Data Quality for the 2007 Survey of Labour and Income Dynamics (SLID)

Jean-François Bastien

Household Survey Methods Division
R. H. Coats Building, Ottawa, K1A 0T6

Table of contents

1. Introduction

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated to produce estimates from 1993 onwards. The survey was designed to measure changes in the economic well-being of Canadians as well as the factors affecting these changes. The target population consists of all persons living in Canada with the following exclusions: persons living in Yukon, the Northwest Territories, and Nunavut, persons living on Reserves, persons living in institutions, and military personnel living in barracks.

The SLID sample is comprised of two panels. Each panel remains in the survey for six consecutive years and a new panel is rotated in every three years. In January following the reference year, SLID sample households are interviewed by telephone. Demographic information is collected for every person in the household while income, education and labour data are collected for every person in the household 16 years or older.

Before reference year 2004, respondents could be contacted for a January interview and a May interview. The May interview was to collect income data for respondents who did not agree to give us permission to link to the income tax records. From 2004 onwards, however, we dropped the May interview to save on collection costs. If a respondent does not grant permission to link to the T1 tax file, we ask them the income questions in January.

Although originally designed as a longitudinal survey, SLID has always maintained the capability of producing cross-sectional estimates. This cross-sectional aspect took on new importance with the cancellation of the Survey of Consumer Finance after the 1997 reference year. At this time SLID became the primary source of cross-sectional household and family income data.

All persons who are members of selected SLID households in the beginning of the first year of a panel's existence are longitudinal sample persons for SLID. As such, it is these individuals that are followed longitudinally. Any (non-longitudinal) person living in a household with a longitudinal person is referred to as a cohabitant. Cohabitants living with cross-sectionally eligible longitudinal persons will also be part of the cross-sectional sample.

For more information about survey concepts, definitions and design please refer to Statistics Canada publication: "*Survey of Labour and Income Dynamics - A survey overview*",

Sample surveys are subject to errors. As with all surveys conducted at Statistics Canada, considerable time and effort is taken to control such errors at every stage of the Survey of Labour and Income Dynamics. Nonetheless errors do occur. It is the policy at Statistics Canada to provide users with measures of data quality so that the user can interpret the data properly. This report summarizes these quality measures for SLID.

The following table shows highlights of data quality indicators for Canada for reference year 2007.

Table 1.1 Main SLID quality indicators for Canada in 2007

Indicator	Statistic
(Individual) Longitudinal sample size	
• Panel 4	29,088
• Panel 5	32,703
Cross-sectional sample size (eligible longitudinal individuals and cohabitants)	
• Panel 4	32,519
• Panel 5	35,789
Coefficient of variation	
• Median total income	0.6%
Slippage rate - person	
• Panel 4	17.0%
• Panel 5	12.3%
Slippage rate - household	
• Panel 4	16.3%
• Panel 5	12.1%
Response rate	
• Cross-sectional - person	69.6%
• Cross-sectional - household	71.8%
• Longitudinal - person	
○ Panel 4	68.9%
○ Panel 5	77.3%
Permission rate	
• Panel 4	91.3%
• Panel 5	85.5%
Tax linkage rate (SIN found)	94.9%
Imputation rate - person	
• Total imputation	2.5%
• Partial imputation	22.4%
Imputation rate - household	
• Partial imputation	38.9%

2. Sample composition/attrition

As mentioned, although originally designed as a longitudinal survey, one can also produce cross-sectional estimates from SLID data. Every non-longitudinal person living with a longitudinal respondent becomes part of the cross-sectional sample and is called a cohabitant. Table 2.1 and 2.2 show the composition of the SLID sample by province and by census metropolitan area (CMA) respectively, in terms of longitudinal sample persons who respond, longitudinal responding persons who are cross-sectionally ineligible (e.g.

deceased or institutionalized persons and those who have moved outside of Canada) and responding cohabitants.

The cross-sectional SLID sample coverage is maintained through the addition of cohabitants each year. The one exception is immigrants who arrive after the beginning of a panel and before the start of the next one and move into their own households; this introduces a small amount of under coverage. The longitudinal sample, however, is subject to attrition. Attrition is the gradual loss of respondents each year through the life of the panel. Table 2.3 shows the respondent status for persons originally selected as longitudinal respondents. In table 2.3 the responding longitudinal sample size is comprised of the in-scope respondents, the individuals who have moved to Yukon, North-West Territories or Nunavut, the individuals who have moved outside Canada, the institutionalized individuals and the deceased individuals.

Table 2.1 Sample composition of SLID by province in 2007

	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally[1]		Cohabitants		Cross-sectional sample size	
Province	**Panel 4**	**Panel 5**	**Panel 4**	**Panel 5**	**Panel 4**	**Panel 5**	**Panel 4**	**Panel 5**
Newfoundland	1,218	1,462	96	44	182	165	1,304	1,583
Prince Edward Island	816	899	55	24	158	116	919	991
Nova Scotia	1,971	1,963	146	66	376	276	2,201	2,173
New Brunswick	1,672	1,874	126	59	355	227	1,901	2,042
Quebec	5,557	6,002	392	181	1,209	916	6,374	6,737
Ontario	8,331	9,232	567	362	1,561	1,214	9,325	10,084
Manitoba	1,983	2,246	161	94	401	289	2,223	2,441
Saskatchewan	2,033	2,405	168	105	402	329	2,267	2,629
Alberta	2,510	3,315	143	102	641	575	3,008	3,788
British Columbia	2,663	3,008	192	109	526	422	2,997	3,321
Not in a province	334	297	0	0	0	0	0	0
Total	**29,088**	**32,703**	**2,046**	**1,146**	**5,811**	**4,529**	**32,519**	**35,789**

0 True zero or a value rounded to zero

1. This includes individuals who are deceased, institutionalized and those who have moved outside the country.

First, with respect to the longitudinal sample size, one can see a difference of 3,600 persons between panel 4 and 5 for Canada. All the provinces had a bigger sample size in panel 5 except Nova Scotia where a small decrease was observed between the two panels. The situation is similar for the cross-sectional sample size.

In all provinces, the number of panel 5 longitudinal persons ineligible cross-sectionally is roughly half that of panel 4. Finally, concerning the cohabitants in Canada, the numbers went from 5,811 to 4,539 between the two panels, a difference of 22%.

Table 2.2 Sample composition of SLID by CMA in 2007

	Longitudinal sample size		Cohabitants		Cross-sectional sample size	
Census Metropolitan Area	**Panel 4**	**Panel 5**	**Panel 4**	**Panel 5**	**Panel 4**	**Panel 5**
Halifax	427	586	90	91	517	677
Quebec City	401	447	141	95	542	542
Montréal	1,100	1,197	267	224	1,367	1,421
Ottawa - Gatineau	753	826	168	123	921	949
Toronto	1,322	1,650	313	241	1,635	1,891
Hamilton	365	432	70	39	435	471
St. Catharines - Niagara	389	364	68	56	457	420
Kitchener	397	427	71	65	468	492
London	360	485	85	94	445	579
Windsor	250	328	48	30	298	358
Winnipeg	857	1,093	207	151	1,064	1,244
Calgary	552	699	177	132	729	831
Edmonton	555	978	157	174	712	1,152
Vancouver	867	1,040	179	153	1,046	1,193
Victoria	232	281	39	49	271	330
Other CMA or CA	9,751	10,950	2,049	1,664	11,800	12,614
Do not live in a CMA	8,130	9,477	1,682	1,148	9,812	10,625
Not available[1]	2,380	1,443	0	0	0	0
Total	**29,088**	**32,703**	**5,811**	**4,529**	**32,519**	**35,789**

0 True zero or a value rounded to zero

1. This information is only available for those individuals who are cross-sectionally eligible

If we compare the longitudinal sample size by CMA, similar to the Canadian increase as explained in the previous paragraph, it was bigger in panel 5 for all the CMA except St. Catharines – Niagara where there's been a small decrease. For the cross-sectional sample size, the differences between the two panels are similar to those observed in the longitudinal sample size with also a reduction for St. Catharines – Niagara. For all the other CMA, the sample size was bigger except Quebec where it remained stable.

The number of cohabitants dropped for all the CMA in panel 5 except Halifax where it was similar and London, Edmonton and Victoria where it increased.

Table 2.3 Person status for the longitudinal sample in 2007

Person status for the longitudinal sample	Panel 4	Panel 5
In scope (respondents)	26,708	31,260
In scope (non-respondents)	3,248	6,839
Moved to Yukon, NWT, Nunavut	9	7
Moved outside Canada	323	288
Institutionalized	657	398
Deceased	1,391	750
Removed from sample[1]	9,859	2,773
Duplicate person/error[2]	37	15
Total	**42,232**	**42,330**

1. Respondents are removed from the sample for one of two reasons. If entire households have refused for two consecutive cycles they are said to be hard refusals and no further attempts are made to enumerate these households. Similarly, if, after two years, we cannot successfully trace households, we no longer pursue them.
2. Respondents who were erroneously included in the household in the beginning of the first year of a panel's existence.

While the total number of persons in panel 4 and 5 was very similar, one can notice major differences between the two panels when looking at the longitudinal status. First of all, the number of in scope respondents and in scope non-respondents was much bigger in panel 5. This is not a surprise since the 5th panel was only in its 3rd wave while the 4th panel was in its 6th and final wave. For the same reasons, the number of people in the category "removed from sample" is much bigger in panel 4 because after six waves, there were many households that couldn't traced and several which were considered as hard refusals.

3. Sampling errors

Sampling errors occur because inferences about the survey population are based on data from a sample of that population rather than the entire population. The sample design, the variability of the characteristic being measured, and the sample size will all contribute to the magnitude of the sampling error.

The standard error is a common measure of sampling error. The standard error measures the degree of variation introduced in estimates by selecting one particular sample rather than another of the same size and design. Another widely used measure of the sampling error is the coefficient of variation (CV), which is the estimated standard error expressed as a percentage of the estimate.

In SLID, the bootstrap approach is used for the calculation of standard errors. This is a resampling method of variance estimation, often used when dealing with estimates from a complex sample design. Table 3.1 shows CV levels at the provincial and national level for a sample of key SLID estimates.

Table 3.1 National and provincial coefficients of variation for certain variables in 2007 (%)

Variable (at the family level unless otherwise stated)	**N.L.**	**P.E.I.**	**N.S.**	**N.B.**	**Que.**	**Ont.**	**Man.**	**Sask.**	**Alta.**	**B.C.**	**Canada**
Median total income	3.0	2.2	2.0	2.2	1.3	1.2	1.7	1.4	1.5	1.7	0.6
Median market income	3.5	3.3	2.4	2.7	1.6	1.2	2.5	3.1	1.9	2.3	0.8
Median wages and salaries	3.1	4.5	2.6	3.1	1.6	1.0	2.5	2.1	1.4	2.8	0.7
Median EI benefits	6.3	4.9	7.1	6.0	3.9	6.4	9.5	8.3	15.1	12.4	2.9
Median social assistance	15.0	14.2	5.8	7.3	6.7	5.0	19.1	12.8	5.8	12.6	4.0
Median other income	12.5	22.7	18.4	17.6	9.3	6.6	15.6	15.7	12.1	10.4	4.4
Number under LICO after tax	9.3	14.8	7.7	7.6	3.9	3.9	6.6	6.7	7.6	5.9	2.1
Counts of employed people	1.8	2.0	1.5	1.4	1.1	1.0	1.9	1.6	1.5	1.7	0.5

For the median total income and the median market income, one can see that Newfoundland showed the highest CV. However, Prince Edward Island recorded the biggest CV for the median wages and salaries, the median other income, the number under LICO after tax and the counts of employed people. Those CV were higher because the lowest sample sizes were found in those two provinces. We noticed two exceptions for the median EI benefits and the median social assistance where the biggest CV was found in Alberta and Manitoba respectively.

4. Coverage errors

To produce good survey estimates, it is necessary that a survey sample adequately represent the survey population. To ensure proper coverage, SLID weights are adjusted using census population projections as control totals. The slippage rate is a measure of the percentage difference between these census projections and the survey estimate using weights prior to the application of this slippage related adjustment. More precisely, slippage is computed as

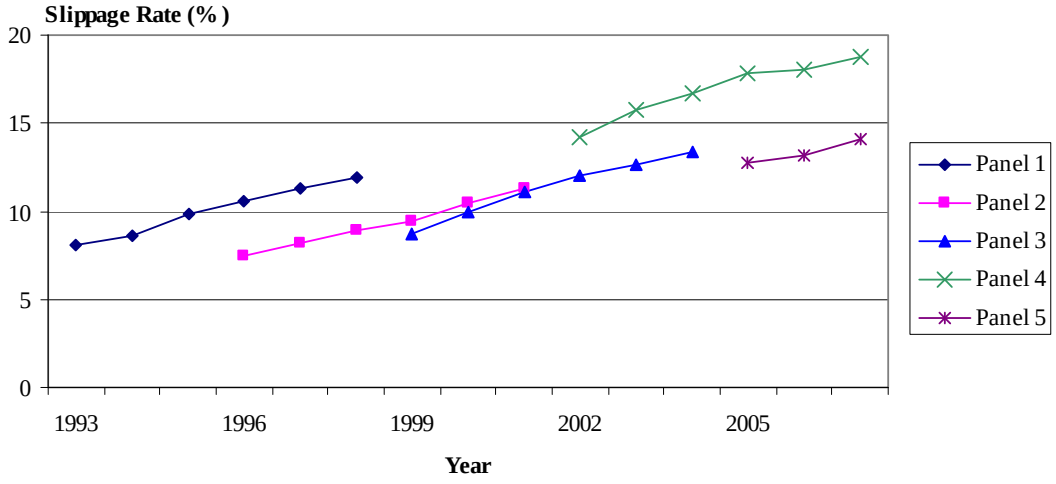
$$slippage_c = \frac{\left(CP_c - \sum_{k \in S_c} w_{kc}\right)}{CP_c} * 100$$

where Class C is the group or class for which we want to calculate slippage rates. For example at a detailed level the groups are based on province, sex and age group.
CP_C is the census population projection for class C
w_{kc} is the survey weight for k_{th} responding unit in class C
S_C is the set of responding sample households in class C

Slippage rates for household surveys are generally positive because of frame under coverage. Figure 4.1 shows slippage rates at the person level by panel. At lower geographic levels, the slippage rate varies more. Table 4.1 shows the person level

slippage rates by province. We also computed slippage rates at the household level (Table 4.2). For household slippage rates for previous reference years, see Figure 4.2.

Figure 4.1 Person-level slippage rate by panel and reference year (%)



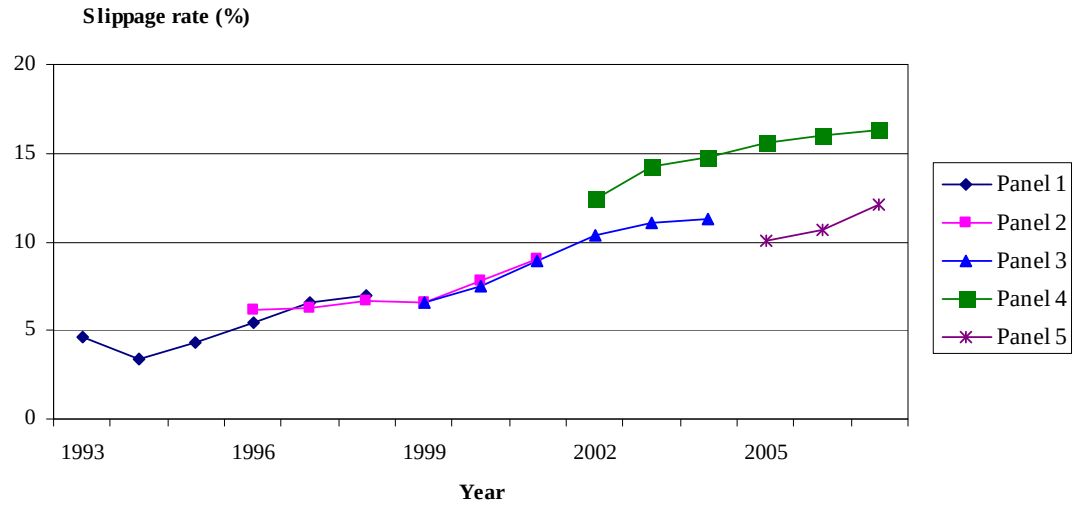
Looking at figure 4.1, one can see that the slippage rate trend is similar for all panels, with the rate always increasing between the first and the last wave. The higher person and household (see Figure 4.2) slippage rates for Panel 4 are due (at least in part) to an improper accounting of households selected to the SLID sample that did not appear on the sample file. At the beginning of a panel, it is belived that the effort to obtain a response from some households would be too high to send them to data collection and, generally, they are deemed non-respondents for the duration of the panel. We estimate the increase in slippage due to the omission of these non-respondent households to be in the neighbourhood of 2%. However, the impact on survey estimates should be negligible, as the error is corrected in part through the calibration of the final weights to census projections.

Table 4.1 Person-level slippage rates by province in 2007 (%)

	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Canada
						%					
Panel 4	8.7	7.7	10.0	12.4	12.9	21.5	15.7	11.9	22.3	25.3	18.8
Panel 5	2.0	4.1	5.8	2.2	5.7	20.0	4.8	0.5	20.1	17.7	14.1

In Table 4.1, the slippage rate is higher in the western provinces (Alberta and British Columbia) and in Ontario, while it is lower in the eastern provinces (Newfoundland and Prince Edward Island) for panels 4 and 5. We also observed a slippage rate close to 0 (0.5%) in Saskatchewan for the 5th panel.

Figure 4.2 Household slippage rate by panel and reference year (%)



Unlike the person slippage rate, the household slippage rate has always increased from one wave to the next, except panel 1 between the first (1993) and second wave (1994) where it decreased slightly.

Table 4.2 Household level slippage rates by province and household size in 2007 (%)

	Panel 4 Household Size				Panel 5 Household Size			
Province	**1**	**2**	**3 or more**	**All**	**1**	**2**	**3 or more**	**All**
Newfoundland	12.1	3.7	4.9	6.0	3.0	-3.0	-1.0	-0.9
Prince Edward Island	-3.5	5.7	4.1	2.9	15.6	2.3	-2.2	3.8
Nova Scotia	-0.9	6.2	13.2	6.6	4.0	10.7	-1.4	4.7
New Brunswick	-3.3	17.2	11.5	9.9	3.6	0.2	-1.9	0.3
Quebec	15.9	13.9	10.4	13.4	11.1	5.5	4.4	7.0
Ontario	13.1	15.1	23.0	17.9	10.3	19.0	19.6	17.1
Manitoba	17.2	13.3	14.5	14.9	-0.2	-0.5	5.3	1.7
Saskatchewan	-10.4	18.6	12.9	8.1	-0.5	-11.9	5.3	-2.6
Alberta	-2.0	13.5	28.8	15.9	6.2	20.3	19.8	16.6
British Columbia	27.7	22.8	23.5	24.5	14.6	17.0	16.4	16.1
Canada	**13.3**	**15.2**	**19.5**	**16.3**	**9.6**	**12.5**	**13.6**	**12.1**

Finally, if we compare the slippage rate by household size for Canada, one notices that the larger the household size, the larger the slippage rate will be.

For the two panels under study, as was the case for the person slippage rate, we observed the highest rates in Ontario, Alberta and British Colombia for households of size 3 or more and for all the households. British Columbia rates in the 4th panel are particularly high while all of them are above the 20% mark no matter what the household size. Still for the 4th panel, for households of size two, the slippage rates are lower in the Maritimes

provinces except New Brunswick which shows a rate higher than the Canadian rate. Also, one can see a negative rate in half of the provinces for households with only one person. Those provinces are Prince Edward Island, Nova Scotia, New Brunswick, Saskatchewan and Alberta.

For the 5th panel, in general, the slippage rates are lower than those of the 4th panel. While the provinces with the high rates are the same, this time, it's in Saskatchewan where the rates are at their lowest, except households of size 3 or more. Manitoba showed the biggest difference between panel 4 and 5 where the rate went from 14.9% to 1.7%.

5. Response rates

Since SLID has taken on the role of both a longitudinal and a cross-sectional survey, respective response rates are calculated. Cross-sectional response rates are calculated both at the person level and at the household level. Since sample persons have the option of giving tax permission thereby avoiding the income questions, it is possible to have complete data for income with no actual contact made during the reference year. Because of this the definition of a non-respondent is not straightforward.

If all persons in a household are non-respondent to both labour and income questions, then these persons (and households) are non-respondent.

With respect to those persons in households which are non-respondent to the labour questions but for whom we have tax data, it is determined whether the person is in the same household as the previous year (as of December 31). If the household is different this means the respondent has split from the original household. Since we have no information at all on the household composition of the new household, such persons are defined to be non-respondent.

Persons in households which are non-respondent to the labour questions but for whom we have income data and for whom the household has not changed since the previous year, are considered non-respondents if the household was a non-respondent household to the labour questions the previous January. Since updates to household composition are collected with the labour questions, this means that the household composition has not been updated for 2 consecutive years. Persons in households that have been non-respondent to labour questions in 2 consecutive January collections are therefore considered to be non-respondents to SLID.

Figure 5.1 shows the cross-sectional person response rates to SLID throughout the years of the survey. The person level response rates are calculated by dividing the number of cross-sectionally eligible respondents to the labour and/or income questions by the total number of cross-sectionally eligible people. An assumption is made that non-respondents are still in the target population unless there is evidence to the contrary. As a result this may somewhat underestimate response rates. Figure 5.2 shows the household response rates by region.

A household is considered a respondent household if at least one person in that household is considered a respondent. Household response rates are calculated by dividing the number of cross-sectionally eligible respondent households by the total number of cross-sectionally eligible households. Once again an assumption is made; non-respondent households are assumed to be still in the target population unless there is evidence to the contrary. As a result this may somewhat underestimate response rates.

Non-response can potentially introduce a bias in the data. A bias is created if characteristics of respondents differ from those of non-respondents and this difference has an impact on the variable being studied. It is difficult to determine whether non-response is introducing bias, because there is a limited amount of information for non-respondents.

Figure 5.1 Cross-sectional person-level response rate by reference year (%)

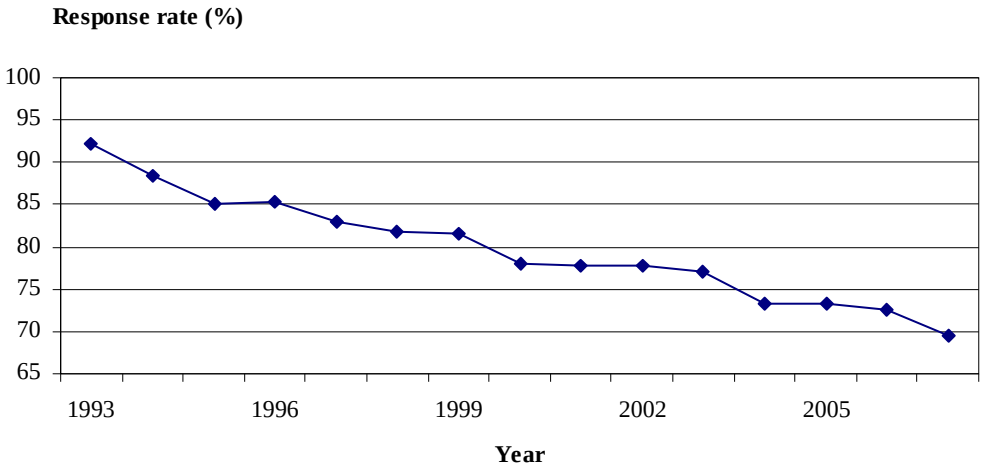
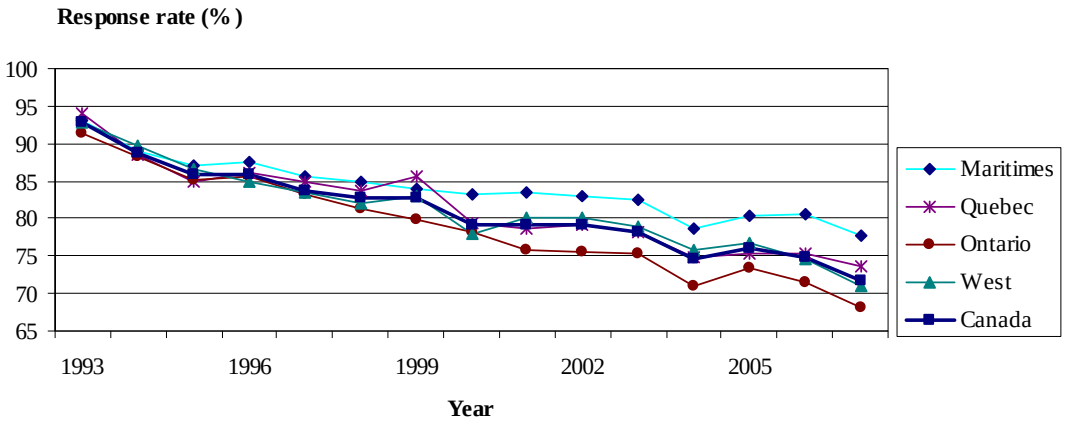


Figure 5.1 illustrates clearly that the person response rate is constantly declining since the beginning of the survey. It began at 92.1% in 1993 to reach a lower limit at 69.6% in 2007.

Figure 5.2 Cross-sectional household response rate by region and reference year (%)



The above graph displays once again the decreasing trend in the household response rate over the years. One can notice a deeper decrease in 2004. After that year, the rate went up a bit in 2005 but went back down to reach a minimum at 71.8% in 2007, for Canada. The response rate curve for Maritimes showed the highest rates while that of Ontario exhibited the lowest rates.

Table 5.1 shows the person response rates by phase. 'Respondent to labour questions' and 'Respondent to income questions' are the percentages of those who responded to only the labour or income sets of questions respectively whereas the 'Respondent to both sets' is the percentage of all those who responded in full or in part to both sets of questions.

Table 5.1 Cross-sectional person response rates by phase and reference year[1] (%)

Year	Response to both Labour and Income	Response to Labour Only	Response to Income Only	Non-response
1993	75.6	10.3	6.2	7.9
1994	75.1	10.5	2.8	11.6
1995	71.7	10.0	3.3	14.9
1996	71.6	10.8	2.9	14.6
1997	68.9	12.2	2.2	16.7
1998	68.8	10.4	2.6	18.2
1999	65.5	13.6	2.5	18.5
2000	56.1	17.3	4.6	22.0
2001	63.3	10.4	4.1	22.2
2002	61.6	10.8	5.4	22.2
2003	63.9	7.9	5.4	22.9
2004	62.3	5.8	5.1	26.8
2005	62.1	8.3	2.9	26.7
2006	59.3	7.2	6.0	27.5
2007	56.9	7.0	5.8	30.4

1. From 2004 onwards, we combined the labour and income interviews into a January interview

Similar to Figure 5.1, Table 5.1 shows a general decrease of the response rate for persons who responded to both Labour and Income questionnaires. The maximum rate was recorded in the first year of the survey (75.6%) and the minimum in the last year (56.9%). Knowing these rates, it's no surprise that the proportion of non-respondents reached their peak in 2007 while it was at his lowest in 1993, when the survey began.

However, if we analyse rates for respondents who answered only one series of questions, the trend is different. For the Labour questions, besides 1997, 1999 and 2000, the rate fluctuated around the 10% mark between 1993 and 2002. Afterwards, it decreased significantly, ranging between 5.8% and 8.3%. For the Income questions, after remaining stable between 1994 and 1999, the rate doubled and stabilised for the remaining years under study, changing from 4.6% to 6.0%, except in 2005 when the rate dropped to 2.9%.

Due to the conceptual difficulty in defining a longitudinal household, only person level longitudinal response rates are calculated. Table 5.2 shows person level longitudinal response rates by panel. These rates are calculated by dividing the number of longitudinal respondents by the original number of longitudinal persons selected in the panel.

Table 5.2 Longitudinal person-level response rates by panel and wave (%)

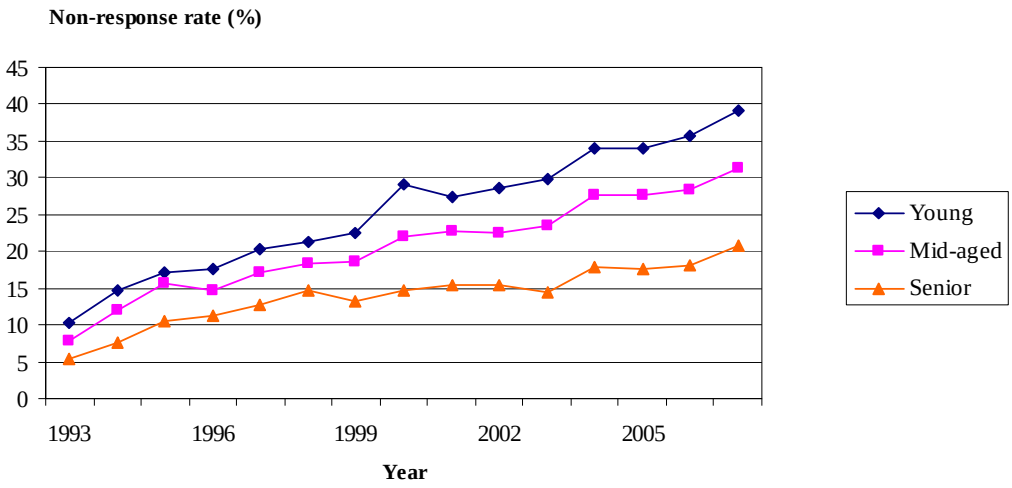
	Wave of panel					
Panel (year panel began)	**1**	**2**	**3**	**4**	**5**	**6**
Panel 1 (1993)	93.3	89.6	86.5	83.9	82.6	81.5
Panel 2 (1996)	89.5	86.8	85.2	82.7	78.5	77.4
Panel 3 (1999)	83.9	83.0	83.0	79.6	76.4	73.7
Panel 4 (2002)	81.2	83.2	78.3	75.0	71.6	68.9
Panel 5 (2005)	78.8	80.6	77.3	…	…	…

… Not applicable

Table 5.2 shows a decreasing trend in the longitudinal response rate. Not only does the longitudinal response rate drop over the life of the panel; it is also lower for each successive panel. For example, the rate went from 93.3% in wave 1 to 81.5% in wave 6 for the first panel while it dropped from 81.2% in wave 1 to 68.9% in wave 6 for the fourth panel.

Figure 5.3 shows the longitudinal non-response rates each year by age group. 'Young' are people at least 16 years of age but less than 30, 'Mid-aged' are people 30 years of age or older but less than 60 years of age and 'Senior' are people at least 60 years of age.

Figure 5.3 Longitudinal non-response rate by age group



Finally, comparing the longitudinal non-response rates by age group, Figure 5.3 shows an increase for all age groups. If we look at the rates between 1993 and 2007, one can see that they are approximately four times bigger in 2007 than in 1993. The young people, those between 16 and 30 years of age, recorded a non-response rate two times bigger than the seniors. As a matter of fact, in 2007, 39.2% of the young people didn't answer the survey compared to 20.2% for senior citizens. This is not surprising since in general, young people are more difficult to reach than senior people, who are more likely to be at home.

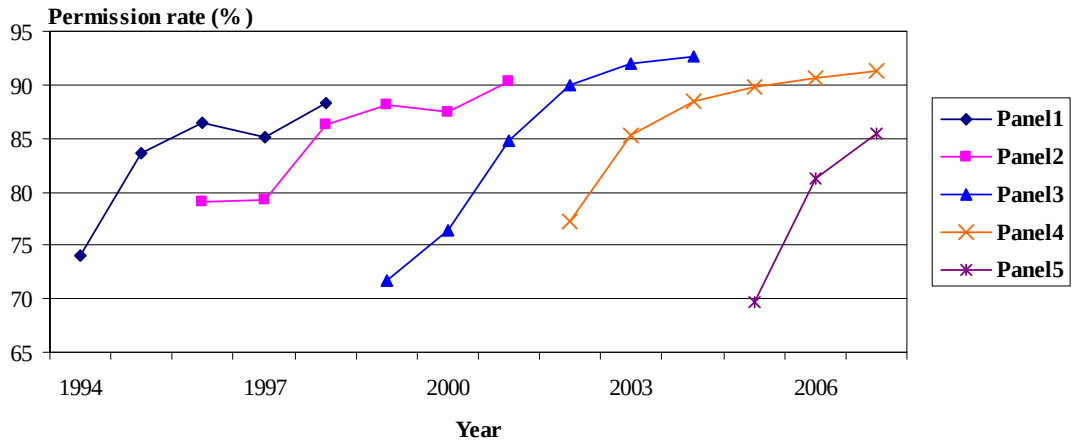
6. Tax permission rates

Prior to reference year 2004, there were two interviews every year: in January the interview was about activities such as working, going to school, looking for work or retirement. The second interview in May was about income, but wasn't necessary if the respondent gave Statistics Canada permission to obtain the required data from tax records. The tax source should provide consistent data of high quality and so a high permission rate should ensure good quality survey income estimates. The respondent was asked for this permission at the end of the January interview. If permission was not given, the respondent was contacted again in May. At this time the respondent was once again asked if he/she would prefer to give permission to access tax records. If permission was not provided, the interview proceeded. Starting in reference year 2004, permission was asked only once, in January. If it was not provided, the interview continued immediately with the income questions.

Figure 6.1 shows permission rates by panel over the years for the survey. The option to give tax permission was given for the first time in the May collection for the 1994 reference year. Prior to this, all income data were collected through interview. Percentages in figure 6.1 are based on the number of respondents over the age of 15 who are cross-sectionally eligible. Permission from the respondent is obtained once and for

the entire panel life duration. Therefore, the cumulative effect of the permission rate may hide the effort made yearly at the collection stage to obtain permission from the new respondents.

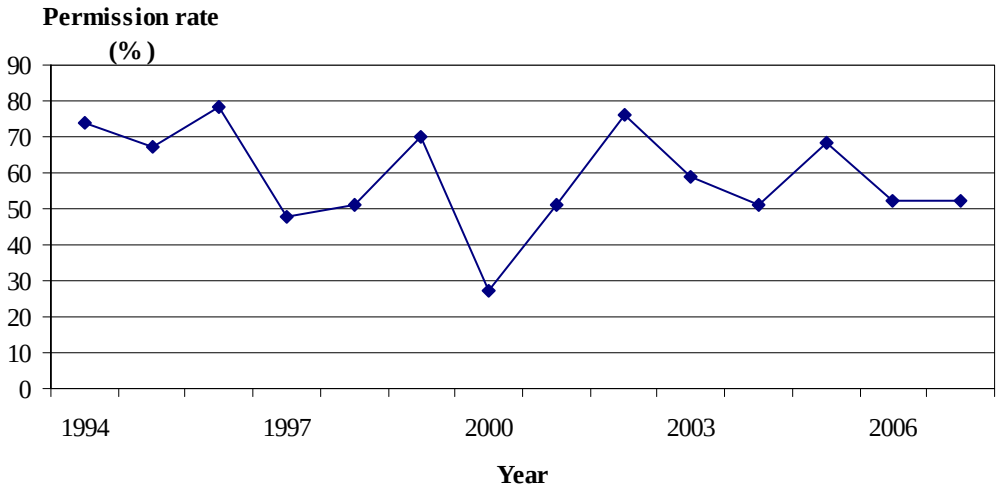
Figure 6.1 Permission rate by panel and reference year (%)



There's a similar trend in the annual permission rate for all panels. The rate showed a strong increase in the first three waves, with the exception of the second wave of the second panel where it remained stable. Then, it continued its rise, but more smoothly in the last three waves. We even observed a decrease in the permission rate between the 4th and the 5th wave for the first two panels, but it went back up in the last wave.

Figure 6.2 below shows the permission rates for new eligible respondents who gave their permission to access their tax data by reference year.

Figure 6.2 Permission rate for new respondents by reference year (%)



The permission rate for new respondents underwent major fluctuations during the period under study. It varied between 27.2% in 2000 and 78.4% in 1996. One can also notice that the year of the introduction of a new panel (1996, 1999, 2002 and 2005) always shows the highest permission rates for new respondents. We also note that the rate was very low in 2000. This corresponds to the first reference year that the May interview was not conducted.

7. Tax linkage rates

While respondents may grant Statistics Canada permission to use their tax data, they are not asked for their Social Insurance Number (SIN). Without a SIN to identify SLID respondents on the tax file, it is necessary to perform a linkage operation to find a respondent's SIN. The generalized record linkage system (GRLS) developed at Statistics Canada is used to perform this linkage.

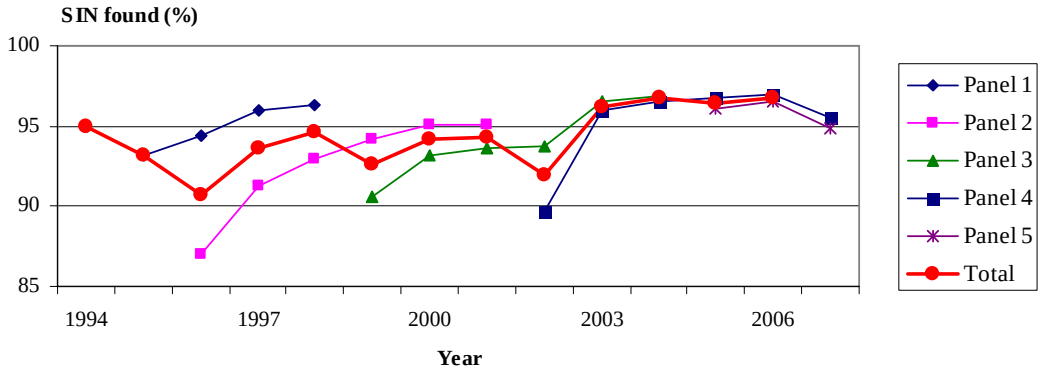
After preprocessing of both the tax file and the SLID file to ensure compatible formatting of all match variables, a direct match is performed using 7 key matching variables. These matching variables are: Sex, province, soundex[1] code for surname, surname, date of birth, postal code and first initial. The SLID record can have no missing data for key matching variables. Output for the direct match is manually reviewed for errors where a SLID record matches to more than one tax record, where more than one tax record matches to a SLID record, and where the first given name is not the same on the 2 sources (only first initial is used in the tax match). The match rate on the direct match is approximately 55 percent.

The unmatched records are then run through a statistical match. Pockets[2] for matching are defined. The files are segmented into pockets with sex, province and surname soundex code defining a pocket. Every record within a pocket on the SLID file is compared with every record within the same pocket on the tax file. Factors of importance are assigned for full agreement, partial agreement, and disagreement. These factors are numeric values and are used to evaluate the likelihood that a pair of records (one from SLID and one from tax) represent the same person. Factors are defined for each of the matching variables. Thresholds are defined whereby records are determined to be definite matches if their total factor is greater than the upper threshold or definite non-matches if their total factor is below the lower threshold. Manual verification is done to ensure the quality of the matches.

Figure 7.1 gives the percentage of the SLID sample giving tax permission for which a SIN can be found. Since some respondents who give tax permission have not filed a tax return not all cases for which a SIN is found will result in successful tax linkages.

1. Soundex is a name coding routine used in order to remove any common spelling errors from the surnames of respondents. This encoding is done based on the sound of the surname.
2. Pockets are groups of individuals on both the tax file and the SLID file with the same sex, province and soundex code.

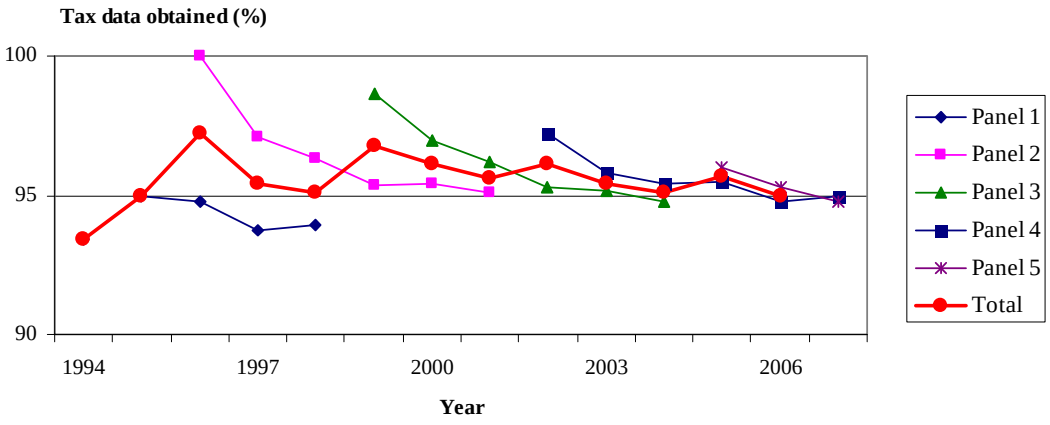
Figure 7.1 People giving permission for which a SIN was found by reference year (%)



In general, the proportion of respondents giving permission and for which a SIN was found showed an increasing trend in the six waves for all the panels. We observed a bigger raise between the 1st and the 2nd wave but the slope of the curve was less important for the subsequent waves. Between the 5th and 6th wave, the rate was stable and it has even decreased a bit for a few panels.

Figure 7.2 gives tax linkage rates for those in the SLID sample for which we were successful in finding a SIN.

Figure 7.2 Tax linkage rates when a SIN was found by reference year (%)



Concerning the linkage rate, it follows the same trend for all panels. In the first wave, the rate was at his maximum but went constantly down to remain stable around the 95% mark in the last wave. The global curve illustrates clearly the scenario where the rate was at its peak for the years where a new panel was introduced (1996, 1999, 2002 and 2005) and it went down the years after.

However, one can see that the initial linkage rates decreased for panels 2 through 5. As a matter of fact, for the first wave of panel 2, the linkage rate was 100% but it was 98.7%, 97.2% and 96.0% for the first wave of panels 3, 4 et 5 respectively.

Finally, table 7.1 compares the proportion of records coming from tax data to those collected during the telephone interview.

Table 7.1 Proportion of respondents coming from tax or interview by reference year[3] (%)

Year	Tax	Interview	Other[1]
1999	71.9	12.0	16.2
2000	74.0	0.0	26.0
2001	78.9	5.0	16.1
2002	74.2	8.8	17.0
2003	81.4	5.2	13.4
2004	83.4	5.0	11.7
2005	73.6	9.8	16.6
2006	78.8	5.9	15.3
2007	79.8	4.7	15.5

0 True zero or a value rounded to zero

1. These are respondents not linked to tax and without responses to income questions.

In the above table, one notices that most of the income data come from tax records. The proportions went from 71.9% in 1999 to 83.4% in 2004. Since 2003, the percentage fluctuated around 80% except in 2005 where it was smaller at 73.6%. It's also in 2005 that we observed a higher proportion of income data from interview, close to the 10% mark while it was around 5% for the other years between 2003 and 2007.

8. Imputation rates

To compensate for non-respondent households in the SLID sample, a non-response adjustment is applied to SLID weights. However, partially responding households are kept in the sample and any income data that is missing for individuals within respondent households is imputed. These individuals may require complete imputation of all income variables or they may require only certain fields to be imputed. Imputation rates in SLID may be thought of as a measure of partial non-response in the survey.

Two methods of imputation are used in SLID: Longitudinal Imputation and Cross-sectional imputation. Cross-sectional imputation of income variables in SLID is done using a nearest neighbour approach. Longitudinal imputation of income is done by using last wave's income to impute for the current wave income. Some variables are also imputed using a deterministic approach.

3 Excluding records non eligible to income imputation.

For the nearest neighbour method, a set of basic consistency rules is defined and for a given record requiring imputation a set of consistent donors is identified. A set of matching variables, each of which are correlated with the variables to be imputed, is also defined. Through combined use of both a score function (for categorical matching variables) and a distance function (for numeric matching variables), the most similar consistent donor record is identified and used to impute data for the record.

The percentage of persons within respondent SLID households that were subject to total or partial imputation is shown in Table 8.1. Recall that a respondent SLID household is one in which at least one household member has responded partially or completely to either the labour or income questions of the survey. In total eighteen income variables are imputed during SLID income imputation. Many individuals require only partial imputation. Partial imputation is when some (but not all) income items are substituted with information supplied by another individual.

Table 8.1 Income-variable imputation for respondents by province in 2007 (%)

Province	Total Imputation[1]	Partial Imputation[2]	No Imputation
Newfoundland	1.7	20.1	78.2
Prince Edward Island	2.1	19.3	78.6
Nova Scotia	2.5	20.7	76.8
New Brunswick	1.8	19.9	78.3
Quebec	1.9	18.4	79.7
Ontario	3.0	24.3	72.7
Manitoba	2.3	24.1	73.6
Saskatchewan	2.1	21.8	76.1
Alberta	3.2	25.1	71.7
British Columbia	2.5	25.7	71.8
Canada	**2.5**	**22.4**	**75.1**

1. No information provided by the respondent. All data items imputed.
2. One or more data items imputed with some information provided by the respondent.

The above table shows that, for Canada, almost one quarter of the records needed some imputation. The lowest imputation rate was found in Quebec where almost 80% of the records weren't subject to imputation. However, in the West (Alberta and British Columbia), the partial imputation rates were at their highest where they broke the 25% mark.

Few records needed total imputation. The rates fluctuated between 1.7% and 3.2% in the different Canadian provinces.

In table 8.2 we compare the percentage of tax data records requiring imputation to the percentage of records for which data is collected through the telephone interview. The need for partial imputation is determined after combining responses to both the labour and income questions. Inconsistencies are corrected through the imputation process.

The table also shows the percentage of individuals subject to partial imputation who require between one and seventeen variables to be imputed.

Table 8.2 Tax or interview records needing partial or total imputation in 2007 (%)

	Data Source			
Imputation	**Tax**	**Interview**	**Other[1]**	**All**
Partial (1 variable)	8.4	13.7	0.0	7.3
Partial (2 to 9 variables)	0.3	37.1	0.0	2.0
Partial (10 to 17 variables)	0.0	0.1	…	13.1
Total imputation	…	…	100	2.5
No imputation	91.3	49.1	…	75.1
Total	**100.0**	**100.0**	**100.0**	**100.0**

… Not applicable.

0 True zero or a value rounded to zero

1. Records that are not linked to Tax and without responses to the income questions. Some of these records are partially imputed based on the information collected from the labour questions.

The above table shows that records for which we could get access to tax data needed almost no imputation as we observed that 91.3% of the records didn't require any imputation. Also, 8.4% only needed partial imputation for one variable. That brings us close to 100%.

For the records collected through the interview, approximately half of them required some imputation and more than one third of them required partial imputation for 2 to 9 variables, which is much higher than the rates observed for tax records.

In 2002, new housing content relevant for housing research and policy development was added to SLID in cooperation with the Canada Mortgage and Housing Corporation (CMHC). The survey now collects information for the following sub-populations beginning with the 2002 reference year: the need for repairs (as determined by the dwelling occupant); the principal heating fuel of the dwelling; and whether a farm or home business is operated from the property. Also from homeowners the amount of regular mortgage payments; the amount of annual property taxes; and whether the dwelling is part of a registered condominium is collected. From renters the following is collected: the amount of monthly rent, what amenities are included in the rent (*e.g.*, heat, water, electricity); and whether the rent is subsidised by government or an employer.

The above information is in addition to information about home ownership and type of dwelling (since 1994) and information on the presence of a mortgage and the number of bedrooms in dwellings (since 1999).

Because of non-response to specific questions, imputation of housing related content was introduced in SLID in 2002. Two methods of imputation were used, longitudinal imputation and cross-sectional donor imputation. The cross-sectional donor imputation uses a similar method to that used in the income imputation, making use of the score function described above. Table 8.4 shows the percentage of responding SLID households that were subject to total or partial imputation.

In total twenty housing variables are imputed during SLID housing imputation. Many households require only partial imputation. Table 8.3 shows the break down of those requiring partial imputation.

Table 8.3 Households requiring imputation by province in 2007 (%)

Province	Total Imputation[1]	Partial Imputation[2]	No Imputation
Newfoundland	…	38.3	61.7
Prince Edward Island	…	38.6	61.4
Nova Scotia	…	35.8	64.2
New Brunswick	…	36.0	64.0
Quebec	…	29.7	70.3
Ontario	…	40.8	59.2
Manitoba	…	44.0	56.0
Saskatchewan	…	40.8	59.2
Alberta	…	43.6	56.4
British Columbia	…	46.0	54.0
Canada	**…**	**38.9**	**61.1**

1. No information provided by the respondent. All data items imputed.
2. One or more data items imputed with some information provided by the respondent.

For household imputation, almost 40 % of the households needed partial imputation. The highest rates have been recorded in the West (Manitoba, Alberta and British Columbia), while the lowest rates were found in Quebec. It was the only province with a rate under the 30% mark.

In total twenty housing variables are imputed during SLID housing imputation. Many households require only partial imputation. Table 8.4 shows the breakdown of those requiring partial imputation.

Table 8.4 Households requiring imputation by number of variables needing imputation and year of reference (%)

	Number of housing variables needing imputation			
Year	**1**	**2 to 5**	**6 to 19**	**One or More**
2004	10.5	9.9	10.9	31.3
2005	10.2	10.1	15.7	36.0
2006	10.0	7.1	22.6	39.7
2007	9.8	6.6	22.5	38.9

Finally, if we look at the partial imputation rates at the household level for the number of variables needing imputation, we notice that in 2006 and 2007, the rates are higher than those of the two previous years. Most particularly, for the category 6 to 19 variables

needing imputation, the proportion of households doubled in two years going from 10.9% in 2004 to 22.6% in 2006. It remained stable in 2007.

9. Rounding of income data

A small percentage of SLID income data comes from data collected in a telephone interview. While data obtained from the tax file is thought to be consistent for the most part, the quality of data coming from collection is not known. While some respondents may give precise amounts, it is possible that many of the responses given are estimates or approximations, which therefore are stated in hundreds or thousands of dollars rather than precise dollars and cents.

To test for the possible presence of rounding, distributions of each of the last 4 digits of reported variables were produced. One would normally expect the distribution to be approximately uniform with the digits 0 to 9 each comprising about 10 percent of the distribution. A prevalence of zeroes in the last digit would indicate rounding to the nearest 10, in the second last digit rounding to 100, etc. Table 9.1 shows the distribution of each of these digits for all reported values greater than ten thousand of the variable wages and salaries from both collected data (e.g. collected by interview) and tax data.

Table 9.1 Distribution of the last four digits of wages and salaries greater than $9,999 in 2007 (%)

	Fourth last digit		Third last digit		Second last digit		Last Digit	
Digit	Collected	Tax	Collected	Tax	Collected	Tax	Collected	Tax
0	33.2	11.2	90.8	12.1	95.8	12.9	96.9	14.2
1	3.6	10.9	0.5	10.0	0.3	9.7	0.4	9.3
2	9.4	10.7	0.9	9.9	0.2	9.8	0.4	9.9
3	5.6	10.2	0.8	9.5	0.5	9.4	0.3	9.4
4	6.2	9.9	0.8	9.6	0.7	9.6	0.3	9.5
5	19.6	10.0	2.8	10.2	0.5	9.8	0.2	9.8
6	5.9	9.8	1.1	9.8	0.4	9.7	0.4	9.6
7	4.8	9.3	0.5	9.6	0.3	9.6	0.4	9.5
8	8.6	9.5	1.1	9.7	0.9	9.8	0.3	9.5
9	3.3	8.4	0.8	9.6	0.4	9.8	0.4	9.2

Table 9.1 shows clearly that wages and salaries equal or higher than $10,000 have been rounded. The third, second and last digit was a zero in 90.8%, 95.8% and 96.9% of the cases respectively for collected records while the distribution is more uniform for each of the numbers between 0 and 9 for data coming from tax records.

Regarding collected data, for the 4th digit, approximately one third of the records displayed a zero and nearly 20% had a five. While these results aren't as striking as for the last three digits, they still indicate a rounding.

Table 9.2 shows the prevalence of zeroes in each of the last 4 digits for all reported non-zero values for a selection of SLID variables.

Table 9.2 Proportion of zeros in the last four digits declared for some variables in 2007 (%)

	Digit			
Variable	**Fourth-last**	**Third-last**	**Second-last**	**Last**
		%		
Wages and salaries	26.7	82.2	93.7	96.1
Investment income	11.5	29.6	59.6	71.9
Social assistance	12.8	25.6	69.8	86.0
EI benefits	6.3	42.2	85.7	95.2
Non-farm self-employment income	37.3	82.8	98.2	97.6

The last results explain without a doubt the constant increase of the number of zeros when we look at the position of the last digits going from the fourth to the last one. For wages and salaries and non-farm self-employment income, one can see a higher number of zeros comparing to the other variables starting at the 4th digit from the end and, with greater extent, for the third one.

For investment income, social assistance and EI benefits, we notice a strong increase in the number of zeros comparing the third digit from the end to the second one. These increases vary from 30.0% to 44.2%.

Finally, all the variables had a zero for the last digit in more than 95% of the cases except for investment income (71.9%) and social assistance (86.0%) but the proportions of zeros were still pretty high.