Income Statistics Division

75F0002MIE - 00006

Cross-Sectional Weighting: Combining Two or More Panels

Prepared by:
Michel Latouche
Johane Dufour
Takis Merkouris

October 2000

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this product and related statistics or services should be directed to: Client Services Income Statistics Division Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; income@statcan.ca) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialing area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 75F0002MIE-00006, is available on internet for free. Users can obtain single issues at http://www.statcan.ca/cgi-bin/downpub/research.cgi.

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 which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Income Statistics Division

Cross-Sectional Weighting: Combining Two or More Panels

Published by authority of the Minister responsible for Statistics Canada

October 2000

Catalogue no. 75F0002MIE - 00006
ISSN 0000-0000

Catalogue no. 75F0002MPE - 00006
ISSN 0000-0000

Frequency: Irr.

Ottawa

La version française de cette publication est disponible sur demande

---

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 co-operation and goodwill.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

TABLE OF CONTENTS


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

Summary

This paper discusses methods and tools considered and used to produce cross-sectional estimates based on the combination of two longitudinal panels for the Survey of Labour and Income Dynamics (SLID). The methodology adopted is similar to standard approaches when dealing with multiple frame design. One makes use of not so standard combined estimator that gives relative importance to the panels according to a panel allocation factors (*paf*). These factors are already available at estimation stage. If need be, the method could easily be extended to integrate a third panel or an additional cross-sectional sample. While several approaches were considered, it was decided to combine the panels such that the variance of level estimate is minimised. The variable *number of persons aged 15 and over* was chosen to compute the panel allocation factors for each province. To simplify the weight calculation, it was decided to derive the panel allocation factors using an external source. For the 1996 reference year, data from the Labour Force Survey (LFS) and SLID were used to compute these factors. Overall, the 1996 data suggest that the use of optimal panel allocation factors leads to interesting gains in precision and can reduce potential attrition bias. The data also suggest that there are some differences in the estimates produced by panel 1 and panel 2 that deserve further investigation.


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

1. Introduction

SLID is a longitudinal survey that produces not only longitudinal data on labour activity of the individuals, but also cross-sectional estimates of individual and family income characteristics on a yearly basis. In 1996 reference year, for the first time, the SLID used two panels to produce cross-sectional estimates. In theory, each panel can produce cross-sectional estimates by itself. Better estimates are however produced by combining the two panels together. Thus, estimates are produced using a bigger sample that allows a reduction in the variance of the estimates and makes it possible to use many more control totals at the calibration stage reducing potential biases and the variability even further.

Combined estimation takes the form of a weighted sum of the panel estimates where the weights are the panel allocation factors (*paf*) (Merkouris, 1999). The use of combined estimator at Statistics Canada is not new. The Labour Force Survey (Singh, Drew, Gambino and Mayda, 1990) has always used allocation factors to combine the six rotation groups present in the survey design. The Survey of Consumer Finances also combines the four rotation groups retained from the LFS in their sample using allocation factors. In the past, both surveys gave the same weight to their rotation groups. In 1998, the Labour Force Survey decided to give less importance to the youngest rotation group to improve trend estimates (Singh, Kennedy, Wu, and Brisebois, 1997).

Because several allocation factor combinations can produce appropriate estimates, one has to decide upon a strategy for the computation of the SLID *pafs*. For example, one could give the same importance to the panels as in the Survey of Consumer Finance. Although simple, this is not recommended for two reasons. First, longitudinal surveys could be subject to attrition, which can impact on the quality of the estimates. Second, since the SLID samples are selected from the LFS samples at different points in time, reliability of the panels could be different especially if they come from different Labour Force designs due to the LFS redesign which occurs every 10 years.

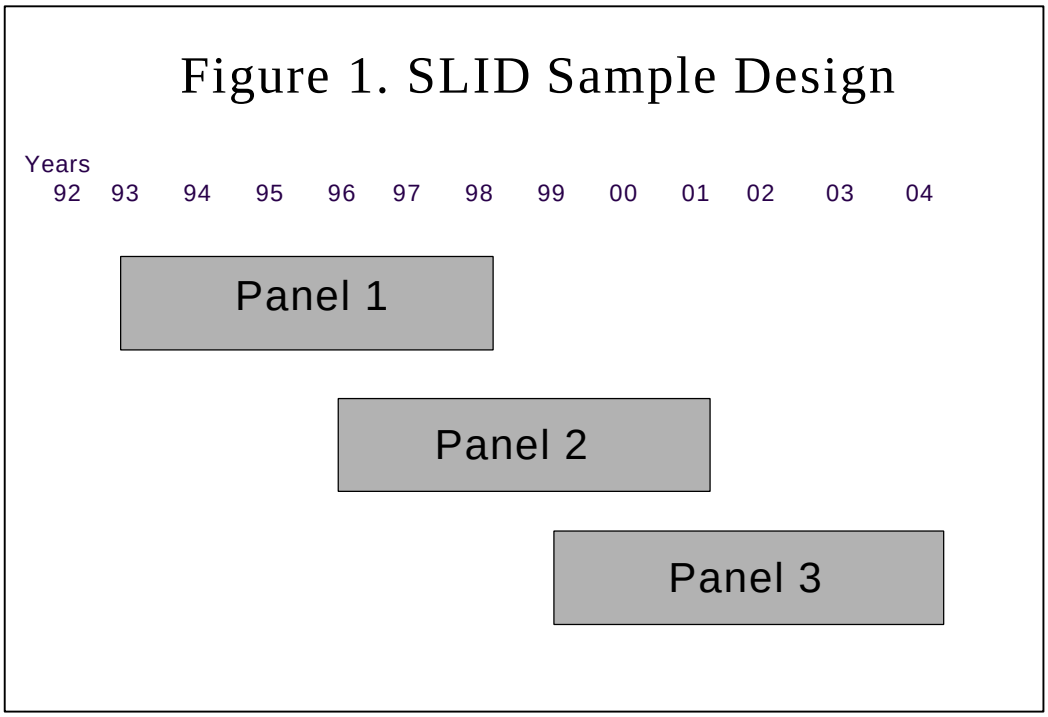
It was decided to calculate *paf* assuming the worst case scenario where the panels have different impacts on data quality. Under this scenario, the *paf* should be such that the mean square error of level or yearly trend estimate is minimised for as many variables of interest as possible. In principle, a set of *pafs* could be calculated for each variable of interest. In fact, for operational reasons and simplicity, only one set of *pafs* is computed hoping that it will produce reasonable mean square errors for all variables. In this way, only one cross-sectional set of weights is required for all variables.

Mean square error involves variability and bias of the estimates. Because it is difficult to have sound estimates of bias, it was decided to consider only variability. We assume that nonresponse adjustments eliminate bias. However, although not proven yet, it is possible that the oldest panel leads to some bias. Even if the bias is not considered, we are conservative and assume that the oldest panel may be more subject to bias than the youngest one. This possibility is taken into account in the way *paf* calculations are done.

The following section provides some information about SLID design. Section 3 discusses the methods considered for the calculation of the *pafs*. Section 4, deals with the practical aspect of the calculations including the choice of the variable to be optimised as well as operational definitions. Section 5 covers calculation for the 1996 reference year and future waves. Section 6 and 7 deal with the impact on the estimates and on their reliabilities. Finally, Section 8 provides some recommendations for future processing of SLID samples.

2. The Survey of Labour and Income Dynamics sample design

SLID is an annual survey made of two panels (Lavigne and Michaud, 1998). The first panel covers persons living in one of the 10 provinces as of January of 1993 (excluding persons living in military barracks, or in institutions). From that population, a longitudinal sample is selected. These longitudinal units are in the sample for six years. The second panel is selected from the 1996 population and also lasts six years. This results in overlapping panels as shown in Figure 1. A new panel will subsequently be selected every three years to replace the older of the two panels. Such a design implies that annual cross-sectional estimates can be produced using two panels, except for the first three years of the survey (1993-1995).



For longitudinal purposes, each panel must represent the population it was sampled from, while for annual cross-sectional estimations the target population changes every year. The introduction of a new panel helps to account for both births and the arrival of new immigrants, an important factor from the cross-sectional point of view.

Each SLID panel consists of a subsample of 15,000 households (approximately 40,000 persons) taken from the LFS. The LFS sample is drawn from an area frame according to a multi-stage random sampling design sample (Singh and al., 1990). The LFS operates on the basis of six panels, one of which rotates every month. The last stage sampling unit is the household. The households selected for the SLID are those that rotated out of the LFS at the beginning of the reference period. Each panel is made of two LFS rotation groups.

The individuals included in the sample at the beginning of the panel are called longitudinal respondents, as opposed to the cohabitant respondents who join the household of a longitudinal respondent later in the panel life time (Lavallée, 1995). The cohabitants are included in the cross-sectional weighting because we are interested in the household characteristics. In addition, the cohabitants help improve the cross-sectional sample representativity.

2.1 Steps in the Weighting Process

Whereas longitudinal weighting is carried out separately for each panel, cross-sectional weighting is done by combining both panels (Lévesque and Franklin, 2000). The weighting process involves six steps:

- Basic weight
- SLID nonresponse adjustments
- Combination of the two panels
- Weight sharing
- Analytical adjustments
- Post-stratification

1- Basic Weight

For each panel, the basic weight is computed using the LFS probabilities of selection. It also includes the nonresponse adjustment to the LFS. The weight is then adjusted to account for the fact that SLID uses only two out of six LFS rotation groups. A special step is also performed for the panel 1 units in order to adjust for nonresponse to the preliminary interview. This special adjustment is carried out for the selection strata in accordance with the LFS design.

2- Adjustment for SLID Nonresponse

Only the LFS respondents are included in SLID sample. Since one has LFS data for all sampled units, it is them possible to model nonresponse using LFS data. Adjustment for nonresponse is thus based on homogenous response groups. The variables used to form these response groups are determined using logistic regression or segmentation modelling (Dufour, Gagnon, Morin, Renaud and Särndal, 1998). The purpose of these adjustment categories is to offset potential biases that could otherwise be introduced by the fact that the response process cannot be overlooked. The modelling and the nonresponse adjustment are done separately for each panel.

3- Combining the panels

At this step the two panels are grouped together to form a large sample using the *panel adjustments factors*. This step is the main purpose of this paper and is explained in section 3.

4- Weight Sharing

This step, which is specific to cross-sectional weighting, is required due to the presence of cohabitants. In fact, cohabitants are included in the sample solely because they have joined households that comprise at least one individual in the longitudinal sample. Since the cohabitants have not been pre-selected using a sampling design with known probability, the weight sharing method must be used in order to obtain weights that yield unbiased estimates (Lavallée, 1995).

5- Analytical adjustments

These weight adjustments are done either because the weights are found to be extremely high compared to others within the same province, or the weighted individual contribution to aggregate income estimates is large. In the former case we talk about extreme weights; the latter refers to outliers (Lévesque and Franklin, 2000). Extreme weights are caused by inter provincial mobility; over time, some respondents move to another province. The outlier adjustment is done for confidentiality reasons and also to insure representativity.

6- Post-stratification

Finally , the weights are post-stratified according to demography counts at the province, age and sex cross-classification for the reference year.

3. Methods considered and data used

The *paf* can give equal or unequal importance to the panels to meet some requirements. Although it is planned to review and adjust the *pafs* at each wave, analytical considerations foster the use of stable *pafs* as much as possible.

The *paf* enters into the calculation of the variance. Although it can also affect bias, this consideration is not fully considered here. Cross-sectional deliverables of SLID include level and yearly trend estimates, and both are affected by the choice of *paf*. For the 1996 reference period, it was decided to calculate the *paf* based on the level estimates for the following reasons:

- More stability over time. For trends, in 1996, only the first panel overlaps with 1995, but in 1997 both panels will overlap with

1996. Optimising trend estimates could lead to considerable differences in the *paf* from 1996 to 1997 (see appendix A). For instance panel 1 would be given much more importance in 1996 and much less in 1997 when the second panel contributes also to the reduction of the trend estimate variance.

- Better cross-sectional estimates: the fact that almost the whole SLID sample will overlap 2 consecutive years two times out of three, and that on the third occasion close to 50% of the sample overlaps reduces the need to optimise yearly trend estimations. On the other hand, SLID was not originally designed to optimise cross-sectional estimates.

The cross-sectional combined estimator can be described as follows: let $\hat{Y}$ be the cross-sectional estimate for a given variable of interest, $\hat{Y}_1$ and $\hat{Y}_2$ be the usual Horwitz-Thompson estimates produced with panel 1 and 2 respectively. Then, the composite or combined estimate is given by:

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2 + \hat{Y}_2' \qquad (1)$$

where p_1 and p_2 are the *pafs* of panel 1 and panel 2. $\hat{Y}_1$ and $\hat{Y}_2$ are produced using the respondents that were eligible to be selected in both panels, while $\hat{Y}_2'$ is produced using the respondents that join the target population after the selection of panel 1 (births, immigrants, etc). That corresponds to a small population which represents yearly only 0.3% of the total population. Most of the time it is not possible to determine precisely when the respondents came into the target population; data collection information do not contain such data. Nevertheless, very few respondents come in after the selection of panel 1. Consequently, one assumes that all respondents are eligible to be selected in both panels and one uses the estimate:

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2 \qquad (2)$$

Note that $\hat{Y}$ is unbiased only if $p_2 = 1 - p_1$ which will be the case. Note that in production, the *paf* is applied at the micro level. That is, each longitudinal respondent weight is multiplied by its associated *paf*. This is done after panel nonresponse adjustment but before weight sharing and calibration, so that the *paf* is included in a cohabitant's weight.

It can be shown that the variance is minimised if

$$p_j = \frac{n_j / deff_j}{\sum_{j=1}^{2} n_j / deff_j} \qquad (3)$$

where n_j is the number of longitudinal respondents (nonzero longitudinal weight) in panel *j* and $deff_j$ are the design effect of the panel *j*. For the moment, there are only two panels, but the formula would be similar if another panel or cross-sectional sample were added (*j*=3).

3.1 Data used

At the time this study was undertaken, SLID data were available only for the first panel. On the other hand, the Survey of Consumer Finances (SCF) data were available for several years. The choice of SCF data was justified by the fact that SCF sample design is similar to SLID (they are both sub-samples of the LFS). Furthermore, their income contents are the same, making it possible to study several SLID variables of interest. For demographic variables, LFS data were used, because it allowed using a larger sample. The use of LFS data is discussed further in section 4.2.

SCF 1993 and 1996 reference year data were used. These years correspond to the years panels 1 and 2 were selected. Note that the LFS redesign took place in 1994. This means that both samples come from a different design hence corresponding exactly to the SLID situation. Finally, the use of these two years corresponds to the time lag between the two SLID panels.

One of the four SCF rotation group had to be dropped because it was found to be very different from the other rotation groups in two provinces (rotation group 6 for the 1993 year and rotation group 2 for the 1996 data).

4. Operational definitions and considerations

In order to determine a useful and practical definition of the *paf*, there are several aspects that have to be considered. These aspects can be divided in the two following elements:

- variables(s) to be considered in the optimisation process
- type and reference period of the data to be used.

These elements are discussed in the next subsections.

4.1 Variable(s) to be considered.

SLID is interested in four categories of data: demographics, labour, income and low income measurement. Each variable in these groups has its own design effect. Table 1 illustrates the variability of the design effect ratios for four variables as estimated by the SCF.

Table 1.
Ratio of 1994 to 1997 SCF design effects
by variable and province[1]

PROVINCE	NUMBER OF PERSON AGE 15+	TOTAL INCOME	NUMBER OF HOUSEHOLDS OF SIZE 2+	NUMBER OF PERSONS BELOW LICO[2]
Newfoundland	1.60	2.33	2.66	1.11
Prince Edward Island	2.19	2.51	3.84	1.05
Nova Scotia	4.45	2.75	4.58	1.53
New Brunswick	2.97	2.58	3.37	0.76
Québec	2.92	4.70	2.90	1.37
Ontario	1.71	2.18	1.77	1.55
Manitoba	4.83	2.70	4.18	1.47
Saskatchewan	3.02	2.67	2.96	1.34
Alberta	2.95	2.30	2.44	1.74
British Columbia	2.71	2.44	2.73	1.78
Canada	2.24	2.48	2.29	1.52

[1] Rotation group 2 is excluded from the calculation for all variables except LICO because it was found to give very different results. Design effect ratios for LICO include rotation group 2, but it is not the reason why ratios are smaller than those for the other variables.

[2] Low Income Cut-Off

Most of the time the ratios are greater than one. This will force the *paf* of panel 2 to be greater than the panel 1 *paf*. Optimising for all variables would lead to the calculation of many *pafs* and sets of weights. That possibility is out of question for operational and analytical reasons given the large number of key variables in SLID products. One then has to choose a unique variable, compute its *paf* and hope that other variables will not suffer too much from that choice. Ideally the *paf* should be available in advance so that *paf* calculations do not slow down the weighting process. In the future when two panels of data will be available, the *paf* can be derived using SLID previous wave data. In the mean-time, one has to use external sources. The current two main external sources of information relevant to SLID are the LFS and the SCF. The variable recommended for the calculation of the *paf* is *number of persons aged 15 or more* which corresponds to the LFS target population. This choice should provide a more stable estimate of the *deff* because it corresponds to a large domain, and because it is a categorical variable. Other factors

explain this choice. First this variable is correlated to all other variables almost equally. Secondly, this variable is produced directly from the LFS hence will always be available in the future as opposed to the data from the SCF which will be integrated with SLID for the 1998 reference year. Thirdly, this variable is defined for all SLID respondents. Finally, the LFS can provide more reliable design effect estimates since it has up to six rotation groups that can be used instead of two for SLID and 4 for the SCF. This assumes that the true design effects do not depend on sample size, which is true when the sample increases with the number of rotation groups.

Moreover, Table 1 also indicates that the ratios of the design effects vary from one province to another. Considering that the ratios of the sample sizes also vary greatly between provinces, it is recommended to compute the *paf* at the provincial level.

4.2 Type and reference period of the data to be used

Several operational definitions of sample size and design effect can be used to calculate the *paf*. Regarding the sample sizes used in equation (3), only longitudinal respondents at time of estimation are considered despite the fact that cohabitants are used in the cross-sectional estimations. Lavallée (1994) indicates that in the weighting process, the two panels must be combined prior to the integration of the cohabitants using the weight share method. Consequently, only longitudinal respondents are used here. In addition, omitting the cohabitants in the *paf* calculation contributes to reduce the importance of the old panel and thus minimising the impact of potential attrition bias.

Since the SLID sampling units are the households, the design effects for the variable of interest are computed at the household level. Hence, design effects are computed assuming a simple random sample of households.

Two types of design effects are considered: one assuming a simple random sample by stratum and another one by province. The latter is used because it reflects both stratification and clustering effects. Note that the design effects are computed using the weights after nonresponse adjustment (subweights), but prior to the calibration.

To insure minimal variance estimates for at least the variable *number of person aged 15 or more,* the *paf* should be computed using SLID data from both panels at time of estimation. This means that first one has to compute individual weights for each panel as described by Latouche, Michaud and Renaud, (1997) or Dufour and al. (1998), then compute the

paf using equation (3), and finalise the individual weights. Such an approach makes the *paf* sample dependent. It is operationally easier to create a sample independent *paf* by using external sources as described in section 4.1.

Since the LFS is used to compute the design effects, the question arises on what reference period should one use. Technical methodologists believe that the reference period should be the one at time the panel is selected. Accordingly, panel one design effects are computed using January 1993 LFS data, and January 1996 data for panel 2. In this way, the old LFS design is used for the panel 1 and the new survey design is used for panel 2. On the other hand, practical methodologists suggest using the design effect corresponding to the reference period one estimates for. This allows not only consideration of the time lag between the two panels, but also design deterioration over time. Hence, panel 2 design effects are estimated by January 1996 LFS data as in the previous approach. Panel 1 design effects are also estimated using January 1996 data. However, the old LFS design was no longer used in 1997. The closest period that can be used is September 1994 which corresponds to the last time the old design was fully used. For the 1996 reference year, it was decided to use the technical approach (1993 for panel 1 and 1996 for panel 2).

5. *PAF* and Frequency of calculation

5.1 1996 reference year

Table 2 presents the *paf* calculation set up for the 1996 reference period. This set up leads to the *paf* presented in Table 3.

Table 2.
1996 *Paf* calculation set up

Elements	Operational definition
Variable of interest	Number of persons aged 15 and over
External Source	Labour Force Survey
Reference Period	January 1993 (panel 1), January 1996 (panel 2)
Level of calculation	Province
Sample sizes	Longitudinal respondents at time of estimation
Type of design effects	Assume random sample of households
Level of design effects	Stratum and cluster, using subweights

Table 3.
Paf used in 1996 reference period
(computed using 1993 and 1996 LFS data)

province	Size of panel 1	Size of Panel 2	$deff_1 / deff_2$	*Paf* Panel 1	*Paf* panel 2
Newfoundland	1,698	1,315	1.71	0.4302	0.5698
Prince Edward Island	575	895	1.92	0.2507	0.7493
Nova Scotia	1,853	2,044	4.02	0.1840	0.8160
New Brunswick	1,768	1,882	1.94	0.3263	0.6737
Québec	4,928	5,853	2.75	0.2344	0.7656
Ontario	7,054	9,174	2.87	0.2113	0.7887
Manitoba	1,821	2,113	1.36	0.3879	0.6121
Saskatchewan	1,945	1,863	0.92	0.3208	0.6792
Alberta	2,406	2,164	2.21	0.2792	0.7208
British Columbia	2,264	2,570	2.87	0.2420	0.7580

If both panels were of same size and had the same design effect in all provinces, the *paf* would be 0.5 for both panels. This would be similar to the old LFS rotation group allocation factor which is *1/6* for each group. For SLID, one can see that the *pafs* are very different from 0.5. This discrepancy is mainly caused by the ratios of the two design effects. Table 4 presents the *pafs* that would result if both panels had the same design effect.

Table 4.
Paf in 1996 reference period
(assuming same design effects)

province	Size panel 1	Size panel 2	$deff_1 / deff_2$	*Paf* Panel 1	*Paf* panel 2
Newfoundland	1,698	1,315	1	0.5636	0.4364
Prince Edward Island	575	895	1	0.3912	0.6088
Nova Scotia	1,853	2,044	1	0.4755	0.5245
New Brunswick	1,768	1,882	1	0.4844	0.5156
Québec	4,928	5,853	1	0.4571	0.5429
Ontario	7,054	9,174	1	0.4347	0.5653
Manitoba	1,821	2,113	1	0.4629	0.5371
Saskatchewan	1,945	1,863	1	0.5108	0.4892
Alberta	2,406	2,164	1	0.5265	0.4735
British Columbia	2,264	2,570	1	0.4683	0.5317

5.2 Later years

It is obvious that the longitudinal sample size of the two panels will change at each wave. Design effect ratios according to the technical approach are assumed to be the same until a new panel comes in. Considering that the design effect ratios and the sample size ratios are stable over time

when computed on a same pair of panels, it is suggested not to re-compute the *paf* for production purpose. This will avoid causing discrepancies in yearly trends. The *pafs* need to be recalculated only when introducing a new panel. Nevertheless, *paf* value should be monitored every year as a quality assurance activity.

When introducing a new panel that comes from the same LFS design as the remaining panel the technical approach would lead to *paf* that favours the oldest panel. This situation will occur in 1999 with the selection of panel 3. One can assume the new panel will have larger design effects caused by a deterioration of the stratum homogeneity. Although, the ratios of design effects should be closer to one than the ratios computed with different design, the ratios would be smaller than one, thus opening the door to more attrition bias. With the practical approach, *paf* would favour the youngest panel and consequently is recommended. When both panels come from the same LFS design, the practical approach can easily be used, and production delays can be avoided by using previous wave design effects.

6. Impact on estimates

Table 5
National estimates as produced
By optimal and equal *paf*

Variable	Optimal *paf*	Equal *paf*	Relative Difference[1]
Number of unattached individuals	3,984,199	4,033,703	1.24
Number of size 2 families	3,407,517	3,445,625	1.12
Number of size 3+ families	4,777,024	4,742,161	-0.73
Number of married persons	12,471,961	12,435,254	-0.29
Number of single persons	6,411,378	6,378,680	-0.51
Number of separated persons	648,956	727,732	**12.14**
Number of persons with unknown marital status	192,204	128,209	**-33.30**
Number of families in rural area	1,311,865	1,315,916	0.31
Number of families in 100000- 499999 area	1,985,941	2,043,483	**8.14**
Number of families in 500000+ area	5,933,449	5,893,937	**-3.57**
Total earnings (X 10^6)	421,029	425,795	1.13
Total investment (X 10^6)	23,863	24,617	3.16
Total government transfers (X 10^6)	76,467	76,196	-0.35
Total other money income (X 10^6)	44,103	45,033	2.11
Total income (X 10^6)	563,583	569,553	**1.06**
Average income: family	56,955	57,290	0.59
Average income: unattached	24,371	24,821	**1.84**
Average income: persons 16+	25,347	25,605	**1.02**
Percentage of persons below LICO	18.60	18.01	**-3.17**

[1] **bold font** means that the difference is statistically significant at the 1% level.

It is of interest to compare the estimates produced by a set of optimal *pafs* and those produced by giving the same importance to both panels. In theory, both sets produce unbiased estimates. What is more, if the estimates produced by both panels separately were the same, then the combined estimates would not depend on *paf* values. Nevertheless, since panels are at least subject to sampling variability, separate panel estimates will be slightly different. If differences fell outside the confidence interval, it would indicate that the panels are different somehow and that the *paf* value could greatly affect the combined estimates. In that case, the panel differences could be caused by change in processing, response error (accustom bias) or sampling coverage including attrition bias.

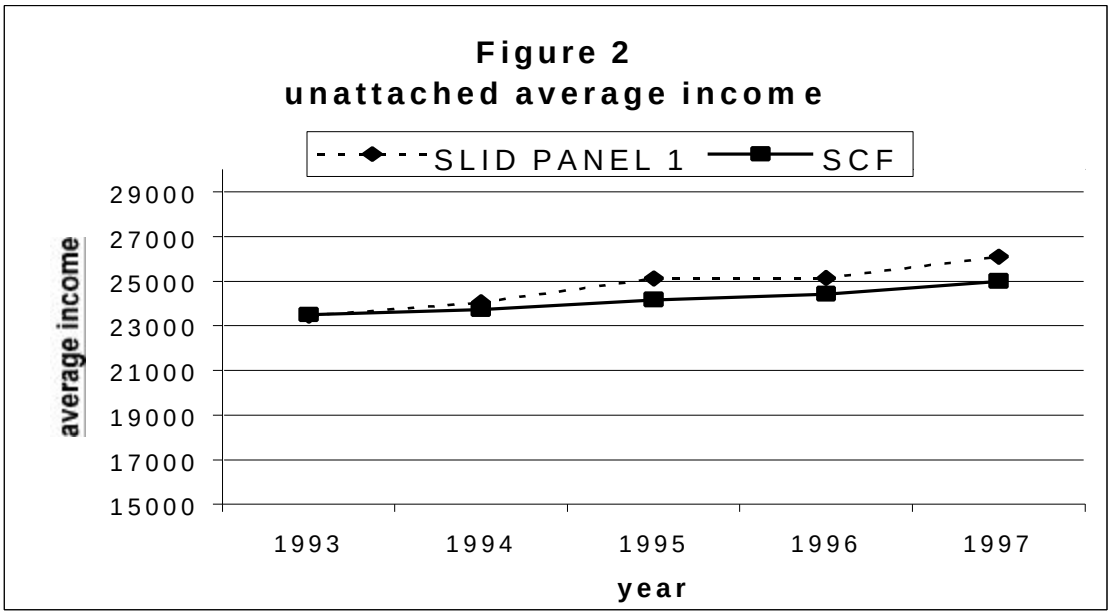
Table 5 shows the 1996 national estimates for key variables produced using optimal *paf* (as provided in Table 3) and by setting the *paf* to 0.5 for all provinces. The calibration used for both sets of *pafs* is the simple province, age, sex post-stratification used in production. Statistically significant relative differences are in bold. The 1996 data suggest that differences occur in four areas: marital status and urban size distribution, average income and low income measurement. The Appendix C presents the approach used to approximate the variances of those differences.

The optimal *paf* estimate suggests that there are much fewer separated persons estimated by panel 2. However it also indicates that there are many more unknown marital statuses. Over time, these differences should vanish with file cleanup.

The urban size distribution discrepancy is observed in the 100 000-499 999 and the 500 000+ groups; the equal *paf* estimate inflates the former and reduces the latter. Because optimal *paf* value for panel 2 is greater than 0.5, this indicates that there are fewer people in the 100000-499999 according to the panel 2 and more in the 500000+ group. This difference is caused by the LFS redesign that took place in 1994. Basically, the new LFS design leads to a bigger sample in urban areas than in rural areas compared to the old design. The SCF was also affected by the LFS redesign.

For the total income aggregate, the unattached and 16+ average total income estimates suggest that panel 1 provides bigger income estimates than panel 2 especially for unattached individuals. This pattern combined with the urban size one can explain why there is a significant difference between low income measurement estimates produced by the optimal and equal *pafs* strategy. It is up to subject matter specialists to judge if these differences are important. The differences could indicate a coverage problem; over time, a panel tends to lose poorer unattached individuals. However, when considering panel 1 only, the average total income for unattached individuals series does not differ from the one produced using

SCF data as shown in Figure 2. Notice that none of the yearly estimates produced by the SLID panel 1 and SCF are significantly different at the 1% level. Further investigations using more precise variance estimates have to be done in order to explain differences between panel 1 and 2 as regard income statistics.



The income distribution graph is presented in Appendix B. The two distributions are very similar.

7. Impact on Variance

It is of interest to know the loss of precision caused by forcing the *paf* to always be equal to 0.5. A simple and useful sensitivity analysis based on the formulation of Appendix A is as follows.

Given that the minimal variance and the optimal *paf* for panel 1 are:

$$V_{\min}(\hat{Y}) = \frac{V(\hat{Y}_1)V(\hat{Y}_2)}{V(\hat{Y}_1)+V(\hat{Y}_2)}$$

$$p_{opt,1} = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1)+V(\hat{Y}_2)}$$

and that the variance of $\hat{Y}$ when the *paf* is set to 0.5 is

$$V_{0.5}(\hat{Y}) = \frac{V(\hat{Y}_2)}{4p_{opt}}$$

then the loss of precision incurred by using a *paf* of 0,5 is

$$\frac{V_{0.5}(\hat{Y})}{V_{opt}(\hat{Y})} = \frac{1}{4\,p_{opt}\,(1-p_{opt})}$$

Table 6 shows for each province the loss of precision caused by forcing the 1996 provincial *paf* to be equal to 0,5. For example, let's consider Ontario. Its optimal *paf* is 0,2113. By setting the *paf* to 0,5 inflates the variance by 19% and the CV by 9%.

Table 6. Loss of precision caused by forcing the *paf* to be equal to 0,5

province	optimal paf	variance loss (%)	cv loss (%)
Newfoundland	0.4302	426.32	129.42
Prince Edward Island	0.2507	177.78	66.67
Nova Scotia	0.1840	96.08	40.03
New Brunswick	0.3263	56.25	25.00
Québec	0.2344	33.33	15.47
Ontario	0.2113	19.05	9.11
Manitoba	0.3879	9.89	4.83
Alberta	0.3208	4.17	2.06
British Columbia	0.2792	1.01	0.50

One can see that the use of an optimal *paf* leads to interesting gains in precision.

8. Recommendations

It is very difficult to determine the best practices for *paf* calculation in the case of SLID. So many variables have to be considered that it is impossible to fix a strategy that will be optimal for all of them. In the light of the work done using 1996 data, it seems that level estimates do not change greatly when computed using optimal or equal *paf*. However, some provincial estimates can be more affected especially for small domains.

In terms of variability, some gains in precision are achieved using optimal *paf*. For stability and bias reasons, it is more appropriate to compute *paf* to minimise the level estimates instead of trend estimates. Although these increases in precision may not seem that important, the use of optimal *paf* results in the reduction of potential attrition bias by giving more importance to the youngest panel. It is therefore recommended to keep using optimal *paf*.

One has seen that taken separately, the two panels could give different estimates. This implies that combined estimates could vary significantly with different *paf* values. It is then suggested to review the *paf* only when introducing a new panel in order to avoid analytical complications. As regards the cause of these discrepancies, it is important that one verifies some hypotheses to explain these differences, and to keep monitoring panel coverage over time.

9. References

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. and Särndal, C.-E. (1998). Measuring the Impact of Alternative Weighting Schemes for longitudinal Data. Proceedings of the American Statistical Association SRMS, pp. 552-557.

Latouche, M., Michaud, S. and Renaud M. (1997). Concerns Pertaining to Weighting of Longitudinal Surveys. American Statistical Association Proceedings of the Section on Government Statistics and Section on Social Statistics. Pp. 111-119.

Lavallée, P. (1994). Ajout du second panel à l'EDTR: sélection et pondération. Statistics Canada internal document.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Household Using the Weight Share. Survey Methodology, Volume 21, No. 1, June 1995: 25-32.

Lavigne, M. and Michaud, S. (1998). General aspects of the Survey of Labour and Income Dynamics, SLID Research Document, Statistics Canada, catalog 98-05.

Lévesque, I. and Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics : 1997 Reference Year. Statictics Canada working paper Catalogue no. 00004.

Merkouris, T. (1999). Cross-sectional Estimation in Multiple-panel Household Surveys. Statictics Canada working paper no.HSMD-99-004E.

Singh A., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, 300-305.

Singh, M.P., Drew, J. D., Gambino, J.G. and Mayda, F. (1990). Methodology of the Canadian Labour Force Survey 1984-1990. Publication of Statistics Canada. Catalogue no. 71-526.

APPENDIX A
VARIANCE ESTIMATION

A1- Optimising *paf* for level estimates

Assuming a complete panel overlap, the cross-sectional combined estimator can be described as follows: let $\hat{Y}$ be the cross-sectional estimate for a given variable of interest, $\hat{Y}_1$ and $\hat{Y}_2$ be the estimate produced by panel 1 and 2 respectively. Then, the composite or combined estimate is given by:

$$\hat{Y} = p_1\hat{Y}_1 + p_2\hat{Y}_2$$

where p_1 and p_2 are the panel one and panel two *paf*. Note that $\hat{Y}$ is unbiased only if $p_2 = 1 - p_1$ which will be the case.

The variance of $\hat{Y}$ is given by:

$$V(\hat{Y}) = {p_1}^2 V(\hat{Y}_1) + {p_2}^2 V(\hat{Y}_2)$$

It can be shown that the variance is minimised if

$$p_1 = \frac{V(\hat{Y}_2)}{V(\hat{Y}_1) + V(\hat{Y}_2)}$$

which gives the minimal variance

$$V_{\min}(\hat{Y}) = \frac{V(\hat{Y}_1)V(\hat{Y}_2)}{V(\hat{Y}_1) + V(\hat{Y}_2)}$$

Assuming that the population variance (S^2) is the same for both panels and that the finite population correction (*1-f*) is negligible, the *pafs* can be expressed as:

$$p_1 = \frac{n_1}{n_1 + n_2 \, deff_1 / deff_2}$$

$$p_2 = 1 - p_1$$

where $deff_1$ and $deff_2$ are the design effect of panel 1 and 2 respectively, and and n_2 are the number of longitudinal respondents (nonzero longitudinal weight) in panel 1 and 2 at time of estimation. One can see that optimal calculation of the *paf* depends only on the longitudinal sample sizes and the ratio of the panel's design effects. Finally, if the sample was made of *K* panel instead of two, the *paf* would be:

$$p_j = \frac{n_j / deff_j}{\sum_{j=1}^{K} n_j / deff_j} \qquad (1)$$

This formula still applies if a panel was replaced by a typical cross-sectional sample.

A2. Variance calculation

SLID variance estimates are produced by jackknifing. It was of interest to know if one has to compute new *paf* values at each jackknife iteration. Considering that the jackknife method is conditional on the sample size, there is no need to recalculate the *pafs*. On the other hand, if one wants to simulate exactly the weighting process at each iteration (implying a change in n_1 and n_2), then the *pafs* should be recalculated. To see what would be the impact on *paf* values, a simulation was performed to see the possible range of provincial *paf*. Computing the *paf* after removing the smallest or the biggest cluster of each province did this. The results are shown in the Table A1. In practice, the changes in the *paf* values are so small that one may decide to omit recalculating the *paf* and use the production *paf* for all jackknife iterations. This also contributes to simplify the jackknife application.

Table A1
Smallest and biggest *paf* values
in jackknife simulation

province	true *paf* value[1]	panel 1		panel 2	
		Min *paf* value	Max *paf* value	Min *paf* value	max *paf* value
Newfoundland	0,37	0,35	0,37	0,38	0,40
Prince Edward Island	0,32	0,28	0,31	0,33	0,36
Nova Scotia	0,19	0,18	0,19	0,19	0,21
New Brunswick	0,34	0,32	0,33	0,34	0,35
Québec	0,25	0,25	0,25	0,25	0,26
Ontario	0,25	0,25	0,25	0,25	0,26
Manitoba	0,42	0,40	0,42	0,42	0,44
Saskatchewan	0,30	0,29	0,30	0,31	0,31
Alberta	0,26	0,25	0,26	0,26	0,26
British Columbia	0,25	0,25	0,25	0,26	0,26

[1] This simulation was produced using preliminary SLID data; this is why the *paf* values are a bit different from the production ones shown in Table 4.

A3- Optimising *paf* for yearly trend estimates

The yearly trend estimator can be described as follows: let $\hat{Y}_t = p_1\hat{Y}_{t,1} + p_2\hat{Y}_{t,2}$ be the cross-sectional estimate for a given variable of interest for the reference period *t*, where $\hat{Y}_{t,1}$ and $\hat{Y}_{t,2}$ are the estimates produced by panel 1 and 2 respectively at time *t*.

Let $\hat{Y}_t = p_1\hat{Y}_{t,1} + p_2\hat{Y}_{t,2}$ and $\hat{Y}_{t+1} = p_1\hat{Y}_{t+1,1} + p_2\hat{Y}_{t+1,2}$ be the estimates of the total at time *t* and *t+1* respectively. Here we assume that the *paf* are the same at *t* and *t+1*. The yearly trend between *t* and *t+1* is given by:

$$\begin{aligned} {}_{t+1}\hat{D}_t &= \hat{Y}_{t+1} - \hat{Y}_t \\ &= p_1(\hat{Y}_{t+1,1} - \hat{Y}_{t,1}) + p_2(\hat{Y}_{t+1,2} - \hat{Y}_{t,2}) \end{aligned}$$

The variance of ${}_{t+1}\hat{D}_t$ is given by:

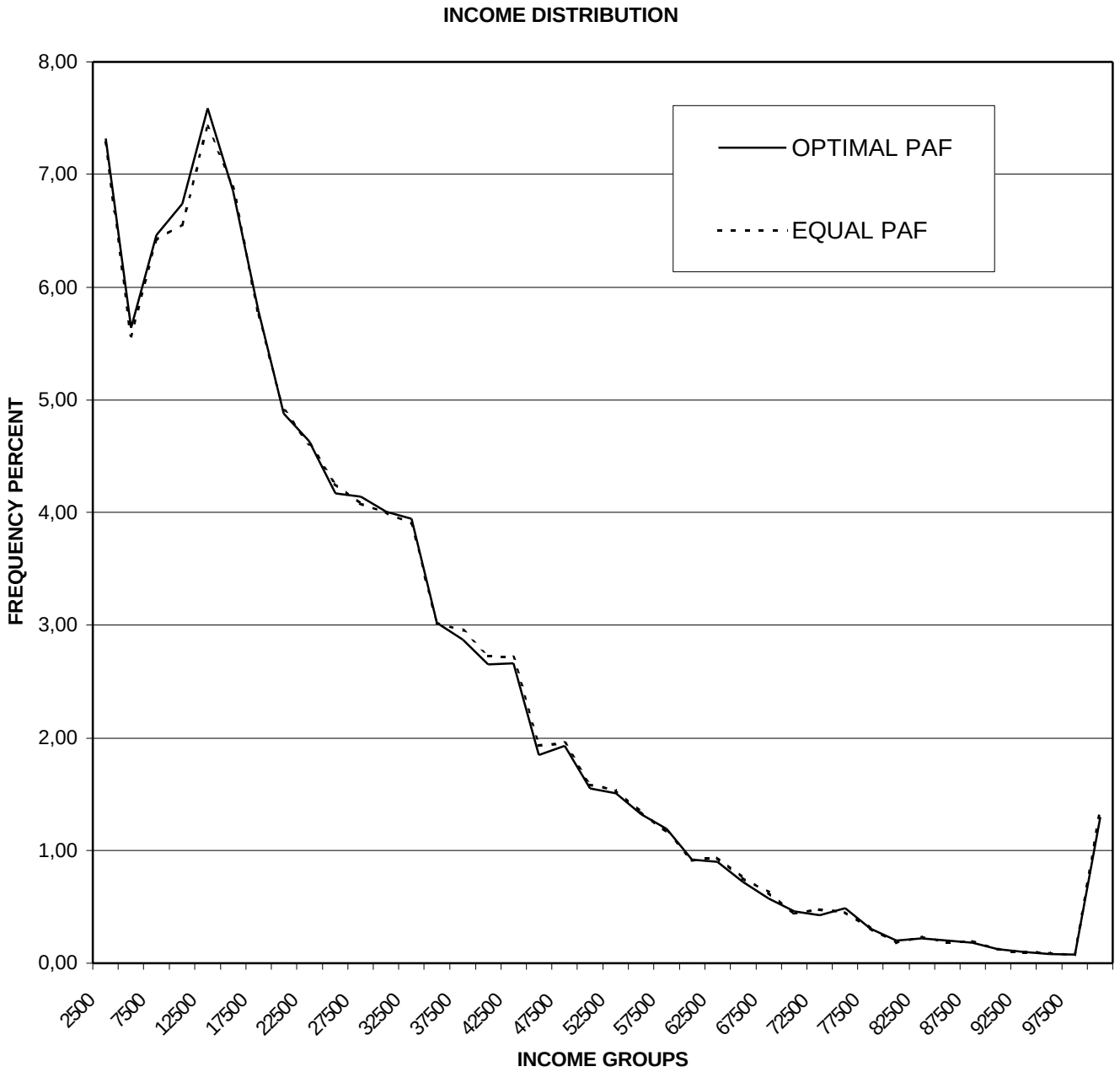
$$\begin{aligned} V({}_{t+1}\hat{D}_t) = {} & p_1{}^2(V(\hat{Y}_{t+1,1}) + V(\hat{Y}_{t,1}) - 2COV(\hat{Y}_{t+1,1}, \hat{Y}_{t,1})) \\ & + p_2{}^2(V(\hat{Y}_{t+1,2}) + V(\hat{Y}_{t,2}) - 2COV(\hat{Y}_{t+1,2}, \hat{Y}_{t,2})) \end{aligned}$$

It can be shown that the variance is minimised if

$$p_1 = \frac{V(\hat{Y}_{t+1,2}) + V(\hat{Y}_{t,2}) - 2COV(\hat{Y}_{t+1,2}, \hat{Y}_{t,2})}{\sum_{j=1}^{2} V(\hat{Y}_{t+1,j}) + V(\hat{Y}_{t,j}) - 2COV(\hat{Y}_{t+1,j}, \hat{Y}_{t,j})}$$

where $COV(\hat{Y}_{t+1,j}, \hat{Y}_{t,j})$ is panel *j* covariance between the two years; this covariance is always positive and should be large. Notice that when a panel in used for the first time, it does not have a covariance. In such a situation, only the oldest panel contributes to the covariance. Hence, the oldest panel has a bigger *paf*, and the estimate is more subject to attrition bias. One year later, both panels have a covariance and the *paf* must be recalculated which may lead to instability of the estimates and analytical problems.

APPENDIX B
NATIONAL AND PROVINCIAL ESTIMATES
AS PRODUCED USING OPTIMAL AND EQUAL *PAF*[1]



[1] The equality of the estimates produced by the two methods has not been tested except for those presented in Table 5.

APPENDIX C
HYPOTHESIS TESTING ON ESTIMATES

In determining which differences between production and equal paf estimates were statistically significant, the following approach was used to approximate the variance of those differences.

Let

$$\hat{t}_p = p_1\hat{Y}_1 + p_2\hat{Y}_2$$
$$\hat{t}_q = q_1\hat{Y}_1 + q_2\hat{Y}_2$$

be the estimates produced by two different sets of *paf*, p and q respectively. $\hat{Y}_i$ is the panel *i* estimate. For instance, *p* could represent the set of equal *paf* (0,50) and *q* the set of production *paf*. Consider the difference

$$\begin{aligned} D_{pq} &= t_p - t_q \\ &= (\hat{Y}_1 - \hat{Y}_2)(p_1 - q_1) \end{aligned}$$

If we assume that $V(\hat{Y}_1) \approx V(\hat{Y}_2)$, then it can be shown that

$$V(D_{pq}) = 2V(\hat{Y}_2)(p_1 - q_1)^2$$

In 1996 production, on averaging over province we have $p_1 - q_1 \approx 0.22$ so

$$V(D_{pq}) = 0{,}0968V(\hat{Y}_2) = 0{,}0968cv^2(\hat{Y}_2)\hat{Y}_2^2$$

where $cv(\hat{Y}_2)$ if the coefficient of variation of the second panel estimate. Finally, one decides that the difference is statistically significant at the 1% level if

$$z = \frac{|D|}{cv(\hat{Y}_2)\hat{Y}_2\sqrt{0.0968}} \geq 2.57$$

Note that this test is equivalent to test the null hypothesis: $\hat{Y}_1 = \hat{Y}_2$. Table C1 shows for some key variables the cv, the z value and the probability under the null hypothesis of observing such a z value.

Table C1
Hypothesis testing on equality of optimal and equal paf estimates

variable	Optimal paf	Equal paf	Relative Difference	Coefficient of Variation (%)	z	probability of a bigger z	cv source
Number of unattached individuals	3984199	4033703	1.24	2.3	1.74	0.0413	SLID jackknife
Number of size 2 families	3407517	3445625	2.24	2.5	1.44	0.0752	intrapolation from SLID
Number of size 3+ families	4777024	4742161	-2.63	2	1.17	0.1204	intrapolation from SLID
Number of married persons	12471961	12435254	-0.29	1	0.95	0.1721	SLID crude table
Number of single persons	6411378	6378680	-0.51	1.8	0.91	0.1812	SLID crude table
Number of separated persons	648956	727732	12.14	7	5.57	0.0000	SLID crude table
Number of persons with unknown marital status	192204	128209	-33.3	11.1	9.64	0.0000	SLID crude table
Number of persons in rural area	3434513	3430380	-0.12	2.7	0.14	0.4430	SLID crude table
Number of persons in 100k-500k area	4759717	4921432	3.4	2	5.46	0.0000	SLID crude table
Number of persons in 500k+ area	14151545	13939624	-1.5	0.7	6.88	0.0000	SLID crude table
Total earnings (in million $)	421029	425795	1.13	1.94	1.88	0.0304	SLID generalized function
Total investment (in million $)	23863	24617	3.16	12.33	0.82	0.2051	SLID generalized function
Total government transfers (in million $)	76467	76196	-0.35	3.07	0.37	0.3553	SLID generalized function
Total other money income (in million $)	44103	45033	2.11	3.56	1.90	0.0285	SLID generalized function
Total income (in million $)	563583	569553	1.06	1	3.40	0.0003	SLID jackknife
Average income: family	56955	57290	0.59	1.21	1.56	0.0591	SCF=0.69. SLID=SCF/0.57
Average income: unattached	24371	24821	1.84	2.21	2.69	0.0036	SCF=1.26. SLID=1.26/0.57
Average income: persons 16+	25347	25605	1.02	1	3.27	0.0005	SLID jackknife. SCF=0.57
Percentage of persons below LICO	18.6	18.01	-3.17	3.1	3.29	0.0005	SLID jackknife