Catalogue no. 75F0002MIE — No. 003
ISSN: 1707-2840
ISBN: 978-0-662-45959-0

Research Paper

Income Research Paper Series

# Data Quality in the 2005 Survey of Labour and Income Dynamics (SLID)



by Christopher Duddek

Income Statistics Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 1-613-951-7355

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 at www.statcan.ca.

| | |
|---|---|
| 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 in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which 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 us > Providing services to Canadians.

Statistics Canada
Income Statistics Division

**Income Research Paper Series**

# Data Quality in the 2005 Survey of Labour and Income Dynamics (SLID)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2007

All rights reserved. The content of this electronic publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it be done solely for the purposes of private study, research, criticism, review or newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form, by any means—electronic, mechanical or photocopy—or for any purposes without prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

May 2007

Catalogue no. 75F0002MIE, vol. 3
ISSN 1707-2840
ISBN 978-0-662-45959-0

Frequency: occasional

Ottawa

La version française de cette publication est disponible sur demande (nº 75F0002MIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

**Table of contents**

1. Introduction ..... 5
2. Sample composition/attrition ..... 6
3. Sampling errors ..... 8
4. Coverage errors ..... 9
5. Response rates ..... 11
6. Tax permission rates ..... 14
7. Tax linkage rates ..... 15
8. Imputation rates ..... 17
9. Rounding of income data ..... 20

# 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 declines to 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*", 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 can interpret the data properly. This report summarizes these quality measures for SLID.

# 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 persons by province, reference year 2005**

| | 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 | 1,305 | 1,468 | 57 | 11 | 159 | 87 | 1,407 | 1,544 |
| Prince Edward Island | 906 | 879 | 36 | 9 | 138 | 64 | 1,008 | 934 |
| Nova Scotia | 2,114 | 1,975 | 101 | 28 | 325 | 155 | 2,338 | 2,102 |
| New Brunswick | 1,793 | 1,893 | 83 | 20 | 287 | 144 | 1,997 | 2,017 |
| Quebec | 5,989 | 5,927 | 285 | 76 | 989 | 431 | 6,693 | 6,282 |
| Ontario | 9,189 | 9,689 | 412 | 165 | 1,285 | 587 | 10,062 | 10,111 |
| Manitoba | 2,234 | 2,352 | 116 | 36 | 310 | 133 | 2,428 | 2,449 |
| Saskatchewan | 2,251 | 2,495 | 118 | 36 | 329 | 165 | 2,462 | 2,624 |
| Alberta | 2,631 | 3,355 | 97 | 33 | 496 | 282 | 3,030 | 3,604 |
| British Columbia | 2,933 | 3,088 | 128 | 49 | 452 | 189 | 3,257 | 3,228 |
| Not in a province | 331 | 248 | 331 | 248 | 0 | 0 | 0 | 0 |
| **Total** | **31,676** | **33,369** | **1,764** | **711** | **4,770** | **2,237** | **34,682** | **34,895** |

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.

## Table 2.2 Sample composition in SLID by CMA, reference year 2005

| Census Metropolitan Area | Longitudinal sample size | | Number of Cohabitants | | Cross-sectional sample size | |
|---|---|---|---|---|---|---|
| | Panel 4 | Panel 5 | Panel 4 | Panel 5 | Panel 4 | Panel 5 |
| Halifax | 450 | 586 | 84 | 57 | 534 | 643 |
| Quebec City | 424 | 425 | 126 | 38 | 550 | 463 |
| Montréal | 1,220 | 1,190 | 210 | 118 | 1,430 | 1,307 |
| Ottawa - Gatineau | 823 | 875 | 134 | 58 | 957 | 931 |
| Toronto | 1,548 | 1,745 | 257 | 127 | 1,805 | 1,872 |
| Hamilton | 388 | 446 | 65 | 24 | 453 | 470 |
| St. Catharines - Niagara | 454 | 447 | 70 | 21 | 524 | 468 |
| Kitchener | 445 | 462 | 59 | 35 | 504 | 497 |
| London | 400 | 498 | 55 | 38 | 455 | 536 |
| Windsor | 317 | 376 | 54 | 16 | 371 | 392 |
| Winnipeg | 1,008 | 1,233 | 165 | 87 | 1,173 | 1,320 |
| Calgary | 611 | 721 | 126 | 52 | 737 | 773 |
| Edmonton | 598 | 1,016 | 130 | 88 | 728 | 1,104 |
| Vancouver | 1,004 | 1,071 | 159 | 51 | 1,163 | 1,122 |
| Victoria | 298 | 308 | 49 | 25 | 347 | 333 |
| Other CMA or CA | 10,854 | 11,469 | 1,746 | 805 | 12,600 | 12,270 |
| Do not live in a CMA | 9,070 | 9,797 | 1,281 | 597 | 10,351 | 10,394 |
| Not available[1] | 1,764 | 704 | 0 | 0 | 0 | 0 |
| **Total** | **31,676** | **33,369** | **4,770** | **2,237** | **34,682** | **34,895** |

0 true zero or a value rounded to zero

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

## Table 2.3 Status of longitudinal persons, reference year 2005

| Longitudinal Status | Panel 4 | Panel 5 |
|---|---|---|
| In scope (respondents) | 29,912 | 32,665 |
| In scope (nonrespondents) | 2,568 | 8,958 |
| Moved to Yukon, NWT, Nunavut | 5 | 5 |
| Moved outside Canada | 324 | 241 |
| Institutionalized | 467 | 186 |
| Deceased | 968 | 272 |
| Removed from sample[1] | 7,953 | 3 |
| Duplicate person/error[2] | 35 | 0 |
| **Total** | **42,232** | **42,330** |

0 true zero or a value rounded to zero

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 after two years, 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.

# 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 (%), 2005**

| 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean total income | 1.6 | 1.8 | 1.4 | 1.2 | 1.1 | 1.2 | 1.6 | 1.6 | 1.3 | 1.6 | 0.6 |
| Mean market income | 2.0 | 2.6 | 1.5 | 1.5 | 1.3 | 1.4 | 1.8 | 1.9 | 1.5 | 1.8 | 0.7 |
| Mean wages and salaries | 2.1 | 2.2 | 1.5 | 1.4 | 1.0 | 1.4 | 1.4 | 1.2 | 1.3 | 1.8 | 0.7 |
| Mean EI benefits | 4.3 | 4.9 | 6.7 | 4.4 | 3.6 | 3.8 | 6.5 | 7.5 | 6.8 | 5.9 | 1.8 |
| Mean social assistance | 5.8 | 11.1 | 6.5 | 6.1 | 4.2 | 3.5 | 9.3 | 9.9 | 7.1 | 5.0 | 2.2 |
| Mean other income | 24.2 | 16.0 | 7.6 | 12.0 | 8.4 | 7.3 | 11.8 | 9.8 | 8.2 | 14.7 | 4.4 |
| Number under LICO after tax | 9.3 | 13.5 | 6.7 | 6.5 | 3.9 | 3.4 | 5.6 | 5.5 | 5.6 | 5.1 | 1.9 |
| Counts of employed people | 1.6 | 2.2 | 1.0 | 1.4 | 1.1 | 0.9 | 1.2 | 1.1 | 1.2 | 1.3 | 0.5 |

# 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

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

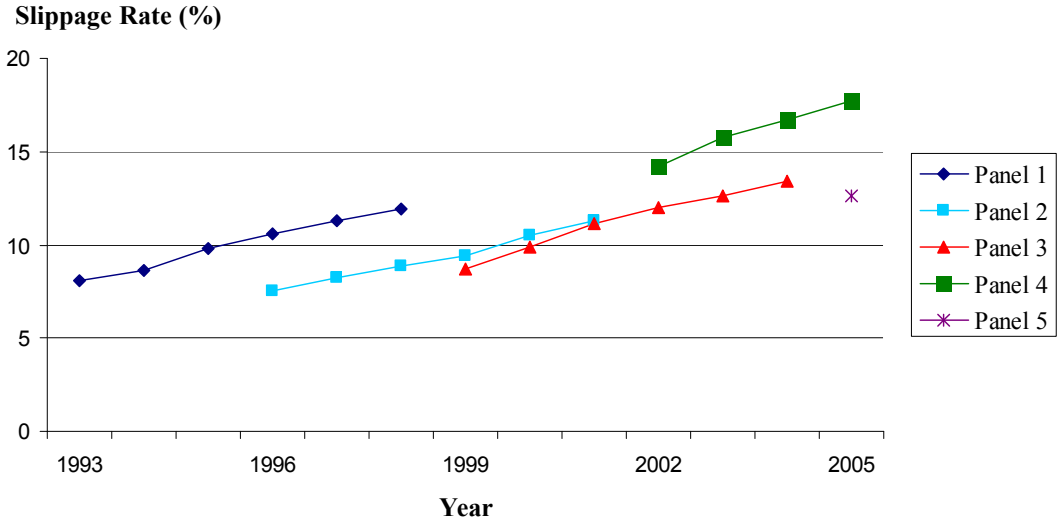
where Class C is the group or class for which we want to calculate slippage rates. For example at a detailed level the groups are based on province, sex and age group.
$CP_C$ is the census population projection for class C
$w_{kc}$ is the survey weight for $k_{th}$ responding unit in class C
$S_C$ is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

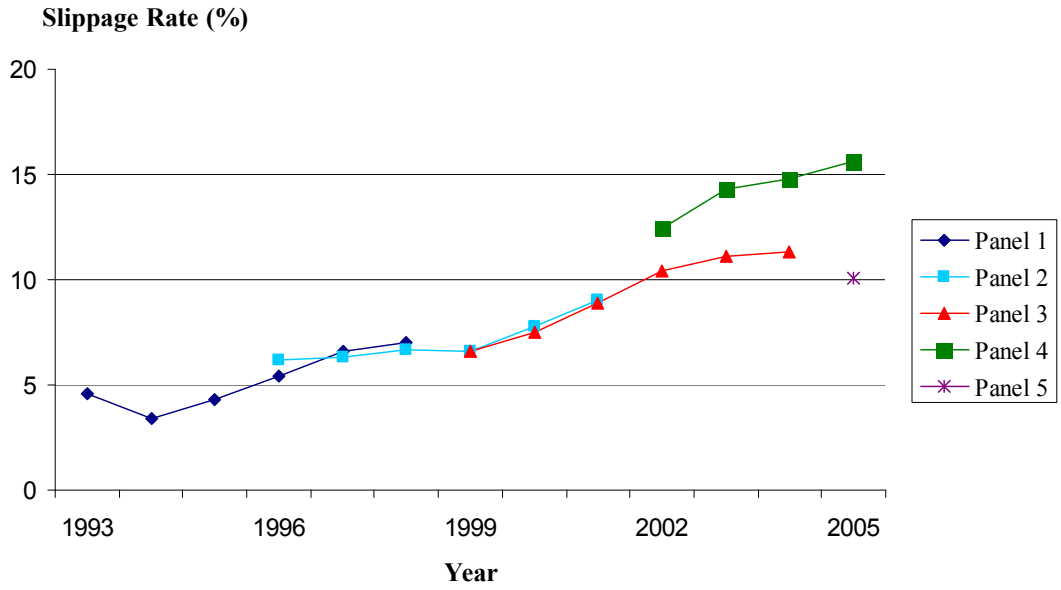
**Figure 4.1: Person-level slippage rate (%)**



## Table 4.1 National and provincial slippage rates by panel for 2005 (%)

| | N.L. | P.E.I. | N.S. | N.B. | Que. | Ont. | Man. | Sask. | Alta. | B.C. | Canada |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Panel 4 | 10.1 | 5.7 | 12.0 | 12.5 | 14.2 | 18.9 | 15.4 | 11.0 | 19.4 | 24.9 | 17.7 |
| Panel 5 | 2.9 | 7.0 | 6.4 | 3.6 | 5.6 | 18.0 | 2.6 | 0.5 | 18.5 | 14.5 | 12.7 |

## Figure 4.2: Household slippage rate (%)




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

| | Panel 4 Household Size | | | | Panel 5 Household Size | | | |
|---|---|---|---|---|---|---|---|---|
| Province | 1 | 2 | 3 or more | All | 1 | 2 | 3 or more | All |
| Newfoundland | 24.5 | 4.2 | 7.1 | 9.7 | 5.5 | 3.7 | -1.9 | 1.7 |
| Prince Edward Island | 8.9 | -0.3 | 3.7 | 3.5 | 11.0 | 2.9 | 4.9 | 5.7 |
| Nova Scotia | 7.4 | 11.4 | 11.4 | 10.4 | 2.2 | 11.7 | 2.4 | 5.8 |
| New Brunswick | 1.1 | 10.7 | 14.3 | 9.6 | 13.3 | 0.6 | 0.5 | 3.7 |
| Quebec | 17.3 | 16.6 | 11.3 | 15.0 | 7.5 | 6.4 | 4.3 | 6.0 |
| Ontario | 14.2 | 14.6 | 18.4 | 16.2 | 13.4 | 15.8 | 16.3 | 15.4 |
| Manitoba | 15.0 | 3.8 | 20.5 | 13.3 | -17.1 | 1.1 | 4.1 | -3.2 |
| Saskatchewan | -18.2 | 15.3 | 15.8 | 5.8 | 6.6 | -12.1 | 3.1 | -1.2 |
| Alberta | -3.1 | 13.8 | 22.4 | 13.2 | -7.1 | 21.7 | 18.2 | 13.3 |
| British Columbia | 20.5 | 24.8 | 23.8 | 23.2 | 1.4 | 11.1 | 15.4 | 9.9 |
| **Canada** | **13.2** | **15.6** | **17.4** | **15.6** | **6.3** | **11.1** | **11.8** | **10.1** |

# 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-responsive 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-responsive 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-responding 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.

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 5.2 shows the household response rates by province.

**Figure 5.1 Cross-sectional person-level response rate (%)**

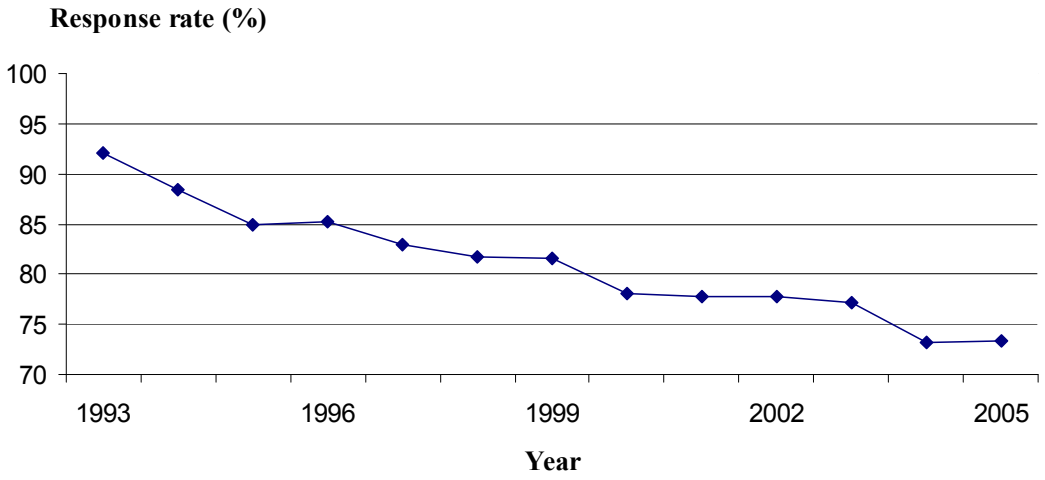
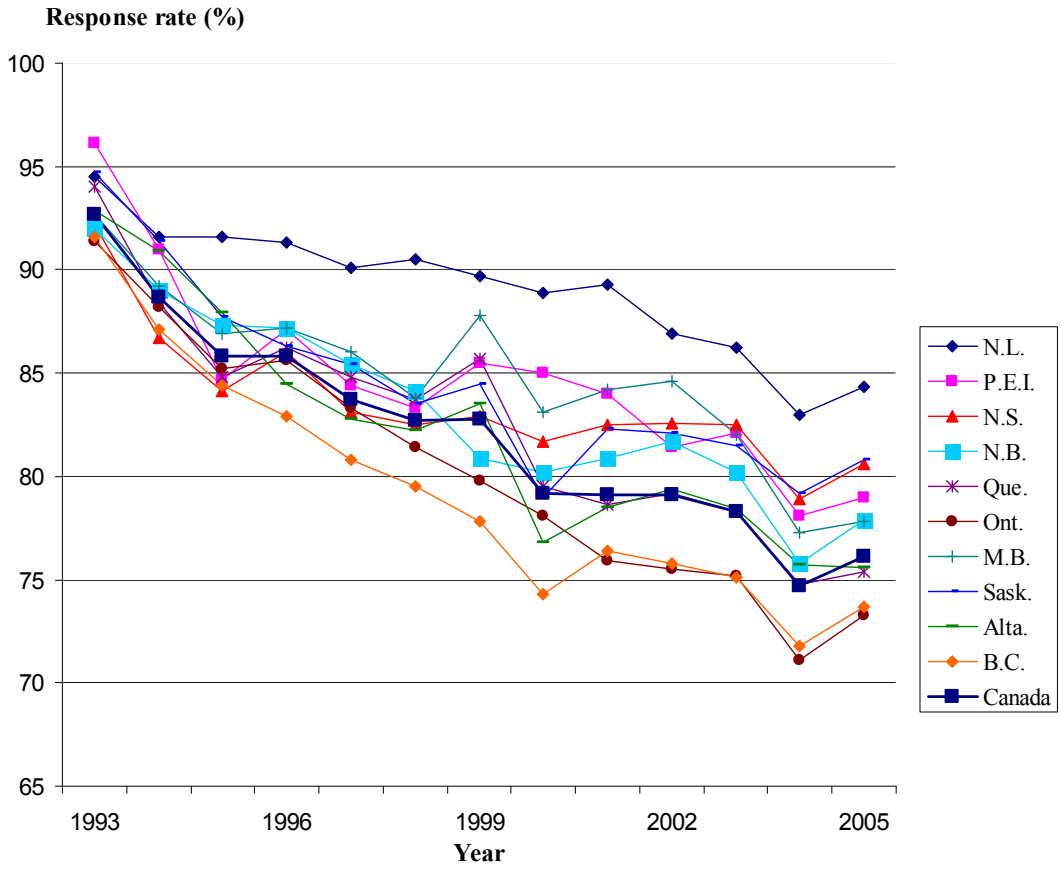



Table 5.1 shows the person response rates by phase. 'Respondent to labour questions' and 'Respondent to income questions' are the percentages of those who responded to only the labour or income sets of questions respectively whereas the 'Respondent to both sets' is the percentage of all those who responded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

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 that panel.

Figure 5.4 shows the longitudinal non-response rates each year by age group. 'Young' are people at least 16 years of age but less than thirty, 'Mid-aged' are people thirty years of age or older but less than sixty years of age and 'Senior' are people at least sixty years of age.

## Figure 5.2: Cross-sectional household response rate (%)




## Table 5.1 Cross-sectional person response rates by phase[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 |

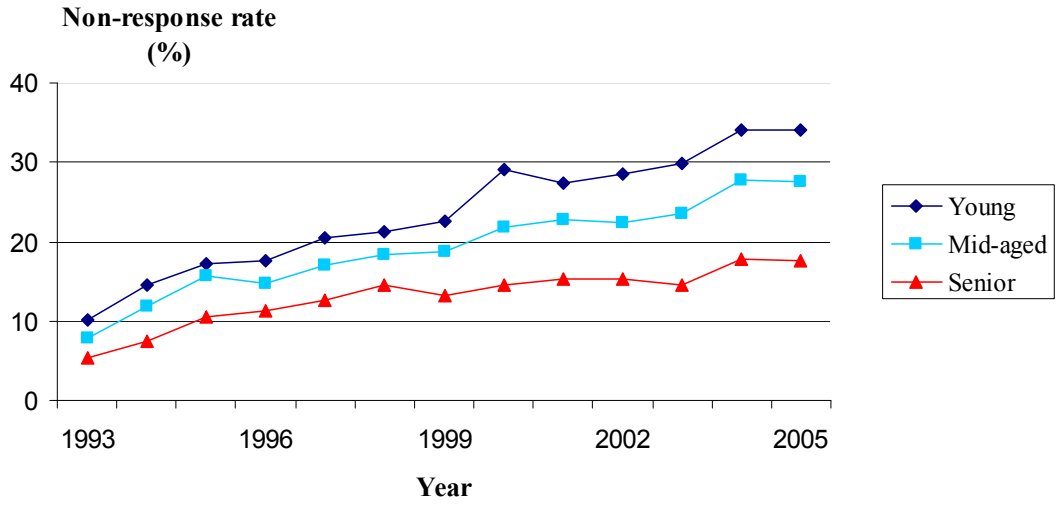
1. From 2004 onwards, we combined the labour and income interviews into a January interview

**Table 5.2 Longitudinal person-level (all ages) response rates by panel (%)**

| | Wave of Panel | | | | | |
|---|---|---|---|---|---|---|
| Panel (and year panel began) | 1 | 2 | 3 | 4 | 5 | 6 |
| Panel 1 (started in 1993) | 93.3 | 89.6 | 86.5 | 83.9 | 82.4 | 81.5 |
| Panel 2 (started in 1996) | 89.5 | 86.7 | 85.2 | 82.7 | 78.5 | 77.4 |
| Panel 3 (started in 1999) | 83.9 | 83.0 | 83.0 | 79.6 | 76.4 | 73.7 |
| Panel 4 (started in 2003) | 81.2 | 83.2 | 78.3 | 75.0 | … | … |
| Panel 5 (started in 2005) | 78.8 | … | … | … | … | … |

… not applicable

**Figure 5.3: Longitudinal non-response rate by age group**




# 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.

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 and 2004 are the same for both January and May because there were no May interviews in those years.

**Table 6.1 SLID permission rates by panel**

| | Wave | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | |
| | Interview | | Interview | | Interview | | Interview | | Interview | | Interview | |
| Panel (start date) | Jan | May | Jan | May | Jan | May | Jan | May | Jan | May | Jan | May |
| Panel 1 (1993) | … | … | … | 58.5 | 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 | … | 87.0 | 90.3 |
| Panel 3 (1999) | 55.0 | 71.7 | 76.4 | … | 78.2 | 84.7 | 86.6 | 90.0 | 89.8 | 91.9 | 92.7 | … |
| Panel 4 (2002) | 58.3 | 77.3 | 80.6 | 85.4 | 88.5 | … | 89.9 | … | … | … | … | … |
| Panel 5 (2005) | 69.7 | … | … | … | … | … | … | … | … | … | … | … |

… not applicable

Note: Permission was asked for the first time in May 1994
The May collection in 2000 was dropped so that the January interview collection period could be extended
Due to the high cost of collection, the May interview for income was dropped from 2004 onwards.

# 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

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.

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. 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 SIN found for respondents giving permission (%)**

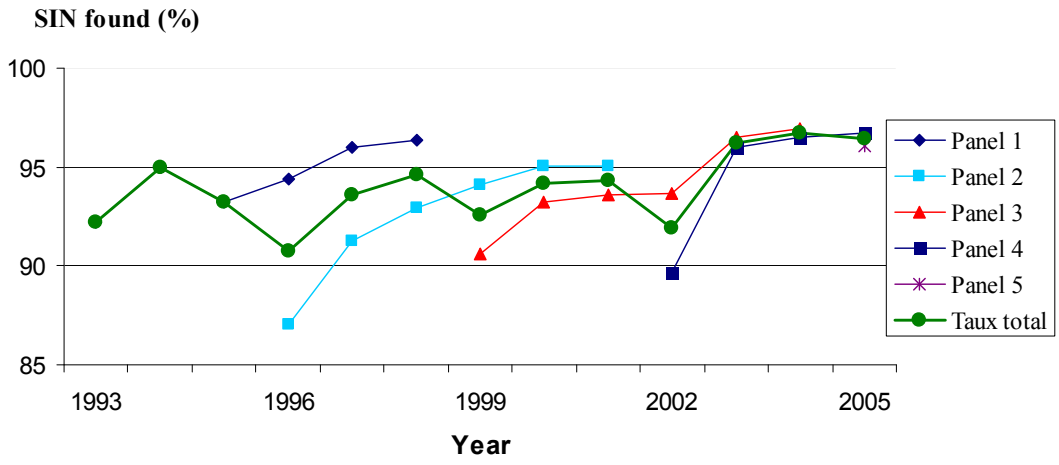


**Figure 7.2: Tax linkage rates where a SIN was found (%)**

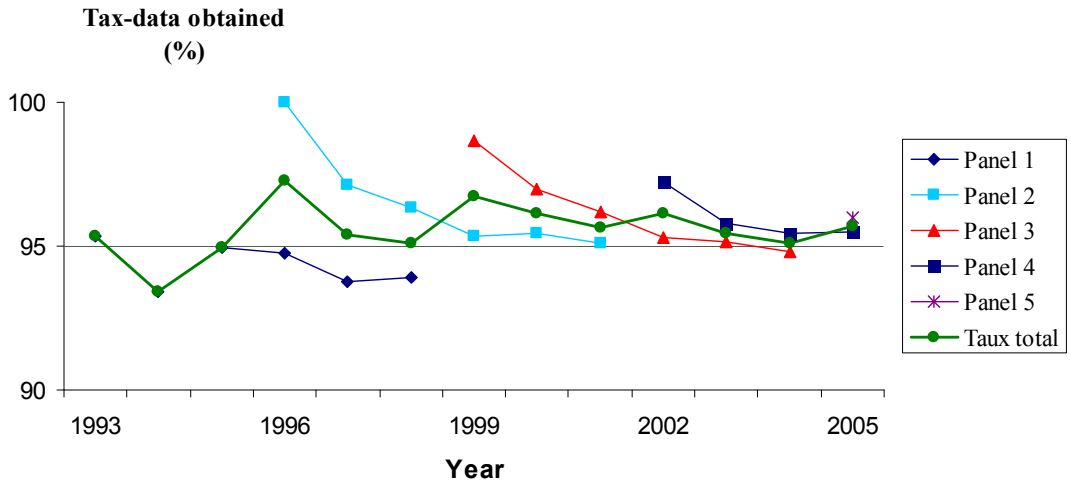



Table 7.1 compares the proportion of records from tax to those collected in the telephone interview. In total eighteen income variables are imputed during SLID income imputation. Many individuals require only partial imputation. Partial imputation is when one or more income items is imputed with some information being supplied by the individual.

**Table 7.1 – Income data coming from tax or interview (%)**

| 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 |

1. These are respondents not linked to tax and without responses to income questions (i.e. usually imputed).

# 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 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 one in which at least one household member has responded partially or completely to either the labour or income questions of the survey.

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.

**Table 8.1 Income-variable imputation for respondents in 2005 (%)**

| Province | Total Imputation[1] | Partial Imputation[2] | No Imputation |
|---|---|---|---|
| Newfoundland | 2.3 | 21.8 | 75.9 |
| Prince Edward Island | 2.2 | 20.0 | 77.8 |
| Nova Scotia | 1.6 | 22.4 | 76.0 |
| New Brunswick | 2.1 | 22.0 | 75.9 |
| Quebec | 2.3 | 19.6 | 78.1 |
| Ontario | 3.5 | 26.2 | 70.2 |
| Manitoba | 2.1 | 23.2 | 74.7 |
| Saskatchewan | 1.8 | 24.2 | 74.0 |
| Alberta | 3.1 | 24.3 | 72.6 |
| British Columbia | 3.3 | 24.7 | 72.1 |
| **Canada** | **2.7** | **23.5** | **73.8** |

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 – Breakdown of partial or total imputation in 2005 (%)**

| | Data Source | | | |
|---|---|---|---|---|
| Imputation | Tax | Interview | Other[1] | All |
| Partial (1 variable) | 6.2 | 14.5 | 0.0 | 6.0 |
| Partial (2 to 9 variables) | 0.2 | 34.2 | 0.0 | 3.5 |
| Partial (10 to 17 variables) | 0.0 | 0.0 | 13.9 | 13.9 |
| Total imputation | ... | ... | 2.7 | 2.7 |
| No imputation | 93.5 | 50.9 | ... | 73.8 |
| **Total** | **100.0** | **100.0** | **100.0** | **100.0** |

... not applicable

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.

**Table 8.3 Household-variable imputation in 2005 (%)**

| Province | Total Imputation[1] | Partial Imputation[2] | No Imputation |
|---|---|---|---|
| Newfoundland | 4.1 | 31.8 | 64.1 |
| Prince Edward Island | 6.3 | 33.3 | 60.4 |
| Nova Scotia | 4.5 | 31.1 | 64.4 |
| New Brunswick | 13.6 | 22.1 | 64.3 |
| Quebec | 3.6 | 29.6 | 66.8 |
| Ontario | 2.8 | 39.7 | 57.5 |
| Manitoba | 3.3 | 38.4 | 58.2 |
| Saskatchewan | 4.2 | 38.1 | 57.7 |
| Alberta | 3.1 | 41.1 | 55.9 |
| British Columbia | 1.6 | 42.7 | 55.7 |
| **Canada** | **3.9** | **36.0** | **60.2** |
| | | | |
| **Canada (2004)** | **4.0** | **31.3** | **64.7** |

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

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 8.3 shows the percentage of responding SLID households that were subject to total or partial imputation.

**Table 8.4 Households requiring partial imputation (%)**

| | 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 |

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

# 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 9.2 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 four digits of wages and salaries greater than $9,999 in 2005 (%)**

| | Fourth last digit (%) | | Third last digit (%) | | Second last digit (%) | | Last Digit (%) | |
|---|---|---|---|---|---|---|---|---|
| Digit | Collected | Tax | Collected | Tax | Collected | Tax | Collected | Tax |
| 0 | 31.8 | 11.6 | 88.7 | 11.6 | 96.0 | 12.8 | 97.1 | 14.2 |
| 1 | 4.8 | 10.9 | 0.6 | 10.0 | 0.2 | 9.7 | 0.2 | 9.4 |
| 2 | 9.0 | 10.6 | 1.2 | 10.0 | 0.4 | 10.1 | 0.3 | 10.1 |
| 3 | 6.1 | 9.8 | 0.4 | 9.9 | 0.3 | 9.7 | 0.4 | 9.1 |
| 4 | 5.7 | 10.4 | 1.6 | 9.6 | 0.7 | 9.6 | 0.3 | 9.2 |
| 5 | 19.5 | 9.6 | 3.7 | 10.0 | 0.6 | 9.8 | 0.4 | 9.7 |
| 6 | 6.2 | 9.8 | 0.9 | 9.9 | 0.5 | 9.7 | 0.4 | 9.6 |
| 7 | 5.0 | 9.4 | 0.6 | 9.4 | 0.4 | 9.6 | 0.4 | 9.4 |
| 8 | 7.8 | 9.2 | 1.5 | 9.9 | 0.5 | 9.6 | 0.1 | 9.6 |
| 9 | 3.9 | 8.8 | 0.8 | 9.7 | 0.3 | 9.4 | 0.3 | 9.7 |

## Table 9.2 Prevalence of zeroes in the last four digits of 2005 reported data (%)

| | Digit | | | |
|---|---|---|---|---|
| Variable | Fourth-last | Third-last | Second-last | Last digit |
| Wages and salaries | 26.3 | 81.8 | 94.6 | 96.4 |
| Investment income | 8.2 | 34.1 | 68.9 | 86.1 |
| Social assistance | 8.7 | 35.8 | 71.7 | 83.3 |
| UI Benefits | 6.5 | 47.0 | 87.8 | 94.0 |
| Non-farm self-employment income | 36.1 | 82.3 | 95.2 | 97.0 |