	124	1	0	0	0	0
Part/comp	14	14	0	0	0	0	0
Part/part	1	1	0	0	0	0	0
Comp/non	0	0	0	0	0	0	0
Part/non	0	0	0	0	0	0	0
Non/comp	1,161	29	0	1,093	1	0	38
Non/part	8	0	0	8	0	0	0
Non/non	0	0	0	0	0	0	0

* Impute amount

** Change flag

Table 24. Reporting of Unemployment Insurance: Survey Route Respondents

Response	Total	No Jan	Jan flag but	Jan non-	No Jan flag	Jan flag &	Other
Total	15,055	11,893	683	1,152	151	1,100	76
Comp/comp	9,446	8,147	105*	2	143**	1,049	0
Comp/part	842	648	137	0	6**	51	0
Part/comp	103	103	0	0	0	0	0
Part/part	1	1	0	0	0	0	0
Comp/non	3,311	2,869	441	1	0	0	0
Part/non	31	30	0	1	0	0	0
Non/comp	770	19	0	677	2**	0	72
Non/part	62	0	0	58	0	0	4
Non/non	489	76	0	413	0	0	0

Table 25. Reporting of Unemployment Insurance: Tax Route Respondents

Response status Jan/May	Total	No Jan flag & no May amount	Jan flag but no May amount	Jan non-resp.	No Jan flag but May amount	Jan flag & May amount	Other
Total	14,627	10,755	54	889	402	2,273	254
Comp/comp	13,318	10,604	53	5	391**	2,265	0
Comp/part	125	115	1	0	1**	8	0
Part/comp	14	13	0	0	1**	0	0
Part/part	1	1	0	0	0	0	0
Comp/non	0	0	0	0	0	0	0
Part/non	0	0	0	0	0	0	0
Non/comp	1,161	22	0	876	9**	0	254
Non/part	8	0	0	8	0	0	0
Non/non	0	0	0	0	0	0	0

* Impute amount

** Change flag

In the case of Unemployment Insurance, 1,100 survey route respondents (7%) reported UI in both the labour and income interview; 683 (5%) reported it in the labour interview but not the income interview and 151 (1%) had the reverse pattern (Table 24). Among the tax route respondents, reporting discrepancies were identified for 3% of the cases (Table 25). The number reporting UI in both the labour interview and tax return was 2,273 (16%).

5. IMPUTATION AND DERIVATION PROCEDURES

Imputation of some missing income values in Wave 1 was done using the Generalized Edit and Imputation System (GEIS). Where the value could be readily calculated based on known parameters, deterministic imputation was used. In broad terms, the strategy is as follows:

- *Imputation by nearest neighbour for most income variables:* using GEIS, the same neighbour will be used to impute all the missing variables, so as to preserve the relationships among the various components.

- *Deterministic imputation for the following variables:* Child Tax Benefit, Goods and Services Tax Credit, Old Age Security, Guaranteed Income Supplement and Spouse's Allowance.
- *Linear regression* was used to model federal and provincial taxes payable for records other than those obtained by the tax file.

5.1 Generalized Edit and Imputation System (GEIS)

The Generalized Edit and Imputation System (GEIS) principally consists of verification, error localization and imputation. The objective of verification is to determine if certain values of a specific record are incorrect, missing, not coherent or outliers. Imputation replaces an erroneous value by a more plausible one. These two functions are linked together by the error localization function, which determines the fields of a given record that need imputation.

5.2 Nearest Neighbour Imputation Using GEIS

For imputation by nearest neighbour, imputation was done in two stages: partial imputation and total imputation. The reason is that, for partial imputation, specific fields are flagged -- that is, we assume that income was received from the source in question -- and we need a donor with income reported for the flagged source. On the other hand, for imputation for total non-response, we do not know the sources that should be imputed. We therefore have less information for the imputation, and we impute both sources and amounts from the donor. This means that if the donor received investment income and pension income, we attribute them to the non-respondent.

All individuals 16 and over eligible for the income interview were divided according to precise rules into 4 groups: donor, partial non-response recipient, total non-response recipient and excluded from imputation. The fourth category consist of records where the data did not require intervention but where there were some grounds for suspecting inconsistent reporting. Thus, only very clean records were used for imputation.

Table 26. Imputation Status of Respondents Eligible for Income Interview

	Donor	Recipient		Excluded from imputation	**Total**
		partial non-resp.	total non-response		
Survey data	8,046 *39.8%*	1,661	117	1,390	11,214
Revenue Canada data	12,160 *60.2%*	0	0	2,467	14,627
No data	0	0	3,829	0	3,829
Total	20,206 *100%*	1,661	3,946	3,857	29,670

5.2.1 Imputation of “Recipients: Partial Non-Response”.

This step imputed missing amounts to the partially responding cases. To select an appropriate donor, the donor pool was divided into the following 8 matching groups:

Males 16-24	Females 16-24
Males 25-54	Females 25-54
Males 55-69	Females 55-69
Males 70+	Females 70+

The discrete matching variables (*edit groups* in GEIS) were as follows:
Region (Atlantic, Quebec, Ontario, Prairies and British Columbia)
Marital status (married or common-law, divorced or separated, widowed and single)
Urban or rural
Level of education (less than secondary, secondary to less than university, university degree or higher)
Full or part-time student
Number of children aged 0-15
Number of children 16-18
Number of adults
Received Unemployment Insurance
Received Workers’ Compensation
Received Social Assistance
CPP/QPP Value Expected
Paid worker flag
Farm self-employed (incorporated) flag
Farm self-employed (non-incorporated) flag
Unpaid worker flag

Employed or not

Farm employment or not

Spouse with welfare income or not

Note that the variables of the work component are not available for the two 70 and over groups.

The continuous matching variables (*must match* in GEIS) were:

Number of jobs held during year

Number of weeks unemployed

Total usual hours worked at all jobs

Total annual wages and salaries

All income variables not to be imputed

Note that the variables of the work component are not available for the two 70 and over groups. After imputation with GEIS, there remained 63 cases that had to be imputed manually.

5.2.2 Imputation of "Recipients: Total Non-Response"

For these recipients, all income variables required imputation. The matching groups and the discrete matching variables were the same as those used for partial imputation. For the continuous matching variables, however, we used only the following:

Number of jobs held during year

Number of weeks unemployed

Total usual hours worked at all jobs

Total annual wages and salaries

Total income from income interview (where available)

After imputation using GEIS, 94 cases still had to be imputed manually.

5.2.3 Deterministic Imputation

For all survey route respondents, deterministic imputation was used for the following variables:
Child Tax Benefit, OAS/GIS/SPA and GST Credit. In the case of tax route respondents, deterministic imputation was used for GST credits.

Deterministic imputation was performed after imputation by the nearest-neighbour method. Values were calculated based on characteristics of respondents and program parameters. Where appropriate, we compared the calculated value with the amount reported by the respondent. If the difference between the two amounts was too large (determined using certain criteria), we used the amount calculated; if not, we kept the reported amount, trying to keep what the respondent reported as much as possible.

The reason for calculating GST credits for tax route respondents is that this amount is not on the tax return. For the same reason, we calculated family allowances for Quebec respondents.

5.2.4 Imputation of Federal and Provincial Income Tax

Federal and provincial taxes payable were modelled using linear regression. This stage followed deterministic imputation. For all records (except where data was obtained from tax information), federal and provincial tax amounts were calculated. For Quebec, provincial income tax was calculated for all records since provincial income tax for Quebec is not available in the tax file.

5.3 Assessment of Nearest-Neighbour Imputation Using GEIS

As one measure of the quality of imputation one can ask: how often was the same donor chosen?

Table 27 shows that 81% of all records picked as donors were selected only once. It also shows that a majority of donors (59%) were selected from tax respondents. This proportion mirrors the actual composition of the donor pool: as shown in Table 27 over 12,000 of the 20,000-odd donors, or 60%, were tax route respondents. A small minority of donors were used many times, which is attributable to the fact that the donor pool itself was quite small in relation to the number of cases requiring imputation. This situation is expected to improve for the 1996 reference year, when the second panel begins, doubling the pool.

Table 27. Source of Donor (Survey or Tax Route Respondent) and Number of Times Specific Donors Were Selected

Source of Data	Number of Times the Same Donor Was Chosen						
	1	2	3	4	5	6+	Total
Survey	1,425 *79.5%*	265 *14.8%*	60 *3.3%*	27 *1.5%*	8 *0.4%*	7 *0.4%*	**1,792** ***41.0%***
Tax	2,096 *81.3%*	383 *14.9%*	65 *2.5%*	17 *0.7%*	7 *0.3%*	10 *0.4%*	**2,578** ***59.0%***
Total	**3,521** ***80.6%***	**648** ***14.8%***	**125** ***2.9%***	**44** ***1.0%***	**15** ***0.3%***	**17** ***0.4%***	**4,370** ***100%***

A number of other steps were taken to evaluate the quality of the imputation. Among them was a comparison of means, medians and ranges for the imputed population, relative to tax data for the same population.

Evaluation of Total Imputation

Among persons aged 16-69, the comparison between tax and imputed data was based on the 2,656 observations that matched to the tax file (Table 28).

There are a few explanatory notes on the table. For each income variable shown in the table, there is one row showing results from the imputation and one row showing the true (tax file) results. The fourth column shows the number of records where the amount imputed was different from zero. The next two columns show the means and median, excluding records where the imputed and true value were both equal to zero.

With the exception of a few variables, the imputation results appear quite reasonable. The variables where the results are somewhat off are those where we experienced difficulty finding a suitable donor. For example, it is hard to impute large negative values for farm self-employment income, and thus the mean imputed amount is quite different (greater) than the mean from the tax data. Results similar to those in Table 28 have also been calculated by 5-year age group and for persons aged 70 and over.

Table 28. Comparison of Mean and Median Income for Components, Imputed Population and Corresponding Tax File Population: Total Imputation, Persons Aged 16-69

	Value	N	Amount≠0	Mean$	Median$	Range
Wages & salaries	tax	2,600	1,923	23,800	17,100	0 to 1,460,000
	imputed	2,600	1,922	23,300	20,200	0 to 175,000
Self-employment (farm)	tax	126	114	4,800	1,300	-25,000 to 60,000
	imputed	126	62	7,400	0	-9,000 to 60,000
Self-employment (non-farm)	tax	373	307	11,600	2,500	-65,000 to 230,000
	imputed	373	206	10,600	0	-90,000 to 300,000
Investment	tax	1,446	997	2,200	100	-85,000 to 155,000
	imputed	1,446	889	1,600	200	-30,000 to 310,000
Capital gains	tax	228	176	5,400	100	0 to 180,000
	imputed	228	124	6,200	100	0 to 260,000
CPP/QPP	tax	323	307	4,600	0	0 to 15,000
	imputed	323	210	3,700	4,700	0 to 45,000
UI	tax	608	587	5,000	3,200	0 to 21,000
	imputed	608	432	3,700	2,400	0 to 19,000
Social assistance	tax	204	168	4,700	3,300	0 to 21,000
	imputed	204	120	3,300	1,400	0 to 18,000
Workers' Compensation	tax	110	106	4,300	1,600	0 to 33,000
	imputed	110	54	2,100	0	0 to 42,000
Employer pension	tax	286	187	7,600	2,800	0 to 76,000
	imputed	286	177	7,900	3,300	0 to 57,000
Alimony	tax	80	41	3,800	300	0 to 59,000
	imputed	80	45	2,700	1,400	0 to 18,000
Other taxable income	tax	649	448	2,500	400	0 to 77,000
	imputed	649	272	1,400	0	0 to 62,000

Evaluation of Partial Imputation

The results for cases requiring partial imputation are given in Table 29, again for persons aged 16 to 69. For wages and salaries, the mean of the imputed amounts is greater than that of tax data, but not enough to be alarming. It can also be seen that the tax maximum was higher than for imputed data. For farm self-employment income, as noted earlier, it is hard to impute small or negative amounts; the mean of imputed values ended up being greater than that for tax data. For the variable non-farm self-employment income, the mean, minimum and maximum amounts imputed were all lower than the tax data. Investment income was underestimated slightly, while the mean for capital gains was overestimated, mainly due to the imputation of a few large amounts. Conversely, over-estimation of CPP/QPP was probably caused by the “flag” from the LFS information, as explained earlier in this report. For social assistance, it is difficult to assess the results obtained at the individual level, since this variable should be examined at the family level. Results were satisfactory for the other variables.

Additional information from the evaluation of the partial imputation step can be obtained on request.

Table 29. Comparison of Mean and Median Income for Components, Imputed Population and Corresponding Tax File Population: Partial Imputation, Persons Aged 16-69

	Value	N	Amount≠0	Mean$	Median$	Range
Wages & salaries	tax	268	113	10,500	4,600	0 to 86,000
	imputed	268	139	14,100	9,200	0 to 52,600
Self-employment (farm)	tax	478	25	-300	-300	-11,000 to 20,000
	imputed	478	13	-1,900	0	-7,000 to 26,600
Self-employment (non-farm)	tax	465	35	2,600	0	-11,200 to 36,600
	imputed	465	30	800	0	-33,300 to 18,200
Investment	tax	478	155	1,700	100	-4,200 to 59,000
	imputed	478	150	1,200	100	-2,700 to 33,700
Capital gains	tax	492	34	7,500	0	0 to 145,100
	imputed	492	28	14,400	50	0 to 260,000
CPP/QPP	tax	101	56	2,900	500	0 to 9,610
	imputed	101	101	4,800	5,000	100 to 11,234
UI	tax	222	158	4,500	3,500	0 to 16,500
	imputed	222	173	4,800	3,700	0 to 19,000
Social assistance	tax	107	34	2,900	800	0 to 18,000
	imputed	107	62	4,300	2,600	0 to 19,500
Workers' Compensation	tax	88	22	3,500	700	0 to 31,000
	imputed	88	37	4,100	1,400	0 to 24,000
Employer pension	tax	478	28	5,600	300	0 to 108,600
	imputed	478	38	6,600	4,300	0 to 30,100
Alimony	tax	479	5	2,000	100	0 to 8,200
	imputed	479	5	2,000	100	0 to 9,400
Other taxable income	tax	485	92	1,600	200	0 to 32,100
	imputed	485	61	900	0	0 to 30,800

5.4 Total Income

Once GEIS imputation for total and partial non-respondents was finished, the next step in processing was to calculate a final total income. This involved a number of steps. To begin with, before calculating total income, certain input components to this total had to be deterministically imputed (i.e., Child Tax Benefit, Old Age Security/Guaranteed Income Supplement and Goods and Services Tax Credit). In order to perform each of these imputation steps, a total income pertaining to that particular step had to be calculated. Table 30, below, outlines the different totals and the steps they pertain to.

The first total income created was a total income for the calculation of Old Age Security/Guaranteed Income Supplement/Spouses Allowance. This allowed for the imputation of the OAS/GIS component for those who were eligible and should have received it (based on standard program qualifications), but who did not report it. Once OAS/GIS was assigned (only to eligible survey route respondents), two other total incomes, for the calculation of GST/CTB and federal/provincial taxes, were computed (OAS was necessary as an input for both of these total incomes). These two totals allowed for the imputation of these components. CTB and federal/provincial taxes were imputed for eligible survey route respondents only, while GST was imputed for all respondents, as tax data was not available for this component. At this point, imputation for all income components was complete on the file and, as a result, a total money income could be derived.

Table 30: Summary of SLID Total Incomes

Total Incomes	Exclusions
Total Income Used to Calculate Eligibility for Old Age Security/Guaranteed Income Supplement/Spouses Allowance	Excludes non-taxable sources of income, such as, Child Tax Benefit, GST credit, social assistance payments
Income Tax Concept of Total Income Used to Calculate Eligibility for GST credit and Child Tax Benefit	Equivalent to line 150 on tax return - does not include items such as Child Tax Benefit and GST credit, which are not reported on the tax form
Total Income Used for Calculation of Federal and Provincial Taxes	Excludes non-taxable sources of income such as, Child Tax Benefit, GST credit, social assistance payments, Guaranteed Income Supplement/Spouses Allowance (but includes Old Age Security)
Total Money Income	Excludes non-money income sources such as capital gains and RRSP withdrawals

6. FAMILY INCOME AND LOW INCOME

Once individual income (components and total) were finalised, the next step was to derive family-level income variables. In SLID, “family” means economic family -- all persons related by blood, marriage, adoption or common-law relationship living in the same dwelling on December 31, 1993. Note that it does not matter whether or not a given individual was part of the family all year; it is sufficient that he or she was in that family at year end. Thus, each individual was assigned a family income amount, representing the sum of the annual incomes of all individuals in that family.

For low income, a low-income cutoff (LICO) was assigned to each person, based on the size of his or her family and the size of their community. The LICOs are the same ones used by the Survey of Consumer Finances and were based on expenditure data from the 1992 Family Expenditure Survey. Average family expenditures on food, shelter and clothing are the basis of these LICOs. The lines represent the income level where the average family of a given size and in a given type of community spends 54.7% of its income or more on these essentials.

There is a different LICO for unattached individuals and for families of 2 through 7+. Similarly, the LICO varies according to whether the family is living in a rural area, small urban area (under 30,000 population), 30,000-99,999, 100,000-499,999 or 500,000+. As a result there are 35 LICOs. The file contains the appropriate LICO for the person’s family size and community size. All members of the same family have the same LICO.

There are some differences between SLID and SCF in how the variable *size of area of residence* (size of community) was derived, which had some impact on the low-income estimates. The derivation procedure is being modified in Wave 2, and will be described in a future ILD Working Paper.