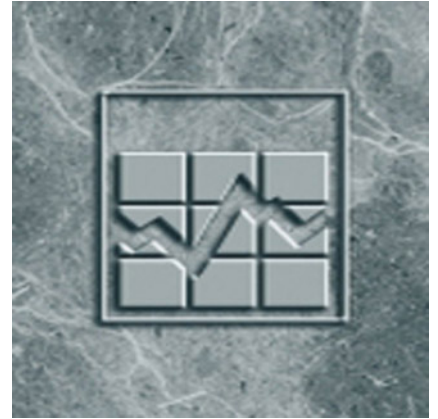# Research Paper

Income Research Paper Series

Data Quality in the 2001 Survey of Labour and Income Dynamics (SLID)

by Barbara Armstrong and Georgina House

Income Statistics Division and Social Survey Methods Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 613 951-7355

Statistics Canada   Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-7355; (888) 297-7355; income@statcan.ca).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 75F0002MIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Income Statistics Division

Income Research Paper Series

Data Quality in the 2001 Survey of Labour and Income Dynamics (SLID)

Note of appreciation

Abstract

Data Quality in the 2001 Survey of Labour and Income Dynamics

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated in 1993. The survey was designed to measure changes in the economic well-being of Canadians as well as the factors affecting these changes.

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 furnish users with measures of data quality so that the user is able to interpret the data properly. This report summarizes a set of quality measures that has been produced in an attempt to describe the overall quality of SLID data. Among the measures included in the report are sample composition and attrition rates, sampling errors, coverage errors in the form of slippage rates, response rates, tax permission and tax linkage rates, and imputation rates.

Table of contents

1. Introduction

The Survey of Labour and Income Dynamics (SLID) is a longitudinal survey initiated 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 2 panels. Each panel remains in the survey for 6 consecutive years and a new panel is rotated in every 3 years.   In January following the reference year, SLID sample households are contacted by telephone interviewers. Demographic information is collected for every person in the household.   Complete survey data are collected for every person in the household 16 years or older. Questions are asked on labour (labour market activity, work experience, jobless spells and job information), educational attainment and income sources. At the end of the January interview, respondents are informed that they will be contacted again in May when they will be asked to supply data on income as well as certain expense items. However, the respondent may elect to grant permission to Statistics Canada to retrieve all the data required from the T1 tax file, thereby avoiding the necessity of a second interview. Collection of income data is deferred until May so that the respondent will be more familiar with the required data (having just filed an income tax return).

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 cross-sectional sample persons.

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.ca:8096/bsolc/english/bsolc?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 furnish users with measures of data quality so that the user is able to interpret the data properly. This report summarizes quality measures that have been produced in an attempt to describe the overall quality of SLID data.

2. Sample composition/attrition

Although originally designed as a longitudinal survey, SLID also has the capability of producing cross-sectional estimates. Every non-longitudinal person living with a longitudinal respondent is also asked to participate in the survey. Such persons are called cohabitants. 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 that are ineligible cross-sectionally (such as deceased, institutionalized and those who have moved outside the country), and responding cohabitants. Historical tables can be found in appendix 1.

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 in SLID by province, 2001

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally[1]		Cohabitants		Cross-sectional sample size	
	Panel 03	Panel 04	Panel 03	Panel 04	Panel 03	Panel 04	Panel 03	Panel 04
	Persons							
N.F.	1,477	1,591	75	42	232	214	1,634	1,763
P.E.I.	1,005	1,014	57	27	195	99	1,143	1,086
N.S.	2,263	2,378	148	63	543	302	2,658	2,617
N.B.	2,024	2,214	109	53	435	306	2,350	2,467
Que.	6,341	6,825	324	217	1,348	871	7,365	7,479
Ont.	10,063	10,376	518	289	2,233	1,335	11,778	11,422
Man.	2,407	2,739	143	79	466	406	2,730	3,066
Sask.	2,087	2,785	140	87	494	421	2,441	3,119
Alta.	2,764	2,910	122	63	759	490	3,401	3,337
B.C.	2,813	3,075	157	95	601	386	3,257	3,366
Moved outside provinces	472	337	472	337	0	0	0	0
Total	33,716	36,244	2,265	1,352	7,306	4,830	38,757	39,722

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

Table 2.2 - Sample composition in SLID by CMA, 2001

Census metropolitan area	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally[1]		Cohabitants		Cohabitants sample size	
	Panel 03	Panel 04	Panel 03	Panel 04	Panel 03	Panel 04	Panel 03	Panel 04
	Persons							
Halifax	586	538	0	0	179	95	765	633
Quebec City	479	563	0	0	123	68	602	631
Montréal	1,333	1,291	0	0	357	205	1,690	1,496
Ottawa - Gatineau	855	935	0	0	203	135	1,058	1,070
Toronto	1,636	2,007	0	0	433	267	2,069	2,274
Hamilton	443	481	0	0	119	32	562	513
St. Catharines - Niagara	433	560	0	0	83	63	516	623
Kitchener	525	549	0	0	140	90	665	639
London	476	504	0	0	87	63	563	567
Windsor	290	387	0	0	71	58	361	445
Winnipeg	1,032	1,270	0	0	231	232	1,263	1,502
Calgary	623	667	0	0	194	135	817	802
Edmonton	751	724	0	0	190	128	941	852
Vancouver	933	1,067	0	0	201	154	1,134	1,221
Victoria	251	303	0	0	64	35	315	338
Other CMA or CA	10,594	11,840	0	0	2,456	1,670	13,050	13,510
Do not live in a CMA	10,211	11,206	0	0	2,175	1,400	12,386	12,606
Not available[2]	2,265	1,352	2,265	1,352	0	0	0	0
Total	33,716	36,244	2,265	1,352	7,306	4,830	38,757	39,722

1. This includes individuals who are deceased, institutionalized and those who have moved outside the country.
2. This information is only available for those individuals who are cross-sectionally eligible.

Table 2.3 - Status of longitudinal persons, reference year 2001

Longitudinal status	Panel 03	Panel 04
In scope (respondents)	31,451	34,892
In scope (nonrespondents)	1,912	4,292
Moved to Yukon, NWT, Nunavut	12	14
Moved outside Canada	457	322
Institutionalized	369	255
Deceased	1,427	761
Removed from sample[1]	7,916	3,147
Duplicate person/error[2]	3	0
Total	43,547	43,683

1. Respondents are removed from the sample for one of two reasons. If entire households have refused for 2 consecutive cycles they are said to be hard refusals and no further attempts are made to enumerate these households. As well after two years households that were not traced are not sent out for further attempts at collection.
2. Respondents who were erroneously included in the household in the beginning of the first year of a panel's existence.

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 selected SLID variables

Variable	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Canada
Average total income	3.1	2.8	1.8	1.8	1.5	1.5	1.7	1.9	1.9	2.1	0.8
Average market income	4.1	3.3	2.1	2.4	1.8	1.6	2.0	2.5	2.1	2.4	0.9
Average wages and salaries	4.7	3.2	2.4	2.6	2.1	1.7	2.2	2.9	2.1	2.4	1.0
Average EI benefits	4.1	8.0	4.9	7.0	3.4	3.4	6.9	6.0	5.9	5.7	1.8
Average social assistance	5.8	12.0	5.0	4.6	2.8	3.9	8.8	6.8	9.5	6.0	2.0
Average other income	11.3	13.1	9.0	9.4	8.6	6.2	11.1	13.3	11.2	13.8	3.8
Prevalence of persons under LICO (after tax)	10.1	19.4	10.0	9.3	4.9	5.0	8.2	7.5	8.0	6.4	2.6
Counts of employed people	1.9	2.3	1.2	1.4	1.1	0.7	1.3	1.2	1.0	1.3	0.4

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.

Table 4.1 shows slippage rates at the person level by province and by age/sex groupings.

As a comparison we will look at the person level slippage rates for the Labour Force
Survey (LFS) by province. We will look at the slippage rates from the LFS at the
beginning of the panel for each panel in tables 4.2 and 4.3. These rates are the rates
associated with the rotation groups used by SLID.

Slippage rates were also computed at the household level and are summarized in Table
4.4. For slippage rates for previous reference years, see Appendix 2.

Table 4.1 - Person level slippage rates (%) by province/sex/age group

Sex	Age Group	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Total
Male	0-6	1.1	15.2	4.3	10.3	11.4	16.3	-6.4	2.7	15.2	18.7	13.3
	7-15	5.1	21.8	10.6	12.0	14.9	25.9	20.6	11.4	22.6	36.7	22.6
	16-17	-32.7	6.8	7.8	-19.9	-5.1	0.2	4.2	-9.4	20.8	4.5	0.8
	18-24	8.7	12.5	16.7	-10.1	2.3	13.7	0.4	0.2	24.3	29.5	12.7
	25-34	20.3	28.3	12.2	17.3	21.6	27.1	6.9	14.6	30.6	28.1	24.6
	35-44	-2.6	10.0	8.3	7.4	18.1	17.8	10.1	2.2	21.7	28.3	18.1
	45-54	-0.1	8.9	0.2	14.7	6.6	7.5	0.9	5.6	16.1	22.2	9.7
	55-59	-1.6	-1.9	-1.0	0.4	7.5	10.4	-6.1	-2.1	17.3	28.1	10.8
	60-64	31.4	26.6	13.9	21.8	-2.5	12.4	15.1	25.1	17.4	19.3	11.1
	65-69	1.7	-4.3	-1.6	16.2	12.0	14.4	-12.1	11.9	10.0	12.6	11.4
	70+	25.4	4.6	6.4	4.9	-0.0	12.4	7.5	8.0	6.7	9.5	7.8
	Total	6.1	13.5	7.5	7.9	10.7	16.6	5.5	6.5	20.5	24.6	15.3
Female	0-6	6.7	1.8	3.1	-0.1	-6.4	10.0	1.8	-6.2	15.1	28.4	7.8
	7-15	2.6	6.1	17.0	6.2	25.8	25.0	21.8	14.1	36.6	29.8	25.4
	16-17	-18.5	1.9	-2.0	-21.0	2.1	-0.6	3.7	7.1	21.5	20.2	4.7
	18-24	-3.1	16.7	9.9	-10.7	-0.3	12.0	-4.2	4.6	19.5	16.0	8.7
	25-34	15.7	19.8	7.9	17.7	10.3	22.8	3.1	-4.6	25.1	17.5	17.4
	35-44	-5.4	11.3	5.1	-5.5	10.8	13.8	1.5	-2.0	21.9	25.2	13.5
	45-54	-4.0	5.0	-1.2	4.7	2.2	5.8	3.2	3.7	11.4	16.6	6.3
	55-59	-3.8	-7.1	3.9	13.4	8.6	6.3	-6.9	5.4	4.0	13.2	7.0
	60-64	16.5	7.1	-2.1	3.6	6.9	18.7	-9.6	4.5	10.7	9.0	11.2
	65-69	1.4	12.6	11.4	-8.5	-6.3	11.6	-25.9	12.7	11.9	4.6	4.2
	70+	18.0	-2.5	2.1	8.7	1.6	8.2	0.2	0.0	-0.6	12.7	5.8
	Total	2.6	7.9	5.5	2.9	6.7	13.6	1.9	2.5	18.6	19.2	11.7
Total		10.0	4.3	10.7	6.5	5.4	8.7	15.1	3.7	4.5	19.6	21.9

Table 4.2 - Person level slippage rates (%) by province of SLID sample coming from the LFS, panel 2

	NL	PEI	NS	NB	Que	Ont	Man	Sask	Alta	BC	Canada
Total	4.5	6.4	3.7	4.2	2.7	7.7	1.0	-2.0	9.5	13.8	6.6

Table 4.3 - Person level slippage rates (%) by province of SLID sample coming from the LFS, panel 3

	NL	PEI	NS	NB	Que	Ont	Man	Sask	Alta	BC	Canada
Total	10.4	10.2	10.5	2.2	6.0	6.6	5.7	6.5	14.1	14.2	8.2

Table 4.4 - Household level slippage rates (%) for provinces by household size

Province	Households			All
	size 1	size 2	size 3+	
N.F.	-20.0	-1.7	9.3	0.7
P.E.I.	-2.7	0.8	15.5	6.5
N.S.	-17.1	7.5	9.7	2.6
N.B.	-9.8	0.7	10.1	2.5
Que.	3.4	5.8	10.3	6.8
Ont.	2.8	16.3	15.1	12.5
Man.	-14.5	-3.2	7.2	-2.2
Sask.	-17.7	3.7	6.6	-1.1
Alta.	-7.6	10.4	23.6	11.9
B.C.	-1.0	14.7	28.3	15.9
Canada	-0.9	10.6	15.6	9.7

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 May interview, 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 nonrespondent is not straightforward.

If all persons in non-responding January households are also nonrespondent in May, then these persons (and households) are nonrespondent.

For those persons in non-responding January households 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 nonrespondent.

Persons in non-responding January households for whom we have May data and for whom the household has not changed since the previous year, are considered nonrespondents if the household was a non-responding household in the previous January. Since updates to household composition are collected in January, this means that the household composition has not been updated for 2 consecutive years. Persons in households that have been nonrespondent in 2 consecutive January collections are therefore considered to be nonrespondents to SLID.

The person level response rates are calculated by dividing the number of cross-sectionally eligible respondents to the January and/or May interviews by the total number of cross-sectionally eligible people. An assumption is made that nonrespondents are still in the target population unless there is evidence to the contrary. As a result this may somewhat underestimate response rates.

Table 5.1 - Cross-sectional person response rates (%) (age>15)

Type	1993	1994	1995	1996	1997	1998	1999	2000	2001
Response	92.2	88.4	85.2	85.4	83.4	81.8	81.6	78.0	77.8
Non-response	7.8	11.6	14.8	14.6	16.6	18.2	18.4	21.9	22.2

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 responding households by the total number of cross-sectionally eligible households. Once again an assumption is made; non-responding 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.

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

Figure 5.1 shows the household response rates by province.

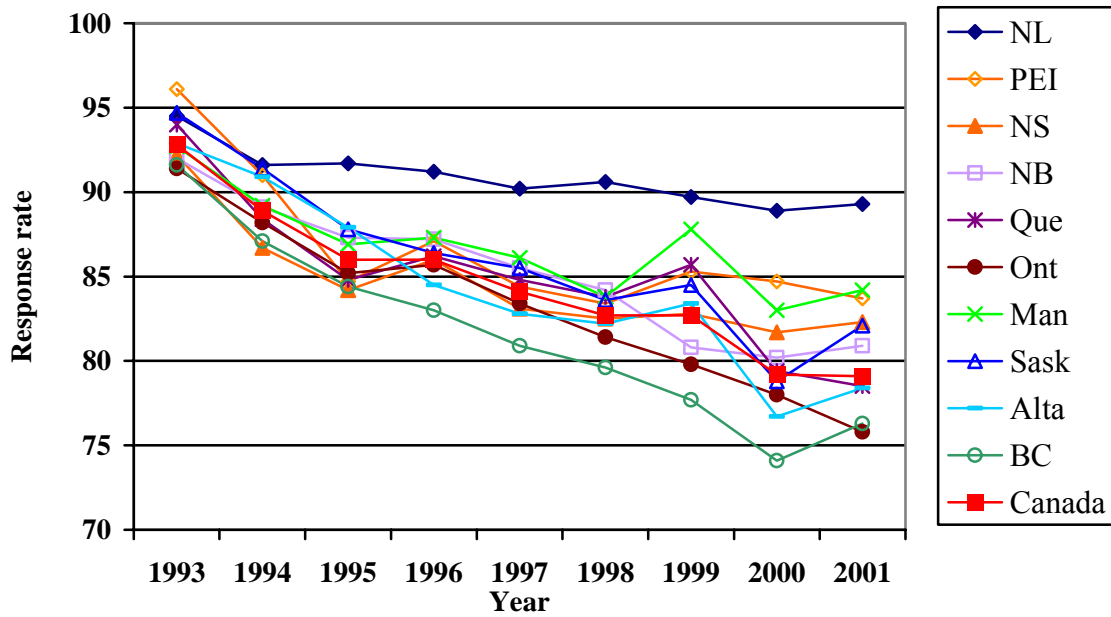**Figure 5.1 - Cross-sectional household response rates by province (%)**

Table 5.2 shows the person response rates by phase. 'Respondent to labour interview' and 'Respondent to income interview' are the percentages of those who responded to only the labour (January) or income (May) interviews respectively whereas the 'Respondent to both interviews' is the percentage of all those who responded in full or in part to both interviews.

**Table 5.2 - Cross-sectional person response rates by phase (%)**

Type	1993	1994	1995	1996	1997	1998	1999	2000	2001
Respondent to labour interview	10.3	10.5	10.0	10.8	12.2	10.4	13.6	17.3	10.4
Respondent to income interview	6.2	2.8	3.3	2.9	2.2	2.6	2.5	4.6	4.1
Respondent to both interviews	75.7	75.1	71.8	71.6	69.0	68.8	65.6	56.2	63.3
Non-response	7.8	11.6	14.8	14.6	16.6	18.2	18.4	21.9	22.2

Due to the conceptual difficulty in defining a longitudinal household, only person level longitudinal response rates are calculated. Table 5.3 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 that panel. Figure 5.2 shows the longitudinal non-response by panel and wave.

Table 5.3 - Longitudinal person response rates (%) (all ages)

	1993	1994	1995	1996	1997	1998	1999	2000	2001
Panel 1	93.3	89.6	86.5	83.9	82.4	81.5	n/a	n/a	n/a
Panel 2	n/a	n/a	n/a	89.5	86.7	85.2	82.7	78.5	77.4
Panel 3	n/a	n/a	n/a	n/a	n/a	n/a	83.9	83.0	83.0
Panel 4	93.3	89.6	86.5	83.9	82.4	81.5	n/a	n/a	n/a

Figure 5.2: Longitudinal non-response by wave
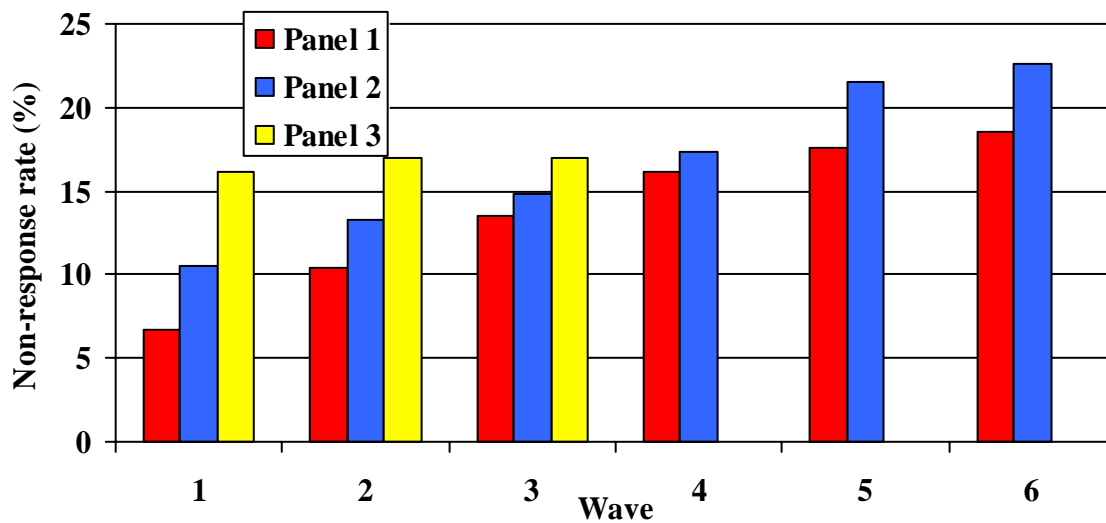
6. Tax permission rates

There are two interviews every year: in January the interview is about activities such as working, going to school, looking for work or retirement. The second interview in May is about income, but it is not necessary if the respondent gives 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 is asked for this permission at the end of the January interview. If permission is not given, the respondent is contacted again in May. At this time the respondent is once again asked if he/she would prefer to give permission to access tax records. If permission is not provided, the interview proceeds.

Table 6.1 shows permission rates by panel for each phase of 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 the table are based on the number of respondents over the age of 15 who are cross-sectionally eligible. Permission rates in reference year 2000 are the same for both January and May because there was no May interview in that year.

Table 6.1 - SLID permission rates (%) by panel

Panel (start date)	Wave											
	1 Interview		2 Interview		3 Interview		4 Interview		5 Interview		6 Interview	
	Jan	May	Jan	May	Jan	May	Jan	May	Jan	May	Jan	May
Panel 1 (1993)	n/a	n/a	n/a	58.5[1]	69.7	75.4	78.8	82.4	83.5	85.1	86.1	88.3
Panel 2 (1996)	53.2	68.9	76.0	79.3	83.3	86.3	85.8	88.2	87.5	87.5	87.0	90.3
Panel 3 (1999)	55.0	71.7	76.4	76.4	78.2	84.7						

1. Permission was asked for the first time in May 1994

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

---

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.

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. Figure 7.2 gives tax linkage rates for those in the SLID sample for which we were successful in finding a SIN.
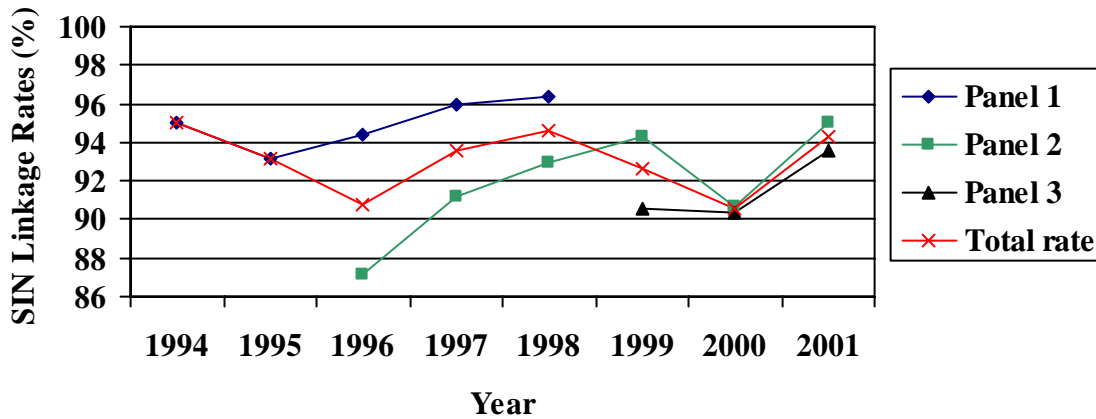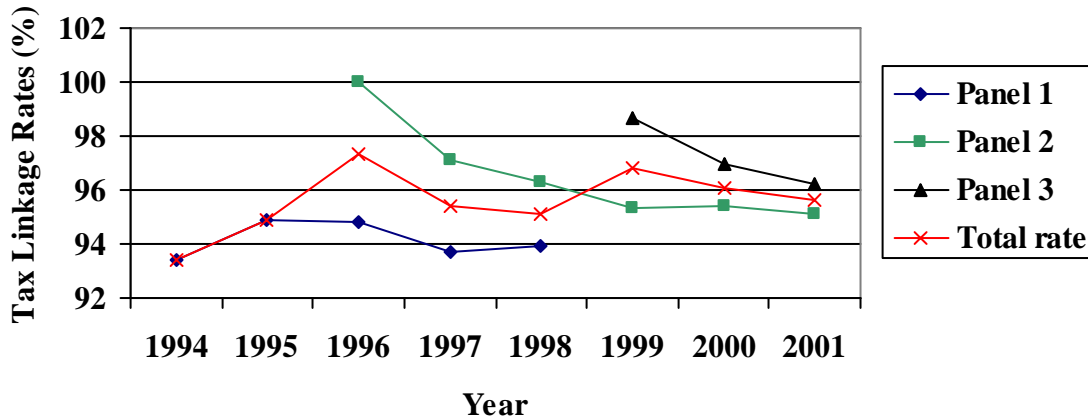
Figure 7.1 Percentage of SINs found

Figure 7.2 SLID/tax linkage rates

8. Imputation rates

To compensate for non-responding 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 responding 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.

Imputation of income variables in SLID is done using a nearest neighbour approach. 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 score function used in SLID income imputation is:

$$s(X,Y) = \sum_{k=1}^{K} p_k \, \mathrm{I}(X_k, Y_k), \quad \text{where } \mathrm{I}(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ 0 & \text{if not.} \end{cases}$$

Note that p_k is a number allowing us to assign more or less importance to the matching variable k. X_k is the value of the receiver's variable k and Y_k is the value of the donor's variable k.

The distance function used in SLID income imputation is the same as the function used in the generalized edit and imputation system (GEIS). Suppose we have two records X and Y. The distance between the two is defined as:

$$d(X,Y) = \max_{j=1}^{J} \left| u_j(X) - u_j(Y) \right|$$

Where $u_j(X)$ is the function of the rank[3] of X_j :

$$u_j(X) = \frac{rank(X_j)}{n_j + 1}.$$

J represents the number of quantitative variables used to calculate the distance, X_j represents the value of the quantitative variable j of the record X and n_j is the number of records with a valid value for this variable. When several records have the same value of the variable j, they are assigned a mean rank. Excluding these cases of equality the $u_j(\cdot)$ are uniformly distributed along the interval (0, 1).

The percentage of persons within responding SLID households that were subject to total or partial imputation is shown in Table 8.1. Recall that a responding SLID household is

---

3. The rank is a method by which a numeric variable can be normalized. This way a numeric variable with a range from 0 to 9 and a numeric variable with a range from -999999 to 999999 have the same level of importance in the distance function.

one in which at least one household member has responded partially or completely in the January and/or the May component of the survey.

Table 8.2 compares the proportion of records from tax to those collected in the telephone interview.

In total eighteen income variables are imputed during SLID income imputation. Many individuals require only partial imputation. Partial imputation is when one or more income items is imputed with some information being supplied by the individual. In table 8.3 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 from both the January and May interviews. Inconsistencies are corrected through the imputation process. Table 8.3 also shows the percentage of individuals subject to partial imputation who require between one and seventeen variables to be imputed.

Table 8.1 - Persons requiring imputation of income variables, by province in 2001

Province	Total imputation[1]	Partial imputation [2]	No imputation
		(%)	
N.F	0.4	19.8	79.9
P.E.I.	0.9	22.5	76.6
N.S.	1.2	23.2	75.7
N.B.	1.0	22.3	76.8
Que.	0.9	21.7	77.4
Ont.	1.5	23.9	74.6
Man.	1.0	22.5	76.6
Sask.	0.9	23.2	75.9
Alta.	1.2	22.1	76.7
B.C.	1.1	25.1	73.8
Canada	1.1	22.9	76.0

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.

Table 8.2: Proportion of respondents coming from tax or interview

	Tax	Interview	Other[1]
		(%)	
Proportion	78.9	5.0	16.1

1. This comprises records that are not linked to Tax and without a May (Income) interview.

Table 8.3: Tax records and interview records requiring partial or total imputation

Record type→	Tax records	Interview	Other[1]	All
		(%)		
No imputation	92.6	59.2	n/a	76.0
Total imputation (18 variables)	n/a	n/a	6.9	1.1
Partial imputation	7.4	40.8	93.1	22.9
1 variable imputed	7.2	16.4	0.0	6.5
2 – 9 variables imputed	0.2	21.0	0.0	1.2
10 – 17 variables imputed	0.0	3.4	93.1	15.2

1. Records that are not linked to Tax and without a May (Income) interview. Some of these records are partially imputed based on the information collected during the January interview

9. Rounding of income data

A small percentage of SLID income data comes from data collected in a telephone interview in May. 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 refer to tax forms and 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 9.2 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 9.1 - Distribution of the last 4 digits of wages and salaries for collected data versus tax data (greater than 10,000)

Digit	Fourth last digit		Third last digit		Second last digit		Last digit	
	Collected	Tax	Collected	Tax	Collected	Tax	Collected	Tax
	(%)							
0	26.8	11.8	80.3	11.3	86.7	13.0	88.1	14.1
1	6.1	10.9	1.5	10.1	1.2	9.9	1.1	9.4
2	10.0	10.7	2.4	9.9	1.8	9.4	1.6	9.7
3	8.1	10.2	2.3	9.8	1.3	9.3	1.5	9.4
4	8.2	9.9	2.3	10.0	1.2	9.6	1.4	9.6
5	15.4	9.4	3.1	9.7	1.9	9.8	1.2	9.8
6	6.8	9.9	2.2	10.2	1.2	9.6	1.3	9.6
7	6.2	9.2	2.0	9.5	1.8	9.7	1.1	9.3
8	7.7	9.0	2.3	9.7	1.5	9.9	1.2	9.4
9	4.7	9.0	1.6	9.8	1.4	9.8	1.5	9.7

Table 9.2 - Prevalence of zeroes in the last 4 digits of reported data for selected variables

Variable	Fourth last digit	Third last digit	Second last digit	Last digit
	(%)			
Wages and salaries	21.4	73.4	85.5	87.6
Investment income	5.7	23.8	47.2	65.0
Social assistance	12.6	37.4	56.5	67.6
UI Benefits	3.0	48.0	75.6	82.7
Non-farm self-employment income	27.1	66.1	81.5	83.6

Appendix 1: Sample composition in SLID by province, 1996 - 2000

1996 Sample composition (persons)

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally		Cohabitants		Cross-sectional sample size	
	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02
N.F.	2039	1692	74	18	290	103	2255	1777
P.E.I.	751	1180	33	5	125	56	843	1231
N.S.	2300	2620	73	26	375	148	2602	2742
N.B.	2118	2441	62	21	322	168	2378	2588
Que.	6146	7537	238	59	923	360	6831	7838
Ont.	9046	11972	335	84	1557	682	10268	12570
Man.	2245	2754	87	19	387	181	2545	2916
Sask.	2415	2468	124	25	373	222	2664	3112
Alta.	3156	2915	89	25	695	222	3751	3112
B.C.	2998	3280	71	27	563	227	3490	3480
Moved outside provinces	149	126	149	126	0	0	0	0
Total	33,352	38,985	1,335	435	5,610	2,312	37,627	40,862

1997 Sample composition (persons)

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally		Cohabitants		Cross-sectional sample size	
	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02
N.F.	1998	1624	87	35	312	148	2223	1737
P.E.I.	734	1120	41	13	140	90	833	1197
N.S.	2234	2500	98	38	410	264	2546	2726
N.B.	2068	2308	79	36	369	258	2358	2530
Que.	6070	7325	270	102	1104	664	6904	7887
Ont.	8831	11550	395	181	1841	1196	10277	12565
Man.	2193	2687	105	48	434	288	2522	2927
Sask.	2368	2406	147	47	436	247	2657	2606
Alta.	3137	2862	102	51	870	397	3905	3208
B.C.	2929	3161	98	60	598	357	3429	3458
Moved outside provinces	196	337	196	337	0	0	0	0
Total	32,758	37,742	1,618	810	6,514	3,909	37,654	40,841

1998 Sample composition (persons)

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally		Cohabitants		Cross-sectional sample size	
	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02
N.F.	1961	1588	89	45	362	153	2234	1696
P.E.I.	708	1076	54	29	149	127	803	1174
N.S.	2206	2456	118	67	484	339	2572	2728
N.B.	2026	2250	97	58	447	287	2376	2479
Que.	6007	7198	310	143	1268	865	6965	7920
Ont.	8682	11253	442	268	2057	1427	10297	12412
Man.	2130	2603	127	72	461	333	2464	2864
Sask.	2318	2332	155	75	470	314	2633	2571
Alta.	3123	2900	97	65	972	539	3998	3374
B.C.	2895	3084	125	78	656	413	3426	3419
Moved outside provinces	472	346	472	346	0	0	0	0
Total	32,394	37,086	1,952	1,246	7,326	4,797	37,768	40,637

1999 Sample composition (persons)

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally		Cohabitants		Cross-sectional sample size	
	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02
N.F.	1550	1578	55	15	179	83	1674	1646
P.E.I.	1065	1005	36	8	165	31	1194	1028
N.S.	2384	2282	102	19	375	136	2657	2399
N.B.	2159	2110	68	15	336	113	2427	2208
Que.	7017	7309	216	86	1048	272	7849	7495
Ont.	10758	10510	347	110	1723	482	12134	10882
Man.	2573	2843	93	27	398	136	2878	2952
Sask.	2265	2783	94	35	369	205	2540	2953
Alta.	2871	2995	85	22	612	208	3398	3181
B.C.	2988	3114	108	34	468	203	3348	3283
Moved outside provinces	375	130	375	130	0		0	
Total	36,005	36,659	1,579	501	5,673	1,869	40,099	38,027

2000 Sample composition (persons)

Province	Longitudinal sample size		Longitudinal sample ineligible cross-sectionally		Cohabitants		Cross-sectional sample size	
	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02	Panel 01	Panel 02
N.F.	1495	1591	66	22	200	129	1629	1698
P.E.I.	1031	1024	46	17	162	71	1147	1078
N.S.	2274	2351	130	36	441	200	2585	2515
N.B.	2060	2194	91	29	359	210	2328	2375
Que.	6493	6970	270	158	1179	526	7402	7338
Ont.	10302	10671	418	191	1913	853	11797	11333
Man.	2402	2747	120	48	409	244	2691	2943
Sask.	2121	2664	116	58	414	268	2419	2874
Alta.	2735	2815	105	40	620	292	3250	3067
B.C.	2809	2977	136	65	513	279	3186	3191
Moved outside provinces	446	235	446	235	0	0	0	0
Total	34,168	36,239	1,944	899	6,210	3,072	38,434	38,412

Appendix 2: Slippage rates over time

Slippage rates (%) by province and year, 1996 to 2000.

Year	N.L.	P.E.I.	N.S.	N.B.	Que.	Ont.	Man.	Sask.	Alta.	B.C.	Canada
1996	4.4	8.0	4.2	4.7	7.0	11.1	7.9	3.8	13.4	17.6	10.3
1997	4.4	8.3	5.2	5.6	7.0	12.2	6.9	3.1	14.9	19.6	11.1
1998	5.3	11.0	5.5	6.5	7.1	13.2	8.9	3.2	15.8	20.2	11.8
1999	6.4	11.0	8.7	7.1	7.3	13.1	3.5	6.0	17.7	19.5	12.0
2000	3.3	10.2	8.4	5.6	7.7	14.5	3.8	6.3	19.0	19.1	12.6
2001	4.3	10.7	6.5	5.4	8.7	15.1	3.7	4.5	19.6	21.9	13.4

Slippage rates (%) for province by household size, 1996

Province	Households			
	Size 1	Size 2	Size 3+	All
N.F.	10.4	21.5	-4.0	6.9
P.E.I.	20.1	12.5	9.3	12.9
N.S.	-0.6	18.5	1.9	7.1
N.B.	13.4	11.0	2.0	7.6
Que.	10.6	14.3	5.8	10.0
Ont.	15.8	20.3	14.1	16.4
Man.	-0.1	17.1	5.8	7.8
Sask.	-18.2	16.0	0.2	0.3
Alta.	12.8	22.8	19.2	18.9
B.C.	12.3	21.5	21.8	19.2
Canada	11.2	18.6	11.9	13.9

Slippage rates (%) for provinces by household size, 1997

Province	Households			
	Size 1	Size 2	Size 3+	All
N.F.	-6.3	-2.0	1.9	-0.5
P.E.I.	7.8	1.9	5.0	4.7
N.S.	-16.4	5.0	3.5	-0.6
N.B.	-1.8	-3.6	4.4	0.5
Que.	-5.5	4.9	4.0	1.6
Ont.	5.7	13.0	7.5	8.8
Man.	-8.4	13.2	2.3	2.8
Sask.	-20.9	8.9	-3.0	-4.0
Alta.	-4.9	9.7	11.7	7.2
B.C.	4.2	18.0	17.9	14.4
Canada	-1.1	10.2	7.5	6.2

Slippage rates (%) for province by household size, 1998

Province	Households			
	Size 1	Size 2	Size 3+	All
N.F.	-15.1	3.0	7.7	2.6
P.E.I.	0.9	4.8	13.4	7.9
N.S.	-24.7	7.0	9.1	0.5
N.B.	-1.7	1.4	7.6	3.6
Que.	-3.0	6.3	8.1	4.3
Ont.	3.5	12.8	12.6	10.5
Man.	-8.2	9.8	10.6	5.2
Sask.	-27.2	14.2	0.5	-2.6
Alta.	-13.6	14.5	16.0	8.6
B.C.	-3.8	17.2	23.5	14.2
Canada	-3.6	11.2	12.4	8.0

Slippage rates (%) for province by household size, 1999

Province	Households			
	Size 1	Size 2	Size 3+	All
N.F.	-19.1	8.4	8.2	3.8
P.E.I.	-6.9	1.4	16.2	6.2
N.S.	-14.4	8.2	11.6	4.3
N.B.	1.6	3.3	8.9	5.4
Que.	-0.7	6.1	8.3	4.9
Ont.	4.0	13.8	12.7	11.0
Man.	-10.2	2.4	6.1	-1.1
Sask.	-21.7	7.7	7.4	-0.5
Alta.	-9.0	11.4	19.1	10.1
B.C.	-3.5	13.2	24.4	13.3
Canada	-1.9	10.1	13.2	8.3

Slippage rates for province by household size, 2000

Province	Households			
	Size 1	Size 2	Size 3+	All
N.F.	-31.7	-4.3	6.2	-3.5
P.E.I.	-9.0	-0.8	13.4	3.7
N.S.	-13.8	3.8	11.6	2.9
N.B.	-3.8	-1.5	10.7	3.4
Que.	4.8	5.4	9.1	6.7
Ont.	0.8	14.4	14.5	11.2
Man.	-10.1	-4.8	9.5	-0.5
Sask.	-17.7	5.1	6.9	-0.5
Alta.	-3.6	14.5	27.2	11.7
B.C.	-3.6	14.5	27.2	14.8
Canada	-1.1	9.6	14.8	9.0