

Data Quality for the 2008 Survey of Labour and Income Dynamics (SLID)

Jean-François Bastien, Caroline Cauchon, Wisner Jocelyn, Ann Lim, Ming Jie Yang

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

Table of contents

1. Introduction.....	5
2. Sample composition/attrition	6
3. Sampling errors	9
4. Coverage errors	10
5. Response rates.....	13
6. Tax permission rates	17
7. Tax linkage rates	18
8. Imputation rates	21
9. Proxy interview rates	24
10. Rounding of income data	26

1. Introduction

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated to produce estimates starting in 1993. 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 used to collect income data for respondents who did not give permission to link to their income tax records. Since 2004, however, the May interview was dropped in order to save on collection costs. If a respondent does not grant permission to link to the T1 tax file, the income questions were asked 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 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*", <http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=75F0011X>

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 presents highlights of data quality indicators for Canada for reference year 2008.

Table 1.1 Main SLID quality indicators for Canada in 2008

Indicator	Statistic
Longitudinal sample size <ul style="list-style-type: none"> • Panel 5 • Panel 6 	42,330 40,912
Cross-sectional sample size (eligible longitudinal and cohabitants) <ul style="list-style-type: none"> • Panel 5 • Panel 6 	33,963 30,204
Coefficient of variation <ul style="list-style-type: none"> • Median total income 	0.9 %
Slippage rate - person	12.1 %
Slippage rate - household <ul style="list-style-type: none"> • Household size 1 • Household size 2 	10.6 % 9.5 %
Response rate <ul style="list-style-type: none"> • Cross-sectional - person • Cross-sectional - household • Longitudinal - person <ul style="list-style-type: none"> ○ Panel 5 ○ Panel 6 	67.5 % 70.6 % 72.8 % 71.0 %
Permission rate <ul style="list-style-type: none"> • Panel 5 • Panel 6 	88.2 % 72.2 %
Tax linkage rate (SIN found)	95.9 %
Imputation rate - person <ul style="list-style-type: none"> • Total imputation • Partial imputation 	3.2 % 23.7 %
Imputation rate - household <ul style="list-style-type: none"> • Partial imputation 	43.1 %

2. Sample composition/attrition

Table 2.1 gives the 2008 status for persons originally selected for the longitudinal sample for panels 5 and 6. The responding longitudinal sample is comprised of in-scope respondents, individuals who have moved to Yukon, the Northwest Territories or Nunavut, individuals who have moved outside Canada, institutionalized individuals and deceased individuals.

Table 2.1 Person status for the longitudinal sample in 2008

Person status for the longitudinal sample	Panel 5	Panel 6
In-scope (respondents)	29,063	28,598
In-scope (non-respondents)	6,245	11,795
Moved to Yukon, NWT, Nunavut	6	5
Moved outside Canada	277	82
Institutionalized	490	135
Deceased	994	230
Removed from sample ¹	5,240	67
Duplicate person/error ²	15	0
Total	42,330	40,912

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 households cannot be traced for two years then they are no longer pursued.
2. Respondents who were erroneously included in the household in the first year of a panel's existence.

While the total number of persons in panels 5 and 6 are very similar, there are major differences between the two panels when looking at the longitudinal status. There is almost twice as many in-scope non-respondents in panel 6 than in panel 5. The number of people who were removed from the sample is considerably larger in panel 5 as a result of many households that could not be traced and several which were hard refusals. This is not an unexpected result given that panel 5 was into its fourth wave in 2008.

The breakdown of the 2008 longitudinal sample by province and CMA is presented in the tables below. Note that the province and CMA is available for in-scope respondents only.

Table 2.2 Longitudinal sample composition by province in 2008

Province	Panel 5	Panel 6
Newfoundland	1,349	1,193
Prince Edward Island	819	792
Nova Scotia	1,788	1,541
New Brunswick	1,715	1,609
Quebec	5,392	5,172
Ontario	8,276	8,400
Manitoba	1,980	1,983
Saskatchewan	2,149	2,150
Alberta	2,947	2,787
British Columbia	2,648	2,971
N/A ¹	13,267	12,314
Total	42,330	40,912

1. This includes individuals who are in-scope non-respondents, moved to the territories, moved outside Canada, are institutionalized, are deceased, removed from the sample or are in error.

Table 2.3 Longitudinal sample composition by CMA in 2008

Census Metropolitan Area	Panel 5	Panel 6
Halifax	543	635
Quebec City	437	407
Montréal	1,130	1,062
Ottawa - Gatineau	784	973
Toronto	1,528	1,536
Hamilton	376	441
St. Catharines - Niagara	351	386
Kitchener	402	389
London	458	464
Windsor	305	370
Winnipeg	1,030	1,123
Calgary	617	615
Edmonton	915	683
Vancouver	963	1,078
Victoria	253	441
Other CMA or CA	10,558	10,257
Not a CMA	8,413	7,738
N/A ¹	13,267	12,314
Total	42,330	40,912

1. This includes individuals who are in-scope non-respondents, moved to the territories, moved outside Canada, are institutionalized, are deceased, removed from the sample or are in error.

Although originally designed as a longitudinal survey, cross-sectional estimates can also be produced from SLID data. The cross-sectional sample is composed of in-scope longitudinal respondents and non-longitudinal persons living with these longitudinal respondents (“cohabitants”). The breakdown of the 2008 cross-sectional sample by province is given in the table below.

Table 2.4 Cross-sectional sample composition by province in 2008

Province	In-scope longitudinal respondents		Cohabitants		Cross-sectional sample size	
	Panel 5	Panel 6	Panel 5	Panel 6	Panel 5	Panel 6
Newfoundland	1,349	1,193	188	64	1,537	1,257
Prince Edward Island	819	792	125	49	944	841
Nova Scotia	1,788	1,541	266	92	2,054	1,633
New Brunswick	1,715	1,609	236	77	1,951	1,686
Quebec	5,392	5,172	1,014	303	6,406	5,475
Ontario	8,276	8,400	1,365	468	9,641	8,868
Manitoba	1,980	1,983	332	97	2,312	2,080
Saskatchewan	2,149	2,150	376	141	2,525	2,291
Alberta	2,947	2,787	576	189	3,523	2,976
British Columbia	2,648	2,971	422	126	3,070	3,097
Total	29,063	28,598	4,900	1,606	33,963	30,204

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 one panel but before the start of the next one and who 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.

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 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 that is often used when dealing with estimates from a complex sample design. Table 3.1 gives CVs at the provincial and national level for selected SLID estimates.

Table 3.1 National and provincial coefficients of variation for selected variables in 2008 (%)

Variable (at the family-level unless otherwise indicated)	NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Canada
Median total income	2.0	2.8	2.1	2.9	1.4	1.2	2.4	2.1	1.5	2.0	0.9
Median market income	3.5	3.2	2.1	3.1	2.0	1.6	2.7	2.1	1.8	2.4	0.9
Median wages and salaries	4.1	4.1	2.4	3.2	1.7	1.4	2.6	2.2	1.7	3.5	0.8
Median EI benefits	7.2	11.4	8.8	8.5	4.6	5.2	13.1	9.8	10.3	8.2	2.3
Median social assistance	10.7	15.8	10.3	9.6	4.9	5.7	14.6	17.5	6.7	8.5	3.6
Median other income	21.6	22.9	14.3	16.3	8.6	11.5	22.1	11.5	11.8	8.5	4.4
Number of persons under LICO after tax	13.7	16.5	8.5	9.8	7.0	5.6	9.5	9.5	11.8	7.2	3.1
Number of persons with some employment ¹	1.7	1.7	1.3	1.4	1.1	0.9	1.4	1.5	1.2	1.5	0.5

1. This includes individuals who were:

- employed all year,
- employed part-year and unemployed part-year,
- employed part-year and not in the labour force part-year, or
- employed, unemployed and not in the labour force during the year.

As Newfoundland and Prince Edward Island have the smallest sample sizes, some of the highest CVs for a particular variable can generally be found in these two provinces. The exception is median EI benefits and median social assistance where the largest CVs were for Manitoba and Saskatchewan 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

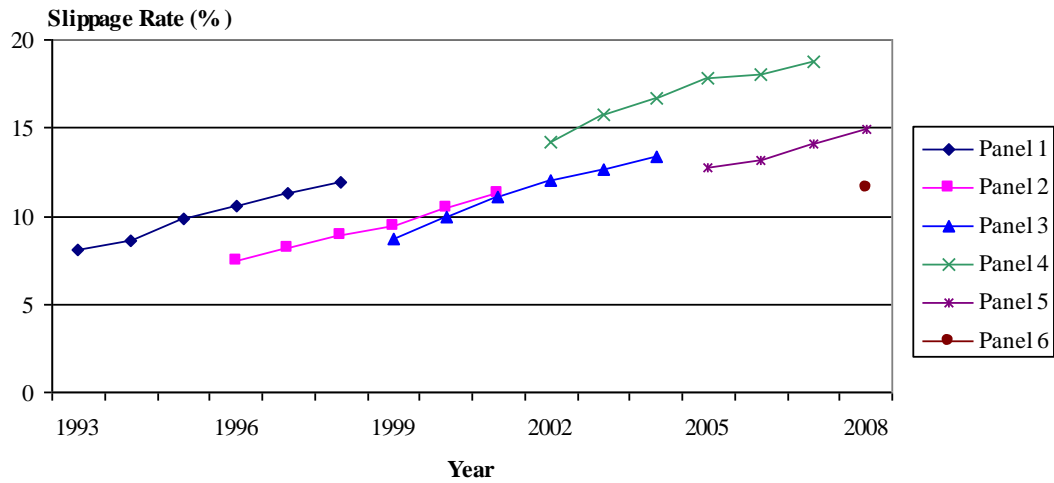
$$slippage_c = \frac{(CP_c - \sum_{k \in S_c} w_{kc})}{CP_c} * 100$$

where Class C is the group or class for which slippage rates are required. For example, the group could be based on province, sex and/or 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.

Slippage rates at the person-level are given by panel in Figure 4.1 and by province in Table 4.1.

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



As can be seen in Figure 4.1, the trend is similar for all panels with the slippage rate always increasing between the first and the last wave. The higher person-level and household-level (see Figure 4.2) slippage rates for Panel 4 are due, in part, to an improper accounting of households selected for the SLID sample that did not appear on the sample file. At the beginning of a panel, it is believed that the effort to obtain a response from some households would be too high to send them to data collection. As a result, they are generally deemed non-respondents for the duration of the panel. The increase in slippage due to the omission of these non-respondent households is estimated to be approximately 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.

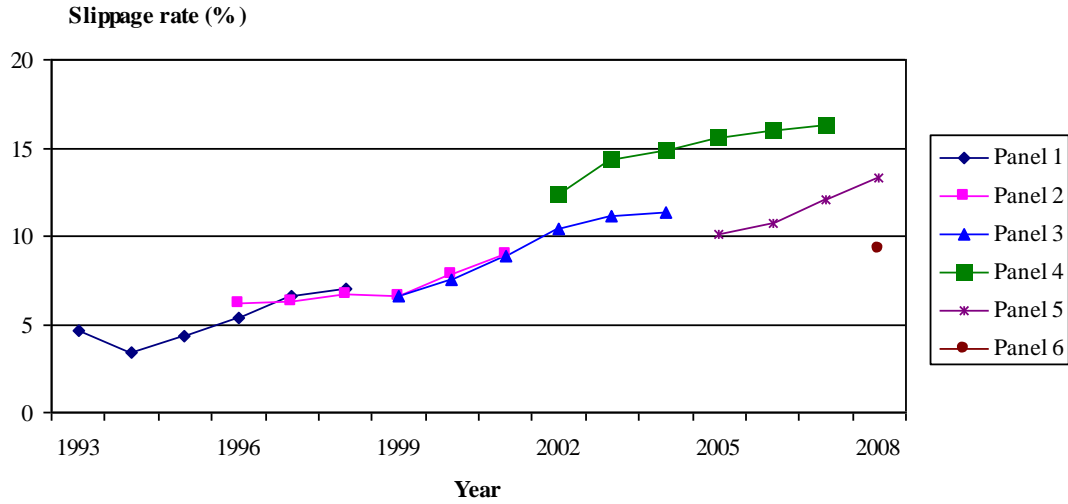
Table 4.1 shows that the highest the person-level slippage rates can be found in Ontario, Alberta and British Columbia.

Table 4.1 Person-level slippage rates by province in 2008 (%)

NL	PEI	NS	NB	QC	ON	MB	SK	AB	BC	Canada
1.0	2.5	6.9	2.9	5.5	13.6	5.0	-0.8	20.9	21.4	12.1

Slippage rates were also computed at the household-level; Figure 4.2 gives these rates by panel while Table 4.2 gives the slippage rates by province and household size.

Figure 4.2 Household-level slippage rates by panel and reference year (%)



The general trend that emerges in Figure 4.2 is that with the exception of one wave in panels 1 and 2 the slippage rate increases between one wave and the next.

Table 4.2 Household-level slippage rates by province and household size in 2008 (%)

Province	Household Size		
	1	2	All
Newfoundland	0.4	0.1	-1.4
Prince Edward Island	14.9	5.7	3.6
Nova Scotia	-0.3	6.8	4.0
New Brunswick	-8.4	0.8	-0.8
Quebec	9.7	5.7	6.1
Ontario	12.4	11.9	11.7
Manitoba	13.2	-1.0	5.0
Saskatchewan	-1.6	-4.4	-2.3
Alberta	2.7	21.9	17.1
British Columbia	21.0	10.8	18.5
Canada	10.6	9.5	10.3

As with the person-level slippage rates, the highest household-level slippage rates can be found in Ontario, Alberta and British Columbia.

5. Response rates

Since SLID has taken on the role of both a longitudinal and a cross-sectional survey, both types of response rates are calculated. Cross-sectional response rates are calculated at the person-level and at the household-level. Since sample persons have the option of giving tax permission and thereby avoiding the income questions, it is possible to have complete income data with no actual contact being made during the reference year. As a result, 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-respondents.

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

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 from the previous year are considered to be non-respondents if the household was a non-respondent 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 for 2 consecutive January collections are therefore considered to be non-respondents to SLID.

Figure 5.1 shows the cross-sectional person-level response rates for SLID by reference year. 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-sectional 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 presents the cross-sectional household response rates by region. A household is considered to be a respondent household if at least one person in that household is a respondent. Household-level 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 the assumption is made that non-respondent households are still in the target population unless there is evidence to the contrary; 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 (%)

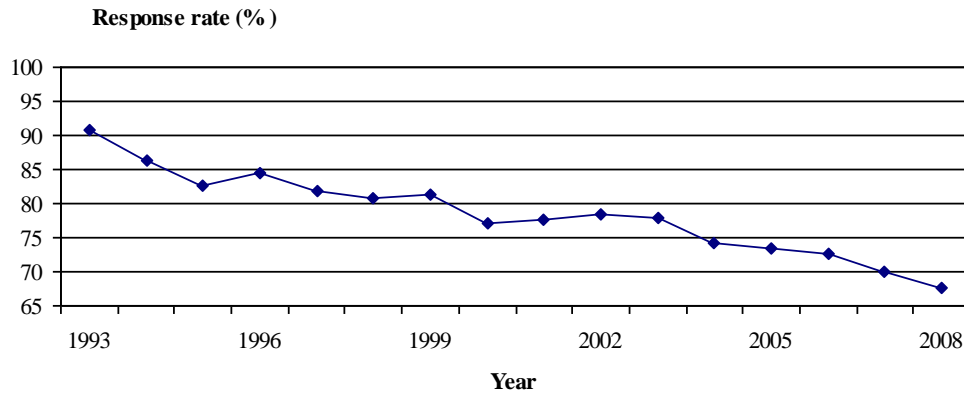
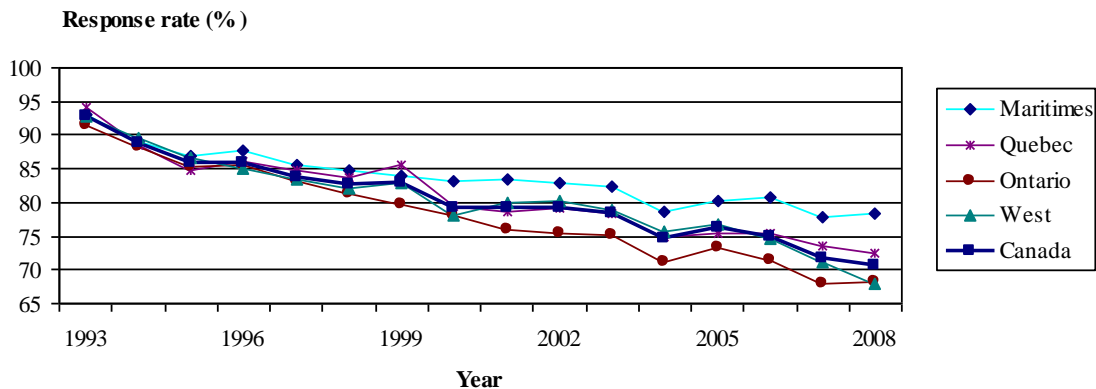


Figure 5.1 clearly illustrates that the person-level response rate has been declining since the start of the survey except for small increases in 1996, 1999 and 2002. The rate was at 90.8% in 1993 and has dropped to 67.6% in 2008.

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



The above graph also shows declining response rates over the years with a significant decrease in 2004. The rates increased slightly in 2005 but dropped afterwards falling to a low of 70.6% in 2008 at the Canada-level. In general, the Maritimes had the highest response rates while Ontario had the lowest.

Table 5.1 shows the person-level response rates by phase. ‘Respondents to Labour Questions Only’ and ‘Respondents to Income Questions Only’ reflect the proportion of those who responded to only the labour or income sets of questions respectively whereas the ‘Respondents to Both Labour and Income Questions’ is the proportion of all those who responded in full or in part to both sets of questions.

Table 5.1 Cross-sectional person-level response rates by phase and reference year¹ (%)

Year	Respondents to Both Labour and Income Questions	Respondents to Labour Questions Only	Respondents to Income Questions 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
2008	56.4	7.2	3.9	32.5

1. Since reference year 2004, labour and income questions were both asked during the January interview.

As in Figure 5.1, Table 5.1 shows a general decline in the response rate for persons who responded to both the labour and income questions. The highest rate was recorded in the first year of the survey (75.6%) and one of the lowest in 2008 (56.4%). Correspondingly, the proportion of non-respondents was highest in 2008 and 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, with the exception of 1997, 1999 and 2000, the rate was around 10% between 1993 and 2002. Since then, it has decreased significantly to a rate between 5.8% and 8.3%. For the income questions, after remaining stable between 1994 and 1999, the rate doubled and has been stable since; the rate is between 3.9% 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 gives 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.

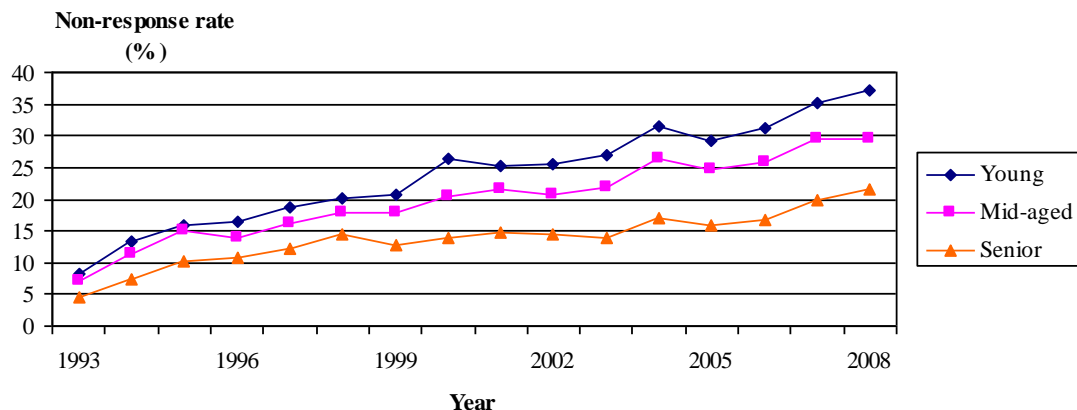
Table 5.2 Longitudinal person-level response rates by panel and wave (%)

Panel (first year of panel)	Wave of panel					
	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	72.8
Panel 6 (2008)	71.0
... Not applicable						

Table 5.2 shows a declining 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 by age group. ‘Young’ is defined as those people between the ages of 16 and 29, ‘Mid-aged’ are between the ages of 30 and 59 and ‘Senior’ are at least 60 years of age.

Figure 5.3 Longitudinal non-response rate by age group (%)



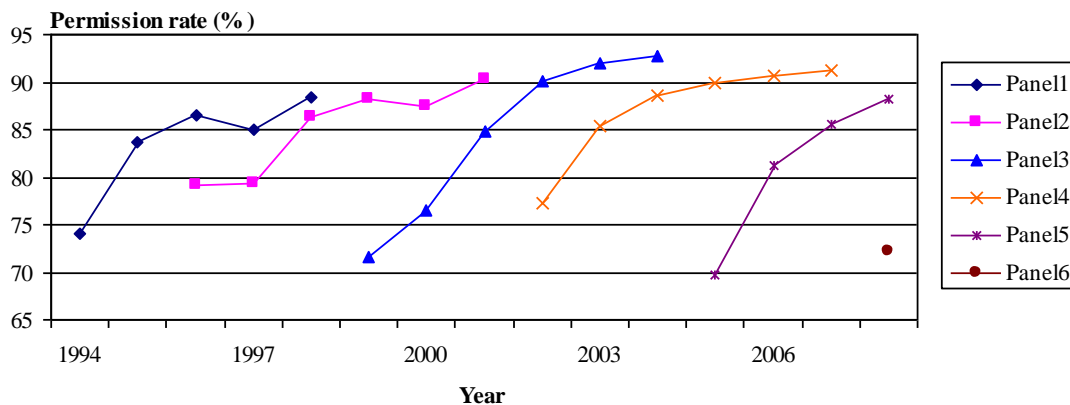
In looking at Figure 5.3, it is clear that there is an increase in non-response for all age groups. The non-response rates have quadrupled in the sixteen years that SLID has been conducted. Young people, those between 16 and 29 years of age, have a non-response rate almost twice that for the seniors. In particular, in 2008, 37.2% of young people did not respond to the survey compared to 21.4% for senior citizens. This is not surprising given that, in general, young people are more difficult to reach than seniors who are more likely to be at home.

6. Tax permission rates

Prior to reference year 2004, there were two interviews conducted every year. In January, the interview concerned activities such as working, going to school, looking for work or retirement. The second interview in May involved income, but respondents would not be contacted if they had given Statistics Canada permission to obtain the required data from tax records. The tax source should provide consistent data of high quality; thus, a high permission rate should ensure good quality income estimates. The respondent was asked for this permission at the end of the January interview. If this was refused, the respondent would be contacted again in May. At the May interview, the respondent was once again asked if he/she would prefer to give permission to access tax records. If the request was rejected, the interview proceeded. Starting in reference year 2004, permission was asked only once, in January. If the respondent declined, the interview continued immediately with the income questions.

Figure 6.1 shows permission rates by panel by reference year. The option to give tax permission was implemented for the May collection for the 1994 reference year. Prior to this, all income data were collected through interview. The figures are based on the number of respondents over the age of 15 who are cross-sectionally eligible. Permission from the respondent is obtained once for the duration of the panel. Therefore, the cumulative effect of the permission rate may hide the effort made annually at the collection stage to obtain permission from new respondents.

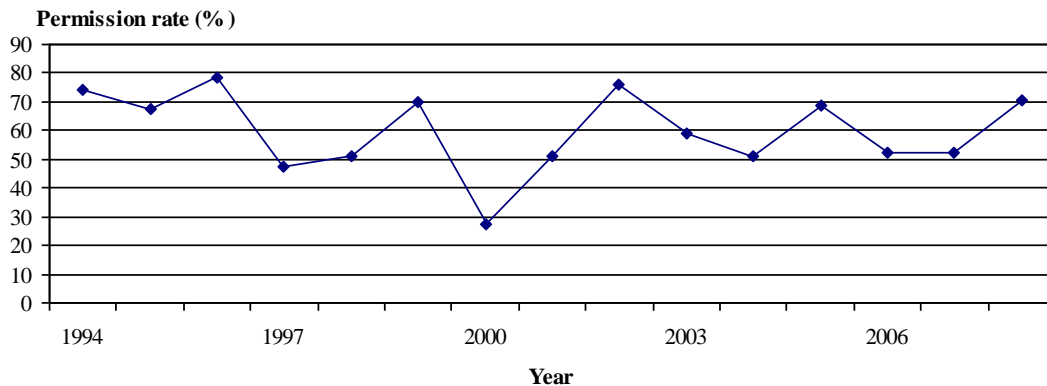
Figure 6.1 Permission rate by panel and reference year (%)



A similar trend appears for all panels in the annual permission rate. There is a strong increase in the rate in the first three waves (except for the second wave of the second panel). Permission rates continued to rise but less dramatically in the last three waves. There was a decrease in the permission rate between the 4th and the 5th wave for the first two panels but increases for the last wave.

Figure 6.2 below shows permission rates by reference year for new eligible respondents.

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



The permission rate for new respondents fluctuated dramatically; it varied between 27.2% in 2000 to 78.4% in 1996. The years in which new panels had been introduced (1996, 1999, 2002, 2005 and 2008) always had the highest permission rates for new respondents. We also note that the rate was very low in 2000.

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 probabilistic match to obtain a respondent's SIN.

The first step is to standardize matching variables on the SLID and tax files to ensure that the formats are compatible. This process includes the removal of all spaces from the address field and the use of phonetic coding such as NYSIIS and SNDX¹. The standardized variables that are available for the linkage process are: address, city, date of birth, first name, surname, sex, province, NYSIIS and SNDX code for surname, postal code, marital status, telephone number and first initial.

A SAS program developed at Statistics Canada compares data from the two data sources (tax and SLID). In order to make the match more manageable, SLID and tax records are grouped into "pockets" based on date of birth, postal code and SNDX code for surname. Every SLID record within a pocket is compared to every tax record in the same pocket. A weight is assigned based on the likelihood that a pair of records represent the same person. Thresholds are defined whereby a pair is deemed to be a definite match if the weight is greater than the upper threshold or to be a definite non-match if the weight is below the lower threshold.

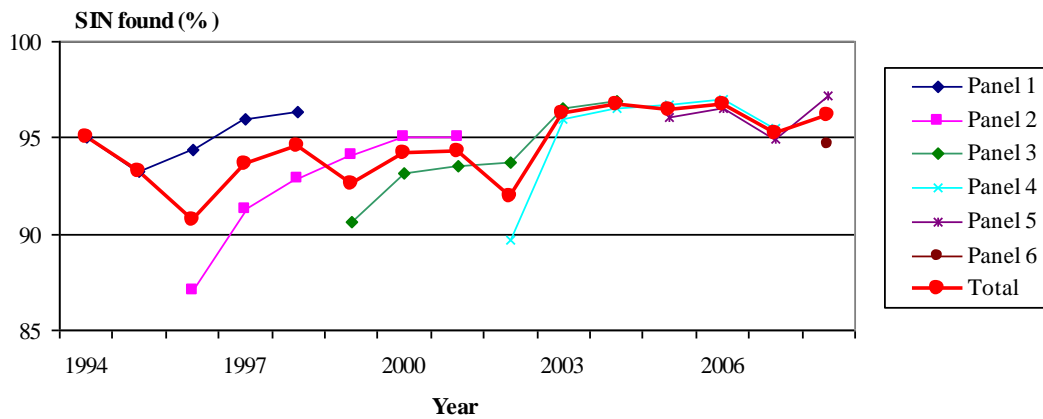
1. NYSIIS and SNDX are name coding routines used to remove common spelling errors from the surnames of respondents. This encoding is based on the sound of the surname.

The linkage of SLID records is undertaken twice for each processing cycle: once to the final tax file for the previous reference year and then again to the preliminary tax file for the current reference year. For potential matches, there are nine possible outcomes depending on whether a definite match or a questionable match (ie. neither a definite match nor a definite non-match) has been made between a SLID record and a tax record (final and/or preliminary). The result is that potential matches are accepted or are manually reviewed. It is possible that two SLID records may be linked to the same SIN; duplicates are resolved at the end of the linkage process.

The newly obtained SINs are then used to obtain tax information for those SLID respondents who gave permission to access their tax data.

Figure 7.1 gives the proportion of SLID respondents who gave tax permission for which a SIN could be found. As some respondents who gave tax permission had not filed a tax return, not all cases in which a SIN is found will result in a successful tax linkage.

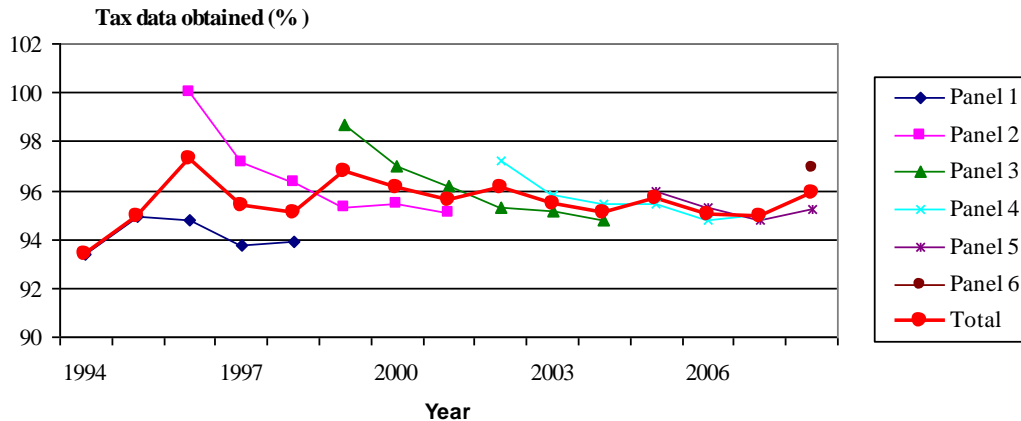
Figure 7.1 People giving permission for which a SIN was found by reference year (%)



In general, the proportion of respondents giving permission and for which a SIN was found showed an increasing trend over the six waves for all panels. Large increases were observed between the 1st and the 2nd wave but the increases were less pronounced for subsequent waves. Between the 5th and 6th wave, the rate stabilized and it actually decreased slightly for a few panels.

Figure 7.2 gives tax linkage rates for those SLID respondents who gave tax permission and for whom a SIN had been successfully assigned.

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



The same general trend is observed for all panels in the above figure. The maximum linkage rate occurs in the first wave but then steadily declines to 95% in the last wave. As can be seen clearly in the overall rate, the linkage peaks in those years in which a new panel was introduced (1996, 1999, 2002, 2005 and 2008).

However, the initial linkage rates decreased from panel 2 through to panel 5; it was 100% for the first wave of panel 2 but was 98.7%, 97.2% and 96.0% for the first wave of panels 3, 4 and 5 respectively. This rate increased slightly to 96.9% for panel 6.

Table 7.1 compares the proportion of records coming from tax data to those collected during telephone interviews.

Table 7.1 Proportion of respondents coming from tax or interview by reference year¹ (%)

Year	Tax	Interview	Other ²
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
2008	74.4	8.9	16.7

1. Excluding records not eligible for income imputation.

2. These are respondents who were not linked to tax data and who did not respond to income questions.

In the above table, it can be seen that most of the income data comes from tax records; the proportion ranged from a low of 71.9% in 1999 to a maximum of 83.4% in 2004. This figure was generally around 80% except for years where a new panel started (1999, 2002, 2005 and 2008). In the first year of a new panel, a greater proportion of income data came from interviews when compared to other years.

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 SLID income variables uses a nearest neighbour approach. Longitudinal imputation of income uses income from the previous wave to impute income for the current wave. Some variables are also imputed using a deterministic approach.

For the nearest neighbour method, a set of basic consistency rules is defined and a set of consistent donors is identified for a given record requiring imputation. A set of matching variables, each of which is correlated with the variables to be imputed, is also defined. Through the combined use of 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 proportion of persons within responding SLID households that were subject to total or partial imputation is given 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, up to eighteen income variables can be imputed during SLID income imputation. Many individuals require only partial imputation where some (but not all) income items are substituted with information from another individual.

Table 8.1 Income variable imputation for respondents by province in 2008 (%)

Province	Total Imputation ¹	Partial Imputation ²	No Imputation
Newfoundland	2.4	20.1	77.5
Prince Edward Island	2.5	21.7	75.8
Nova Scotia	2.3	21.5	76.2
New Brunswick	1.9	21.6	76.4
Quebec	3.4	20.5	76.1
Ontario	3.6	26.3	70.1
Manitoba	2.2	22.3	75.5
Saskatchewan	2.7	21.5	75.8
Alberta	4.7	26.3	69.0
British Columbia	3.4	26.8	69.7
Canada	3.2	23.7	73.1

1. No information was provided by the respondent. All data items were imputed.

2. One or more data items were imputed with some information provided by the respondent.

The above table shows that almost three-quarters of the records did not require any income imputation at the Canada-level. The lowest imputation rate was found in Newfoundland where 77.5% of records did not undergo any imputation. However, the highest partial imputation rates (approximately 26%) were found in Ontario, Alberta and British Columbia.

Few records needed total imputation; the rates ranged from 1.9% to 4.7% at the province-level.

In Table 8.2, the income imputation rates were compared for tax data records and records which were collected through telephone interviews. The need for partial imputation is determined after combining responses from the labour and income questions. Inconsistencies are corrected through the imputation process. This table also gives an indication of the extent to which partial imputation was employed (1 variable, 2 to 9 variables and 10 to 17 variables).

Table 8.2 Tax or interview records requiring partial or total income imputation in 2008 (%)

Imputation	Data Source			All
	Tax	Interview	Other ¹	
Partial (1 variable)	7.2	14.4	0.0	6.6
Partial (2 to 9 variables)	0.3	37.1	0.0	3.5
Partial (10 to 17 variables)	0.0	0.4	...	13.5
Total imputation	100.0	3.2
No imputation	92.5	48.1	...	73.1
Total	100.0	100.0	100.0	100.0

... Not applicable.

1. Records that are not linked to tax data and do not have 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 very few of the tax records required any imputation; 92.5% of records which could be linked to tax data did not undergo any income imputation. Imputation of one variable only was required for 7.2% of the tax records. These two scenarios cover almost all the tax records.

For records collected through interviews, approximately half required some income imputation and more than a third required partial imputation of 2 to 9 variables. This is much higher than the rates observed for tax records.

Because of non-response associated with specific questions, imputation of housing related content was introduced to SLID. Two methods of imputation were used: longitudinal imputation and cross-sectional donor imputation. The cross-sectional donor imputation method is similar to that used in income imputation and involves the use of a score function.

Table 8.3 gives the proportion of responding SLID households that underwent imputation of housing variables.

Table 8.3 Households requiring imputation of housing variables by province in 2008¹ (%)

Province	Total Imputation ²	Partial Imputation ³	No Imputation
Newfoundland	...	39.2	60.8
Prince Edward Island	...	43.0	57.0
Nova Scotia	...	39.1	60.9
New Brunswick	...	37.8	62.2
Quebec	...	36.5	63.5
Ontario	...	44.3	55.7
Manitoba	...	45.3	54.7
Saskatchewan	...	43.4	56.6
Alberta	...	48.7	51.3
British Columbia	...	50.8	49.2
Canada	...	43.1	56.9

1. For reference year 2008, the variable that indicated whether a dwelling was a condominium was excluded from the calculation of the partial imputation rate as there had been a change in the way this variable was defined. Keeping this variable would have artificially boosted the partial imputation rates.
2. No information was provided by the respondent. All data items were imputed.
3. One or more data items were imputed with some information provided by the respondent.

At the Canada-level, 43.1% of households needed partial imputation of housing variables. The highest rates were found in Manitoba, Alberta and British Columbia with half of the B.C. households requiring some imputation. The lowest rate was found in Quebec at 36.5%.

In total, up to twenty variables are imputed during SLID housing imputation. Table 8.4 gives a breakdown of those requiring partial imputation.

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

Year	Number of housing variables needing imputation			
	1	2 to 5	6 to 19	One or More
2004	10.5	10.2	10.6	31.3
2005	10.2	10.6	15.2	36.0
2006	10.0	7.4	22.3	39.7
2007	9.8	6.9	22.3	39.0
2008	8.8	5.9	28.4	43.1

The number of variables needing imputation has increased annually. In particular, the proportion of households with 6 to 19 imputed variables has almost tripled from 10.6% in 2004 to 28.4% in 2008.

9. Proxy interview rates

A proxy interview occurs when an interviewer obtains information for a given person in the household from another household member who is willing to respond on his/her behalf. Information on the number of proxy interviews has been available since the reference year 2000. A variable is used to indicate if the interview information for a particular person was provided by proxy. Prior to the reference year 2004, respondents were interviewed twice a year, once in January and again in May so two proxy variables were created. Since reference year 2004, the May interview was dropped; therefore, only one proxy variable was created corresponding to the January interview. For comparison purposes, only the proxy rates from the reference year 2004 onwards are presented.

Proxy rates are calculated based on the number of SLID respondents aged 16 and over who furnished responses either directly or by proxy for the given reference year (overall number of respondents). Proxy rates are obtained by dividing the number of respondents by proxy by the overall number of respondents.

Figure 9.1 below shows the proxy rates starting from 2004. The rates went from 38.1% in 2004 to 33.9% in 2008. The rate decreased from 2004 to 2005, remained fairly stable from 2005 to 2007 before decreasing again.

Figure 9.1 Proxy rate by reference year (%)

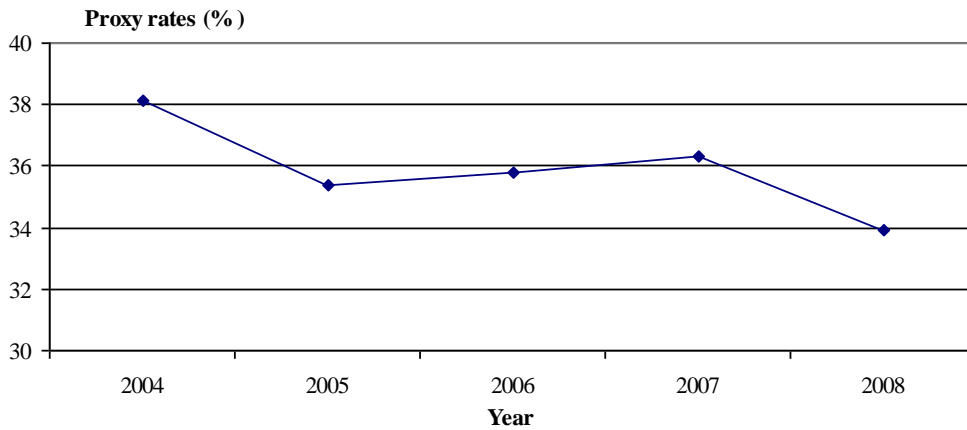
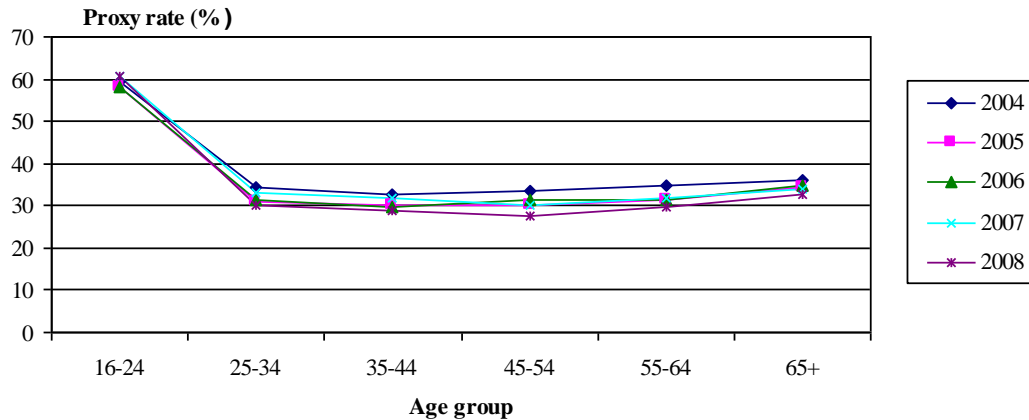


Figure 9.2 below shows the proxy rate by age group for the reference years 2004 to 2008. In general, the same trend can be observed for each year under study. The proxy rate is very high among respondents aged 16 to 24 years, around 60%, while the other age groups have rates around 32%. This difference can be explained by the fact that young people are more difficult to reach than the rest of the population and that most are still living with their parents who tend to respond for them. Support for this comes from the fact that 12% of the sample is composed of young people (age 16 to 24) which as a group has a response rate of 67% compared to 74% for the rest of the population (age 25 or more).

Figure 9.2 Proxy rate by age group and reference year (%)



10. Rounding of income data

A small percentage of SLID income data is collected from telephone interviews. 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 and, as such, are stated in hundreds or thousands of dollars rather than precise dollars and cents.

To test for the possible presence of rounding, the distribution for each of the last 4 digits of reported variables were produced. The distribution would normally expected 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 10.1 gives the distribution of each of these digits for all reported values of at least \$10,000 for the wages and salaries variable from both collected data (e.g. collected by interview) and tax data.

Table 10.1 Distribution of the last four digits of wages and salaries greater than \$9,999 in 2008 (%)

Digit	Fourth last digit		Third last digit		Second last digit		Last Digit	
	Collected	Tax	Collected	Tax	Collected	Tax	Collected	Tax
0	34.9	11.4	89.3	11.8	95.5	13.2	96.7	14.3
1	4.2	10.6	0.8	10.0	0.2	9.5	0.2	9.3
2	8.9	10.8	1.1	10.0	0.7	9.8	0.4	9.6
3	5.3	10.0	0.6	9.6	0.5	9.4	0.4	9.2
4	5.4	10.3	1.0	9.8	0.5	9.8	0.3	9.5
5	18.0	9.9	3.1	9.7	0.8	9.7	0.6	9.7
6	6.5	9.7	1.5	9.9	0.4	9.9	0.5	9.5
7	5.7	9.4	1.0	9.7	0.5	9.6	0.3	9.7
8	7.6	9.1	1.0	9.6	0.5	9.6	0.4	9.6
9	3.4	8.9	0.7	9.8	0.5	9.6	0.3	9.7

Table 10.1 clearly shows that collected wages and salaries equal to or higher than \$10,000 have been rounded. The third, second and last digit was a zero in 89.3%, 95.5% and 96.7% 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.

For the fourth last digit of collected data, more than a third of the records displayed a zero and 18% had a five. While these results are not as striking as for the last three digits, this is still an indication of some rounding.

Collected data was further examined to see if there was a difference between data gathered directly from the respondent and data obtained by proxy. For the wages and salary variable, the third last digit was a zero 90.7% of the time for collection by proxy

and 88.5% for direct collection. In the case of the last digit, we found that it was a zero in 97.6% of the proxy cases and 96.3% directly. Similar results were observed for the other digits considered. Therefore, we conclude that the respondents whether they were answering for themselves or providing a proxy response tended to round the reported amount of wages and salary.

Table 10.2 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 10.2 Proportion of zeroes in the last four digits declared for selected variables in 2008 (%)

Variable	Digit			
	Fourth-last	Third-last	Second-last	Last
Wages and salaries	30.0	83.1	94.1	96.2
Investment income	10.3	29.0	59.6	72.0
Social assistance	7.5	33.7	77.0	90.8
EI benefits	4.0	45.5	84.5	93.5
Non-farm self-employment income	36.4	85.8	96.5	96.6

These last results demonstrate the constant increase in the proportion of zeroes when proceeding from the fourth last digit to the last digit. For wages and salaries and non-farm self-employment income, there is a higher proportion of zeroes compared to the other variables; this is especially evident in the third last digit.

For investment income, social assistance and EI benefits, there is a strong increase in the proportion of zeroes when comparing the third last digit to the second last. These increases vary from 30.6% to 43.3%.

All variables had a zero in the last digit in at least 90% of the cases except for investment income (72.0%) but the proportion of zeroes here was still quite high.