Analytical Studies: Methods and References

Big Tax Data and Economic Analysis: Effects of Personal Income Tax Reassessments and Delayed Tax Filing

by Derek Messacar

Release date: January 11, 2018

Statistics Canada   Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service   1-800-263-1136
- National telecommunications device for the hearing impaired   1-800-363-7629
- Fax line   1-514-283-9350

Depository Services Program

- Inquiries line   1-800-635-7943
- Fax line   1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. 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.gc.ca under "Contact us" > "Standards of service to the public."

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.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.   not available for any reference period
..   not available for a specific reference period
...   not applicable
0   true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
p   preliminary
r   revised
x   suppressed to meet the confidentiality requirements of the *Statistics Act*
E   use with caution
F   too unreliable to be published
*   significantly different from reference category (p < 0.05)

Big Tax Data and Economic Analysis:
Effects of Personal Income Tax Reassessments and Delayed Tax Filing

by

Derek Messacar

Social Analysis and Modelling Division
Statistics Canada

January 2018

Acknowledgements

Table of contents

Abstract

Amid an increasing reliance on administrative tax data for economic analysis, the extent to which such data are confounded by income tax reassessments and delayed tax filing requires examination. This article provides novel insight into this issue using population records of initial and delayed Canadian tax filers from 1990 to 2010. The results show that 3.5% to 4.8% of tax filers delay filing their returns each year. However, the consequences of this behaviour are generally small, and do not bias estimates of income distributions, aggregate statistics, or inequality. These findings inform discourse about the relative merits of using administrative versus survey data for economic analysis.

Keywords: administrative tax records; survey data; income tax reassessment; delayed tax filing; economic analysis; income inequality.

Executive summary

This study investigates the extent to which income tax reassessments and delayed tax filing affect the reliability of Canadian administrative tax datasets used for economic analysis. The study is based on individual income tax records from the T1 Personal Master File and Historical Personal Master File for selected years from 1990 to 2010. These datasets contain tax records for approximately 100% of initial and all income tax filers, who submitted returns to the Canada Revenue Agency (CRA) before specific processing cut-off dates. The results of this analysis indicate that:

1. Each year, approximately 3.5% to 4.8% of individuals do not submit their tax returns to the CRA in time to be included in conventional datasets.

2. Delayed tax filing tends to be more prevalent among younger tax filers; residents of Ontario, Alberta, British Columbia, and the territories; non-residents; emigrants; low-income earners; and those with final tax balances close to zero.

3. Delayed tax filing is often a repeat behaviour. For example, among individuals who delayed filing their 2005 tax returns, approximately 34.4% of them also delayed filing their 2006 tax returns, and 21.3% of them delayed filing their 2007 tax returns. This may stem from individuals consistently filing several months after the cut-off dates, or filing multiple years of outstanding tax returns at once.

Although delayed tax filing occurs regularly, the consequences for economic analysis based on administrative tax records are generally small:

4. Income tax reassessments and delayed tax filing do not bias estimates of income distributions, aggregate income statistics, or top-income cut-offs obtained from tax data derived only from initial tax filers. This is true for employment earnings, Employment Insurance (EI) income, and Old Age Security income.

5. The only notable exception is business self-employment income. In this case, reassessments appear to be slightly more prevalent among initial tax filers, although the magnitude of the difference in reassessments between business and other sources of income is small. There are several possible explanations for this, including the difficulty of precisely calculating business self-employment income.

6. The probability of delayed tax filing is not highly correlated with changes in individuals' labour earnings or receipt of EI. For example, the prevalence of delayed tax filing for many individuals who had relatively constant labour market earnings from one year to the next was 2.2%. This compared with 3.6% among those whose earnings fell by 50% or more.

7. Top-income cut-offs are predicted to be biased upward in administrative tax datasets derived only from initial tax filers. The results of this analysis show that this bias is negligible and rarely exceeds 1%.

These findings are relevant for policy analysts, practitioners, and researchers who rely on the accuracy of individual income tax records. More broadly, the findings inform ongoing discussions about the relative merits of using administrative versus survey data for economic analysis.

1   Introduction

Over the past several decades, economic science has shifted toward greater reliance on empirical research (Einav and Levin 2014). Alongside the digital revolution and rise of personal computing, large-scale probability surveys were "the 20th-century answer to the need for wider, deeper, quicker, better, cheaper, more relevant and less burdensome official statistics" (Citro 2014, p. 138).[1] National surveys became a primary means of estimating unemployment, poverty, inflation, and other statistics of policy relevance, while serving as an important source of data for economic research (Meyer, Mok, and Sullivan 2015). However, in recent years, the quality of survey data has in some cases declined amid secularly declining response rates (Heffetz and Reeves 2016). Many economists now advocate greater use of administrative data or linked survey–administrative data to overcome this issue (Citro 2014; Varian 2014; Meyer, Mok, and Sullivan 2015; Heffetz and Reeves 2016; Jarmin and O'Hara 2016; Lane 2016).

Large administrative tax datasets ("big tax data") offer several advantages over conventional probability surveys. In particular, their size and granularity permit economic outcomes to be measured with precision, new patterns of behaviour to be identified, novel and innovative research designs to be implemented, and treatment effects of different policies to be estimated credibly across groups when such effects are heterogeneous (Einav and Levin 2013). With this in mind, it is not surprising that, among research articles published in some of the top economics journals over the past few years, the percentage of studies that use survey data has significantly declined, and the percentage of studies using restricted-access administrative data is increasing (Chetty 2012; Einav and Levin 2014).

Previous research has assessed the extent to which survey data measure economic outcomes reliably, benchmarked against administrative tax records.[2] A related issue—which has not yet received attention in this literature—is how accurately tax records measure economic outcomes. This issue is relevant because tax data may be confounded by two types of behaviours. First, tax filers may misreport income, either inadvertently or intentionally, to avoid paying higher income taxes given that tax systems in many countries operate by means of voluntary compliance. Income misreporting results in measurement error if the records are not reassessed, either by the tax filers themselves or by tax authorities, before statistical agencies compile the data. Second, some tax filers delay submitting their tax returns, which means that administrative datasets consist of a selected sample of initial tax filers whose records were available to statistical agencies at the time the data were compiled. In many countries—including Canada—no penalties are incurred for filing after the deadline if income taxes are not owed, hence big tax datasets may underrepresent relevant socioeconomic groups with comparatively low tax liabilities who have weak incentives to file on time. The consequences of income tax reassessments and delayed tax filing on the reliability of big tax data for economic analysis is an underexplored empirical issue, and is the focus of this analysis.

This study makes two contributions. First, the prevalence of delayed tax filing by individuals is assessed, and potential causes of such behaviour are investigated. To this end, the analysis uses the T1 Personal Master File (T1 PMF) and T1 Historical Personal Master File (T1 HPMF) for selected years spanning 1990 to 2010. These datasets, produced by the Canada Revenue Agency (CRA), offer detailed information about demographics, employment, income, and taxes and transfers for the populations of initial and delayed tax filers, respectively. The findings show that, each year, from 3.5% to 4.8% of filers delayed submitting their returns. Delayed filing was

---

1. For example, Duncan and Shelton (1978) and Harris-Kojetin (2012) review the expanded use of probability surveys for official statistics in the United States.
2. Lemieux and Riddell (2015) show that estimates of top-income shares in Canada are similar, using both census and administrative data. By contrast, Heffetz and Reeves (2016) caution against using surveys with low response rates to draw population inferences of outcomes such as labour force participation or unemployment rates in the United States.

most prevalent among younger tax filers; residents of Ontario, Alberta, British Columbia, and the territories; non-residents; emigrants; low-income earners; and those with final tax balances close to zero.

The second contribution of this study is an evaluation of the consequences of income tax reassessments and delayed tax filing for economic analysis that uses big tax data. The extent of the biases introduced in standard estimates of aggregate income and income inequality is considered. On balance, the results from this analysis are favourable: reassessments and delayed tax filing are sufficiently rare that biases are negligible. A notable exception is business self-employment income. Among initial tax filers, the aggregate value of such income observed in the T1 PMF is 97.5% of the corresponding amount observed in the T1 HPMF; this compares with 99.9% for employment earnings and 100.0% for Employment Insurance (EI) income and Old Age Security (OAS) income. Hence, initial tax filers' valuations of their business self-employment income are systematically being increased after their taxes are filed, and this is occurring to a greater extent than for other types of income. There are several explanations for this finding, such as business self-employment income being difficult to measure or tax evasion.

This paper proceeds as follows. The next section describes the data used in the study. Then, Sections 3 and 4 present the findings from the analyses of personal characteristics associated with delayed tax filing and the resulting implications of such behaviour for economic analysis that uses big tax data, respectively. Section 5 concludes.

2   Data and sample selection

This section begins by describing the datasets and defining what constitutes an initial versus delayed tax filer in the context of this study. The sample selections for the cross-sectional and longitudinal analyses are subsequently described.

2.1   Data

This study uses the T1 PMF and T1 HPMF tax registers, constructed by the CRA and obtained by Statistics Canada through a data partnership. The T1 PMF is a cross-sectional dataset consisting of the T1 personal income tax records of approximately 100% of Canadian tax filers who submitted their returns before an assessment date. It contains a wide set of information about these individuals, including demographics (e.g., year of birth, sex, marital status, province or territory of residence); income (e.g., employment, self-employment, investments, capital gains); and many federal and provincial amounts for taxes, transfers, credits, and allowances. The T1 PMF constitutes the source file from which Statistics Canada constructs several analytical datasets commonly used by academics, analysts, consultants, and governments, including the Canadian Employer–Employee Dynamics Database, Intergenerational Income Database, Longitudinal Administrative Databank (LAD), Longitudinal Worker File, and T1 Family File (T1FF). The T1 PMF also feeds into survey datasets, such as the Survey of Labour and Income Dynamics and the Longitudinal and International Study of Adults to the extent that survey respondents agree to have their income information retrieved from tax records rather than providing it through the questionnaire. The reliability of the T1 PMF is, therefore, of wide-spanning importance for many stakeholders who use these administrative datasets to inform policy discourse or conduct research in Canada.

The T1 PMF also provides the filing and assessment dates of each tax record. Appendix Table 1 reports the latest observed dates for the relevant years analyzed in this study, which indicate when individuals had to file to be included in the T1 PMF. For example, only those who filed their 2010 tax returns on or before December 22, 2011, are included. There is some variation in this cut-off date by year, which suggests that the sample composition of T1 PMF tax filers could be

changing over time. Such an effect is not likely a significant concern; however, because the cut-off dates are all very similar across years. Whether an individual files his or her taxes early, on time, late, or very late is likely correlated with many personal characteristics. However, for an increasingly narrow time interval, the degree to which this behaviour is expected to be random increases. Nevertheless, the differences in cut-off dates across years is relevant to note when drawing comparisons from this dataset over time. In this study, an "initial" tax filer is defined as an individual who appears in the T1 PMF and who filed before the cut-off date. (This concept differs from that of filing "on time," where a tax filer submits a return to the CRA on or before the cut-off date for when interest charges begin to accrue on outstanding tax balances owed, typically April 30 or June 15 of the next year for personal income tax filers or self-employed individuals, respectively.)

The T1 HPMF is a superset of the T1 PMF, containing the same wide set of information about demographics, income, taxes, transfers, credits, and allowances for approximately 100% of records filed within several years of the reference period. For example, a return for the tax year 2010 is observed in the T1 HPMF if it was filed on or before December 18, 2012, nearly one year later than the corresponding cut-off date for the T1 PMF. As a result, the T1 PMF and T1 HPMF are both snapshots of all tax returns received by the CRA within a fixed time interval of the end of the reference year, but the interval is wider for the T1 HPMF than for the T1 PMF. In contrast with the T1 PMF, the cut-off dates to be included in the T1 HPMF appear to vary substantially, shown in Appendix Table 1. This variation likely has an impact on changes in the composition of tax filers that needs to be taken into consideration when drawing inferences about how the results of this study vary over time. A tax return is considered "delayed" if it appears in the T1 HPMF but not in the T1 PMF.

One approach to resolving the problem of changing compositions in the T1 PMF and T1 HPMF is to condition the analysis on consistent cut-off dates. This would provide insight into how delayed tax filing has changed over time, for reasons that include the introduction and gradual adoption of electronic filing. Such analysis is left for future work, because this study generally relies on pooled data spanning multiple years. In addition, differences over time from the yearly data tend to be small. Because the T1 HPMF is a snapshot of tax returns received by a fixed cut-off date, records that are received after that date are never included. There could be a non-trivial amount of income missing from the T1 data even after several decades, such as what the CRA did not detect through audits or what tax filers did not revise (e.g., income that remains undeclared). The effects of undetected income misreporting and non-tax filing on the reliability of tax data for economic analysis are beyond the scope of this study but represent a promising direction for future work.

Together, the T1 PMF and T1 HPMF constitute the only source of data for investigating the effects of income reassessments and delayed tax filing on the quality of administrative datasets in Canada. A key feature of such data is that they are designed for tax purposes and not typically with economic research in mind. This study addresses the reliability of Canadian tax records to be used in conducting innovative research and, more broadly, contributes to the growing discussion in academic research about the relative merits of using survey versus administrative data for economic analysis.

2.2   Sample selection

When this study was undertaken, the most recent years of tax data available were 2013 for the T1 PMF and 2011 for the T1 HPMF. With this in mind, the following conditions are imposed to make the analysis tractable, given the large size of the datasets being used. First, the analysis primarily centre on the tax years 1990, 1995, 2000, 2005 and 2010, which span a wide range of cohorts, time periods, and macroeconomic conditions in Canada. Hence, this study offers a meaningful snapshot of the causes and consequences of income tax reassessments and delayed tax filing for multiple years spanning nearly the full range of available data. Second, although the

administrative records provide rich information on a wide array of individual characteristics, the following variables are primarily used throughout this study: year of birth, sex, marital status, employment income, business self-employment income, EI income, OAS income, total income, disability tax deductions, and tuition credits.[3] These personal characteristics and sources of income are used commonly in economic analysis and are of direct relevance for policy discourse concerning such issues as demographic change, employment estimates, worker mobility, income inequality, and more.[4]

Table 1 shows the numbers of initial and delayed tax filers by year, from the repeated cross-sectional data. Each year, approximately 3.5% to 4.8% of tax filers are observed in the T1 HPMF but not in the T1 PMF and are, therefore, considered to be delayed tax filers. Although the percentage of delayed tax filers is the highest in 1990, this may arise from the fact that electronic filing was not introduced on a national basis until the 1993 tax year (its adoption was even more gradual). On balance, although the prevalence of delayed tax filing is low in aggregate, there are approximately one million tax filers who are known to be excluded from the T1 PMF annually.

Table 1
Prevalence of initial and delayed tax filing, selected years, 1990 to 2010

	Number of initial tax filers	Number of delayed tax filers	Total number of tax filers	Percentage of delayed tax filers
	count			percent
1990	18,566,069	941,602	19,507,671	4.8
1995	20,504,412	838,133	21,342,545	3.9
2000	22,189,409	1,091,789	23,281,198	4.7
2005	23,772,773	1,115,600	24,888,373	4.5
2010	25,371,932	908,038	26,279,970	3.5

Note: Results are based on cross-sectional data. An initial tax filer is an individual who appears in the T1 Personal Master File and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

A limitation of the repeated cross-sectional data is that longitudinal effects of income tax reassessments and delayed tax filing cannot be investigated. To address this issue, a panel of tax filers who are observed in the T1 HPMF in every year from 2005 to 2010 was constructed to assess the effects of various life events—changes in marital status, migration, and income shocks—on delayed tax filing, as well as the prevalence of repeat delayed filing. This time period was chosen to provide a longitudinal analysis that pertains to the most recent cohort of tax filers. It is important to note that the restriction that individuals appear in the T1 HPMF in every year over this time period means the panel dataset contains a reduced sample of tax filers. This restriction ensures that the longitudinal analysis is not affected by changes in the composition of the data over time, because individuals may not need to file taxes in every year or they may choose to file after the date when the CRA created the T1 HPMF data. Approximately 85.3% of the T1 HPMF tax records satisfy this selection criterion.

---

3. Business self-employment income refers to net income (i.e., income minus expenses) from an unincorporated business, which is a component of self-employment income.
4. Despite the availability of many demographic variables, the T1 PMF and T1 HPMF data do not provide data on whether the tax filer is an immigrant. This variable would be a useful addition to the analysis, as the economic well-being of immigrants is an important issue in both academic research and policy discourse. Future research could expand upon this work by linking the T1 records to the Immigrant Landing File.

3   Delayed tax filing behaviour

This section investigates the associations between delayed tax filing and observed personal characteristics, with the objective of providing insight into potential underlying causes of such behaviour. To this end, the analysis proceeds in three stages. The first stage documents the relationship between delayed tax filing and various demographic and labour market characteristics based on the repeated cross-sectional data, for selected years from 1990 to 2010. Second, using longitudinal data spanning 2005 to 2010, the effects of various life events on the prevalence of delayed tax filing are assessed. The section concludes with an examination of the relationship between delayed tax filing and individuals' income tax balances due or tax refunds.

3.1   Personal characteristics

Table 2 presents descriptive statistics for the populations of initial and delayed tax filers. The results indicate that delayed tax filers tend to be younger than initial tax filers (40.1 years old versus 45.7 years old, on average), more likely to be male (59.7% versus 49.2%), and less likely to be married or in common-law relationships (36.9% versus 56.2%). Delayed tax filers are much less likely to have OAS income (4.7% versus 16.9%), although this is likely due to the age gap between the two groups. In contrast, delayed tax filers are slightly more likely to have business self-employment income (13.4% versus 7.2%).

Table 2
Characteristics of initial and delayed tax filers, 1990 to 2010 (combined)

	Initial tax filers	Delayed tax filers
	years	
Demographic characteristics		
Average age	45.7	40.1
	percent	
Sex		
Female	50.8	40.3
Male	49.2	59.7
Marital status		
Single	27.8	45.0
Married or common-law	56.2	36.9
Other	16.0	18.1
Income sources		
Employment	66.8	67.7
Business self-employment	7.2	13.4
Employment Insurance	11.8	11.0
Old Age Security	16.9	4.7
Total	98.3	94.1
Tax credits and allowances		
Disability deduction	2.1	1.3
Tuition credits	8.7	7.5
	2010 constant dollars	
Average conditional income		
Employment	38,000	32,350
Business self-employment	11,500	12,450
Employment Insurance	6,000	6,500
Old Age Security	5,650	5,400
Total	38,950	32,250

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010. An initial tax filer is an individual who appears in the T1 Personal Master File and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.
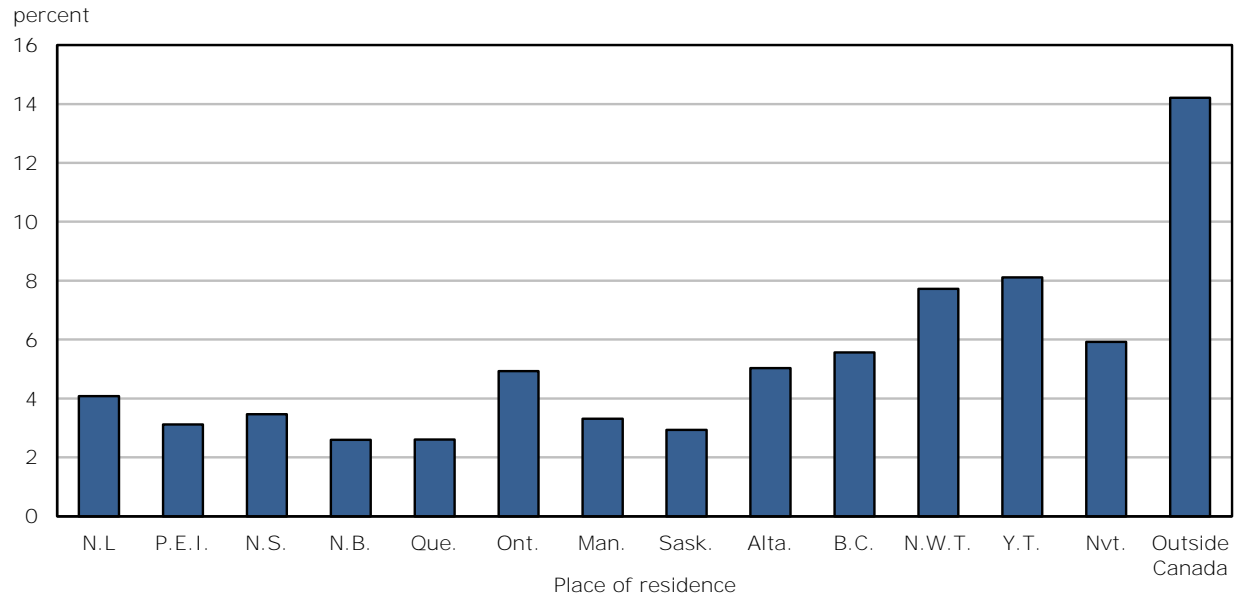**Sources:** Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Despite these demographic differences, both groups are approximately equally as likely to have employment income or to collect EI. Conditional on the incomes being strictly positive, both groups also have similar earnings from business self-employment income, EI and OAS. The only type of income that differs is employment earnings, where initial tax filers earn more, on average, than delayed tax filers ($38,000 versus $32,350, respectively, in 2010 constant dollars).

To further explore how the incidence of delayed filing varies across groups of tax filers, Charts 1 to 3 plot the shares of delayed tax filers by jurisdiction, age group, and income levels, respectively. First, delayed tax filing is the most prevalent among residents of Ontario, Alberta, British Columbia, the territories, and non-residents, shown in Chart 1. A possible explanation for these differences is that some provinces and territories have more tax filers in self-employment than others, which affects delayed tax filing, discussed later. This finding has implications for economic analysis of these tax filer groups. For example, a recent study by Finnie, Gray and Zhang (2016) show that non-residents are less likely than residents to enter into, and more likely to exit out of, the receipt of Guaranteed Income Supplement (GIS) income than residents, based on an analysis of the LAD. These estimates could be mitigated or exacerbated by the inclusion of non-resident delayed tax filers depending on the correlation between GIS usage and the timing of tax filing.

Second, Chart 2 illustrates that the prevalence of delayed filing declines with age, which may arise for reasons including gradual learning or an increasing incentive to claim tax credits and allowances. For example, tax filing is encouraged among GIS recipients to avoid the need to fill out benefit renewal forms. This finding may also help explain the interprovincial differences in delayed tax filing observed in Chart 1 to the extent that some provinces and territories—notably in Atlantic Canada—tend to have older populations, on average, than others. Note that the analysis does not distinguish between alive and deceased tax filers (e.g., taxes filed by a spouse or relative on behalf of a deceased person). The slight upturn in the prevalence of delayed tax filing at older ages, shown in Chart 2, may stem from several factors. It may be that tax liabilities decrease with age, which affects the balance due or refund and, in turn, the incentive to file taxes in a timely manner. Other explanations include changes in sample composition by age group, increasing tax illiteracy, cognitive decline, or deceased tax filers being more likely to file late. In contrast, the slight drop in the probability of delayed tax filing among those aged 20 to 24 (relative to those aged 0 to 19 and 25 to 29) may stem from the incentive to file to take advantage of tuition credits for postsecondary education. As Table 2 shows, initial tax filers have a slightly higher prevalence of tuition credit claimants than delayed tax filers.

Chart 1
Prevalence of delayed tax filing, by place of residence, 1990 to 2010 (combined)

percent

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005 and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Chart 2
Prevalence of delayed tax filing by age group, 1990 to 2010 (combined)

percent
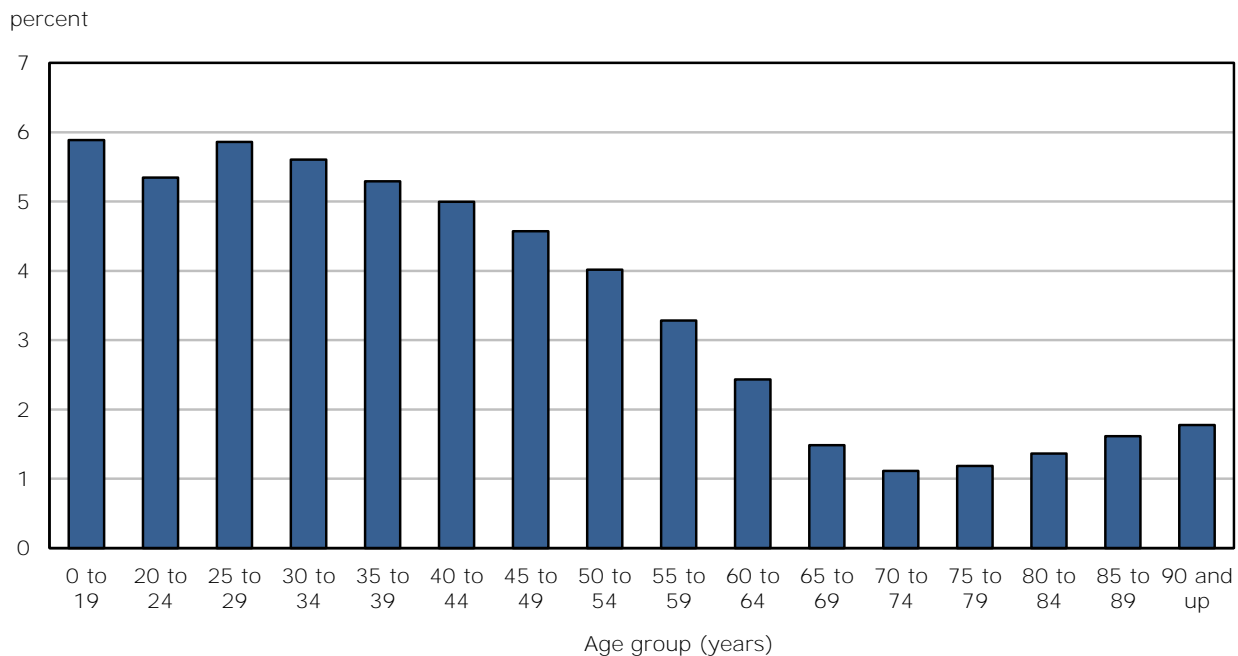
Note: Results are based on pooled cross-sectional data for the years 1990, 1995, 2000, 2005 and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Third, whereas Table 2 shows that delayed tax filers have lower employment earnings than initial tax filers, Panel A of Chart 3 shows that this effect is predominantly driven by workers with very low earnings. In contrast, the prevalence of delayed filing is nearly uniform for every employment earnings bracket above $10,000. Because low-income tax filers are the least likely to have a positive balance outstanding with the CRA, the incentive for these individuals to file in a timely

manner to avoid interest charges is weak. Furthermore, as Panel B of Chart 3 shows, tax filers with zero total income are the most likely to delay filing, although most individuals declare income from at least one source on their tax returns.

Chart 3
Prevalence of delayed tax filing, by employment earnings and total income brackets, 1990 to 2010 (combined)
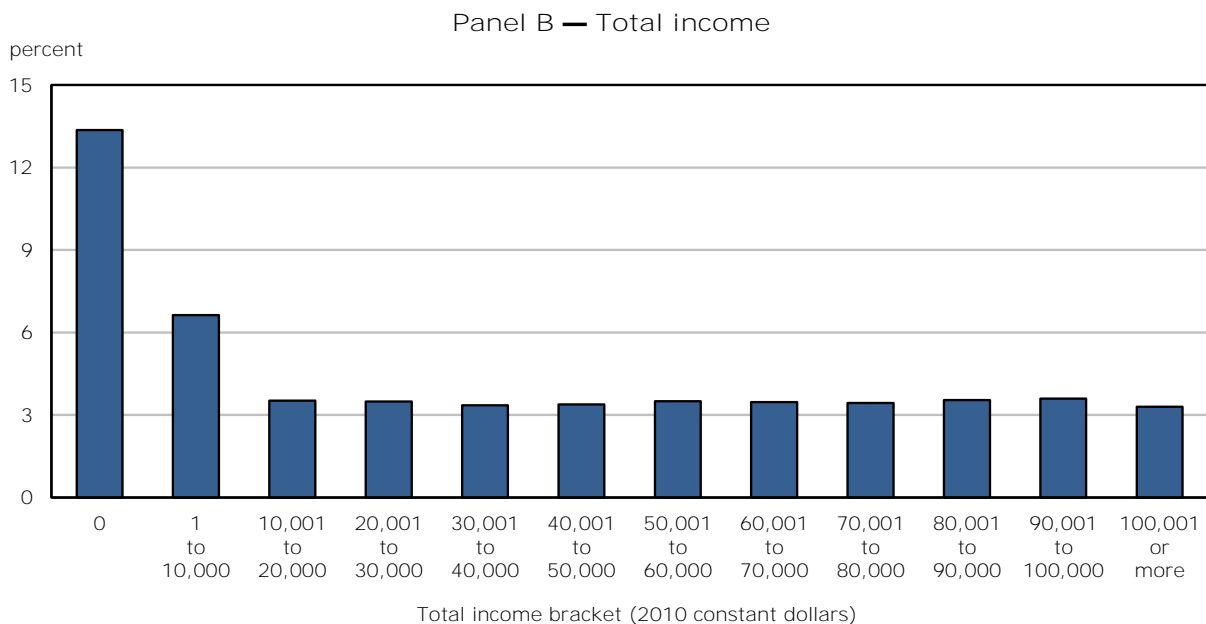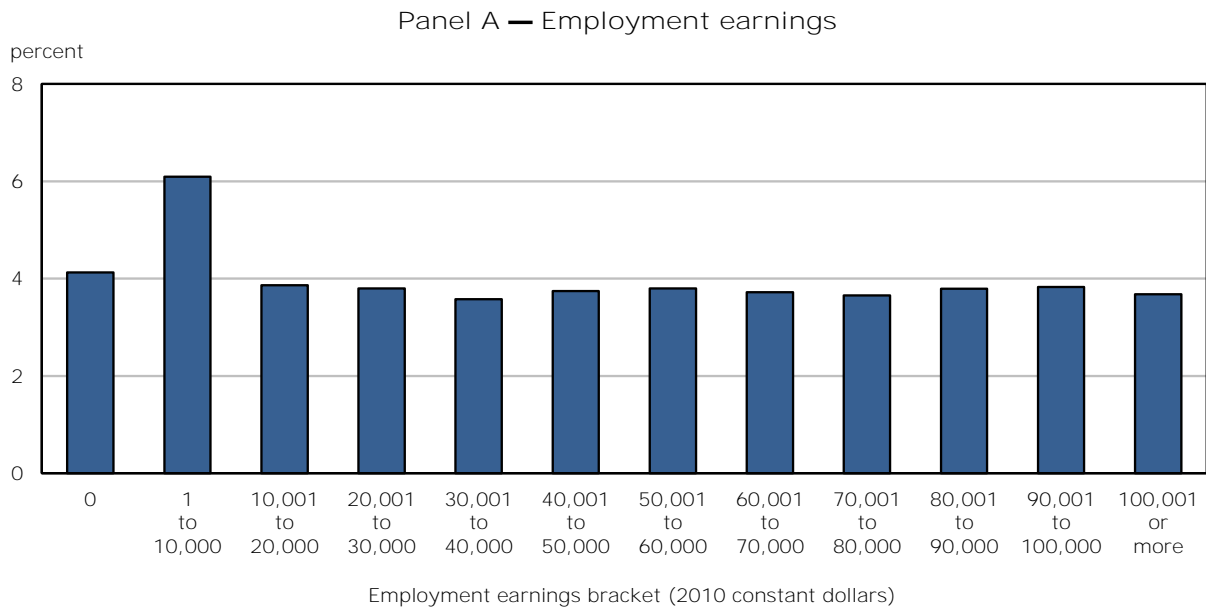
Panel A — Employment earnings

percent

Employment earnings bracket (2010 constant dollars)

Panel B — Total income

percent

Total income bracket (2010 constant dollars)

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005 and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Taken together, the results of this analysis suggest that, on balance, many characteristics of tax filers in the T1 HPMF are adequately measured in the T1 PMF. Although delayed tax filing is slightly more prevalent in certain age groups and jurisdictions, this behaviour appears to be

homogeneous across many personal characteristics. A few notable exceptions include younger tax filers, very low employment income earners, and non-residents.

3.2 Longitudinal analysis

Biases when measuring economic outcomes in the T1 PMF resulting from delayed tax filing may be especially prevalent among certain groups. For example, individuals who experienced a job loss or moved between provinces or territories could be prone to delay filing while they undergo these adjustments, the result being that they are underrepresented in the T1 PMF.

To explore this issue, Table 3 shows the prevalence of delayed tax filing based on the panel of individuals who are observed in the T1 HPMF every year from 2005 to 2010. Although this sample is not necessarily representative of the full population of Canadian tax filers, the sample restriction that individuals are observed repeatedly is necessary to conduct a longitudinal analysis, as discussed earlier. Moreover, this restriction corresponds to an individual-level fixed effect model specification that would typically be employed in panel data to estimate the impacts of life shocks on economic outcomes. The first column of Table 3 shows that, among the full sample of tax filers in the panel, the incidence of delayed tax filing is 2.2%.

Table 3

Prevalence of delayed tax filing, by selected life events in past year, 2005 to 2010 (combined)

Life event in past year	Event frequency	Percentage of delayed tax filers
	percent	
Total	…	2.2
Change in marital status		
Single to married or common-law	1.60	3.2
Married or common-law to separated or divorced	0.85	5.6
Married or common-law to single	0.29	9.1
Married or common-law to widowed	0.37	1.3
Migrant		
Interprovincial migrant	0.98	6.1
Emigrant	0.02	17.2
Change in employment earnings		
Increased by 50% or more	8.28	3.1
Increased by 25% or more	14.15	2.9
Increased by 10% or more	23.16	2.8
Changed by -9% to 9%	27.93	2.2
Decreased by 10% or more	15.05	3.1
Decreased by 25% or more	10.45	3.3
Decreased by 50% or more	7.50	3.6
Change in Employment Insurance receipt		
New recipient	4.42	2.8
Former recipient	4.07	3.1

… not applicable

Note: Results are based on a sample of tax filers observed every year from 2005 to 2010. A delayed tax filer does not submit an income tax return to the Canada Revenue Agency before the cut-off date.

Source: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

This analysis considers the following life events: change in marital status, migration, and employment income shocks. First, delayed tax filing is slightly more prevalent among individuals who entered into marriage or a common-law relationship in the past year (3.2%) relative to the sample average. This behaviour may arise for many reasons, such as time constraints, coordination difficulties, or an increased complexity associated with understanding the relevant parameters of the tax code. Similarly, individuals who were married or in common-law

relationships in the past year but become single, separated or divorced—which occurs approximately 1.1% of the time—are significantly more likely to delay filing. The prevalences of delayed tax filing are 5.6% for recently separated or divorced tax filers and 9.1% for tax filers who recently became single. Although such behaviour is relatively uncommon, studies that seek to understand the effects of separation and divorce on labour market outcomes would benefit from taking a medium-run or long-run perspective—such as that used by LaRochelle-Côté, Myles, and Picot (2012) for studying the effects of widowhood and divorce in Canada based on the LAD—to ensure that measurable short-run effects are not driven by sample selectivity.

A limitation of administrative data is that job-related transitions—separations, work interruptions, changes in occupation, changes in hours worked (e.g., hours worked, full-time versus part-time, full-year versus part-year), changes in hourly wage—cannot be observed. However, by exploiting the longitudinal component of the data, an investigation of how delayed tax filing coincides with income shocks or take-up of EI benefits is possible. Table 3 shows that the prevalence of delayed tax filing is 2.2% among individuals whose earnings changed by less than 10%, equivalent to the full-sample average. The prevalence of delayed filing is very homogeneous for individuals who experienced both increases and decreases in their earnings of 10% or more, ranging from 2.8% to 3.6%. Although these values are slightly larger than the overall average, delayed tax filing was not particularly high among individuals who experienced large income fluctuations. Similarly, neither becoming an EI recipient nor ceasing to receive these benefits in the past year were associated with the prevalence of delayed taxfiling, these being 2.8% and 3.1%, respectively. Overall, these findings suggest that the full population of tax filers who experience labour income shocks are well represented by the T1 PMF.

Exploiting the longitudinal structure of these data, Chart 4 considers the extent to which the same individuals delay filing repeatedly. The results suggest that such behaviour is common. For example, among individuals who delayed filing their 2005 tax returns, approximately 34.4% also delayed filing their 2006 tax returns and 21.3% delayed filing their 2007 tax returns. Whether this behaviour is the result of individuals repeatedly filing taxes too late to be captured by the T1 PMF data, or filing multiple years of outstanding tax returns at once, cannot be determined by this analysis.

Chart 4
Prevalence of repeated delayed tax filing, by year, 2005 to 2010

percent

Years since base period

— 2005 — 2006 — 2007 — 2008
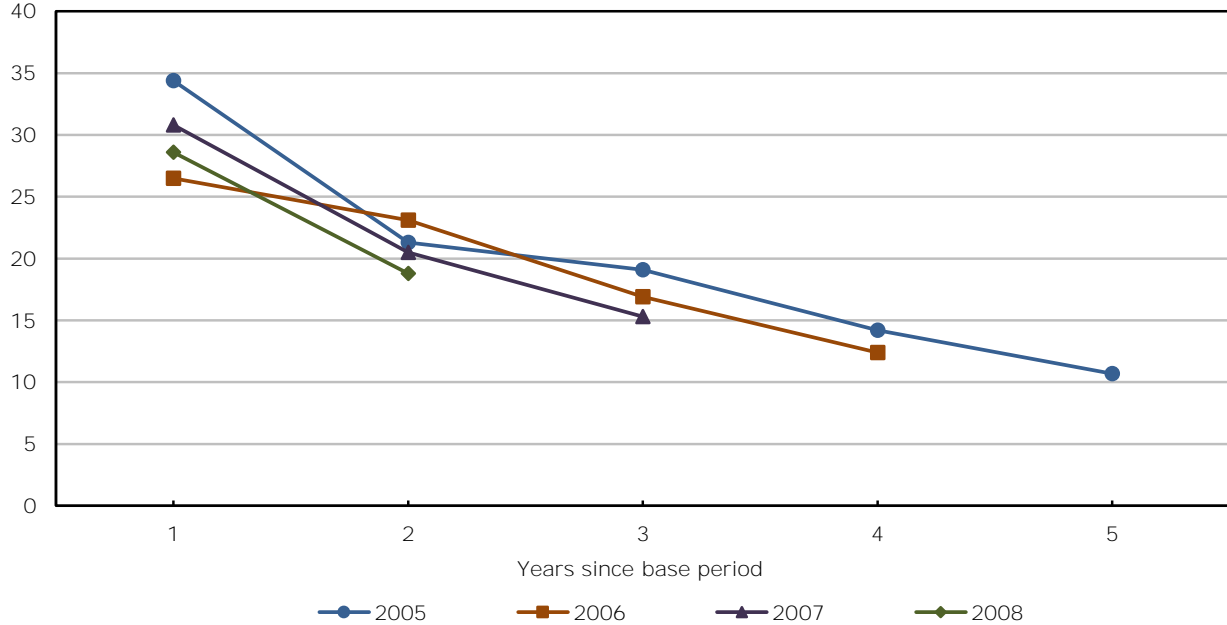
Notes: Results are based on the sample of tax filers observed every year from 2005 to 2010 inclusive, and show the percentage who delayed filing one, two, three, four, and five years after being delayed in the base year.
Source: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

3.3   Delayed tax filing and tax balances due

As noted, delayed tax filing was most prevalent among tax filers with (close to) zero income tax balances due, likely because no interest penalties are incurred when a balance is not owed, nor is there an incentive to file promptly when a refund is not anticipated. To investigate this issue empirically, Chart 5 plots the relationship between the value of the balance due and the probability of delayed filing, for balances ranging from $3,000 owed to $3,000 expected as a return. Consistent with expectations, the results show that individuals with balances closer to zero are indeed the most likely to delay filing their taxes.

Chart 5
Prevalence of delayed tax filing by tax balance, selected years,
1990 to 2010 (combined)
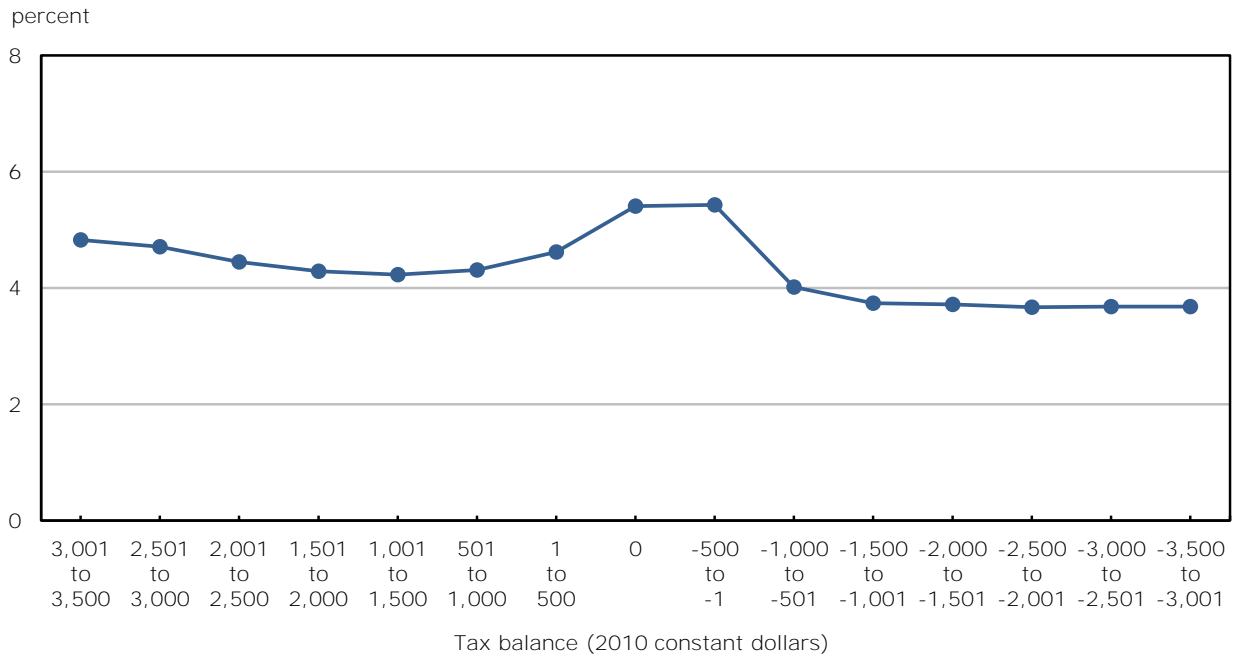
percent

Tax balance (2010 constant dollars)

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

To explore this issue further, Chart 6 plots the relationship between the balance due and the prevalence of delayed filing by level of total income: $20,000 or less, from $20,001 to $75,000, and more than $75,000. For those in the middle-income and high-income categories, the prevalences of delayed tax filing appear relatively homogeneous irrespective of the balance due, although those with balances close to zero are slightly more inclined to delay filing. For individuals in the low-income group, the increased prevalence of delayed tax filing at balances close to zero continues to hold, but the overall relationship is "U-shaped." Although the factors behind this relationship are unclear, low-income tax filers who have large balances due may delay filing in order to delay making a payment, irrespective of the interest costs. The fact that low-income tax filers who are owed a refund tend to delay filing is striking, and may be driven by issues of tax illiteracy.

Chart 6
Prevalence of delayed tax filing by total income and tax balance,
selected years, 1990 to 2010 (combined)

percent

Tax balance (2010 constant dollars)

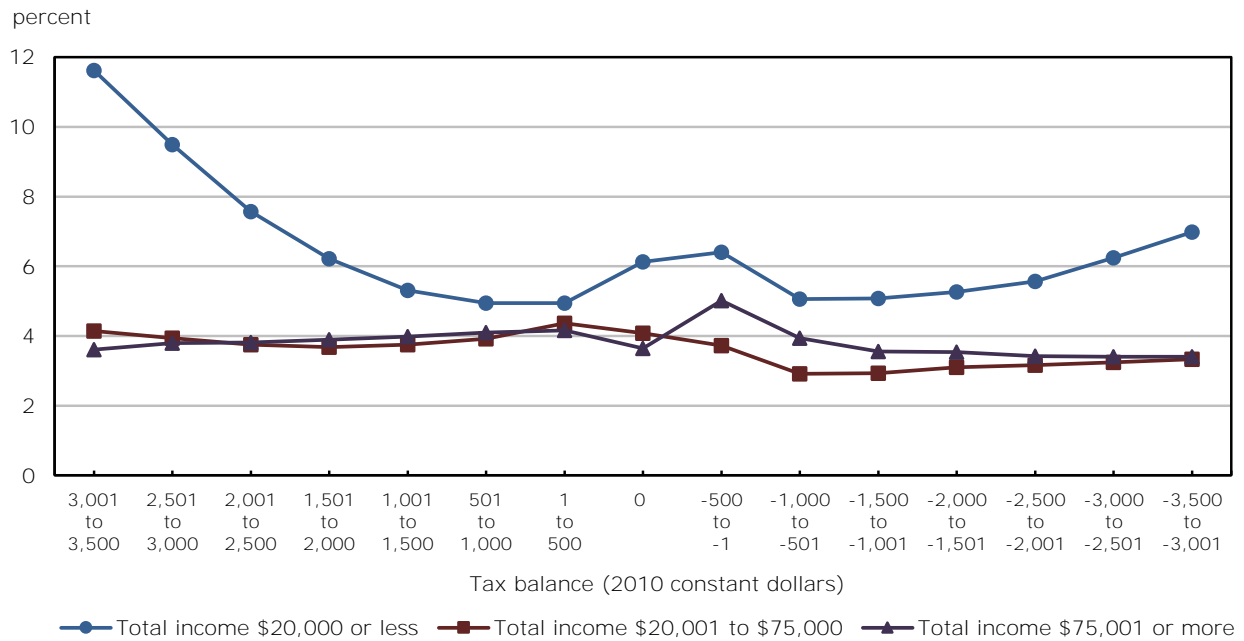— Total income $20,000 or less   — Total income $20,001 to $75,000   — Total income $75,001 or more

Note: The results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010.
Source: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

These results relate to empirical research that exploits tax refund data to draw inferences about consumer theory. For example, based on data from the Internal Revenue Service (IRS) of the United States, Feldman (2010) investigates the extent to which tax filers respond to exogenous changes in their tax refunds through savings adjustments to individual retirement accounts. To this end, the author exploits a 1992 reform that decreased federal income tax withholding rates, which shifted the timing of income tax payments forward while leaving total tax liabilities unchanged. The analysis shows tax filers have a greater marginal propensity to save out of the lump-sum income tax refunds than out of normal flows of funds, suggestive of a mental accounting effect. Using similar data, Rees-Jones (2014) analyzes the distribution of income tax refunds (or balances due) in the United States from 1979 to 1990, and considers the implications for loss aversion. In particular, the author posits that a loss-averse tax filer will use tax shelters to manipulate total tax liabilities to a greater extent when a balance is due than when a refund is owed, resulting from a discreetly steeper marginal utility of a dollar under loss framing. This behaviour is predicted to result in excess mass ("bunching") at the gain/loss threshold where the tax refund is exactly zero, which the author shows is consistent with actual behaviour observed in IRS tax data. These findings raise the question of whether such effects as bunching or savings responses to income tax refunds are in some way affected by changes in the composition of observed tax filers as a result of delayed tax filing. At a minimum, these findings are important to note for future research seeking to extend this line of inquiry to the Canadian context.

4 Implications for economic analysis

This section investigates the influence of income tax reassessments and delayed tax filing on the accuracy and reliability of T1 PMF estimates of income distributions and aggregate statistics at federal and provincial levels. Then, a closer look is taken at the effects of income tax reassessments on sources of income for which reassessments may be especially prevalent: commissions, farming, fishing, professional, and rental. Last, the implications of reassessments and delayed filing for the measurement of income inequality using T1 PMF data are examined.

4.1 Income distributions

The descriptive statistics in Table 2 show that delayed tax filers tend to have less labour market income than initial tax filers. To explore this issue in more detail, Chart 7 plots the distributions of employment earnings for initial, delayed, and all (initial plus delayed) tax filers using the repeated cross-sectional data, over a range of income from $1 to $100,000 (2010 constant dollars). Panel A indicates that the difference in the average earnings of initial and delayed tax filers arises because a large share of delayed tax filers—19.3%—have earnings less than $2,500—compared with 9.4% for initial tax filers. Although this discrepancy shifts the distribution of earnings downward for delayed tax filers, the distributions for both groups are otherwise very comparable.

Because delayed tax filers with earnings less than $2,500 represent a small fraction of all tax filers, the effect of using the T1 PMF to infer the earnings distribution of all tax filers is negligible, as shown in Panel B, which compares initial and all tax filers. The difference in the share of individuals with earnings less than $2,500 based on the T1 PMF versus T1 HPMF data is only 0.4%, and the earnings distributions of the two groups closely overlap above this threshold.

Chart 7
Distribution of employment earnings, by type of tax filer, selected years,
1990 to 2010 (combined)
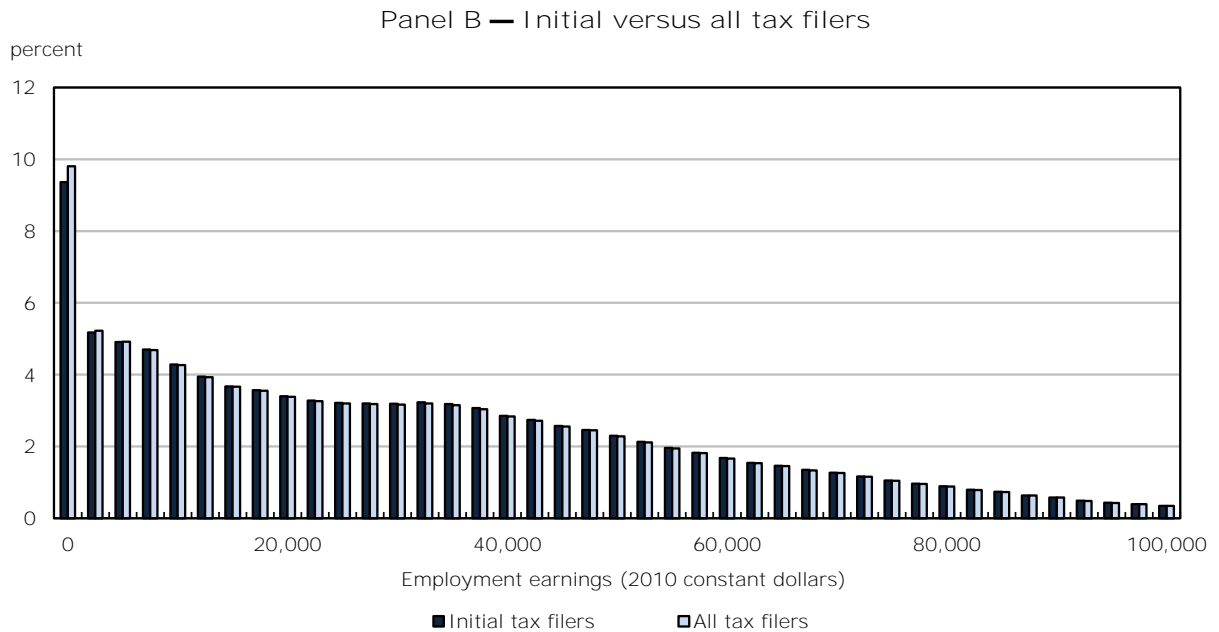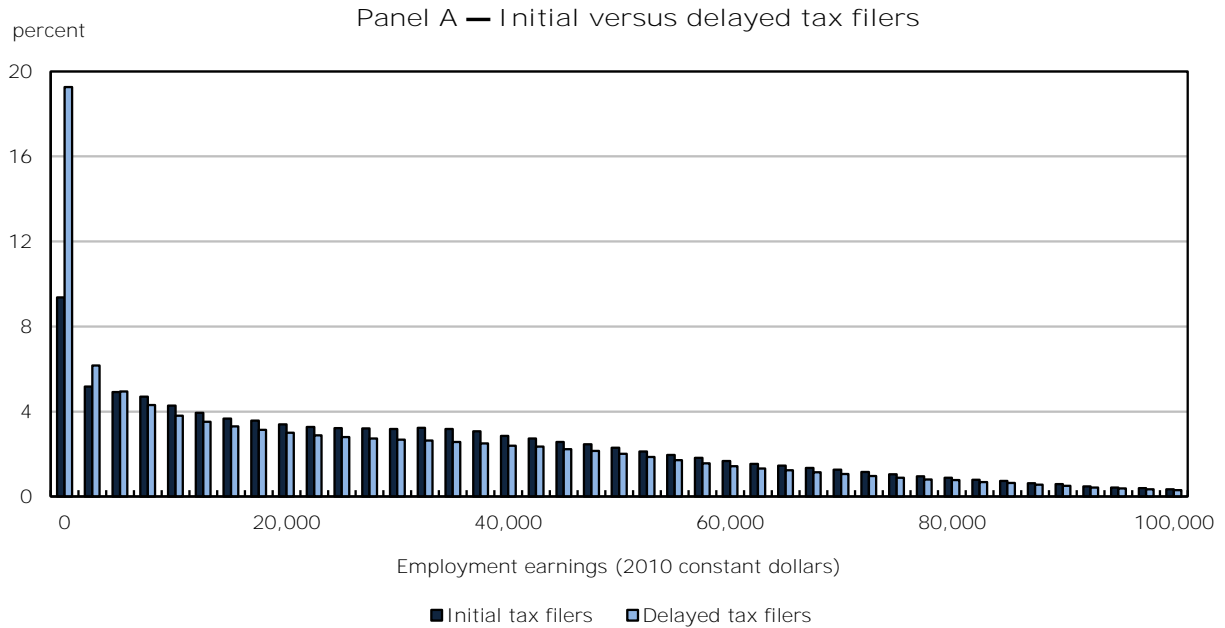
Panel A — Initial versus delayed tax filers

percent

Employment earnings (2010 constant dollars)

■ Initial tax filers   □ Delayed tax filers

Panel B — Initial versus all tax filers

percent

Employment earnings (2010 constant dollars)

■ Initial tax filers   □ All tax filers

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Chart 8 shows that, in contrast with the previous findings, delayed tax filers tend to have slightly higher business self-employment income than initial tax filers. In Panel A, the distribution of business self-employment income is right-shifted for delayed tax filers relative to initial tax filers over the full range of income from $1 to $100,000. This indicates that the T1 PMF slightly understates business self-employment income, but that the difference in the distributions of business self-employment income for initial versus all tax filers is negligible, as shown in Panel B.

Chart 8
Distribution of business self-employment income, by type of tax filer, selected years, 1990 to 2010 (combined)
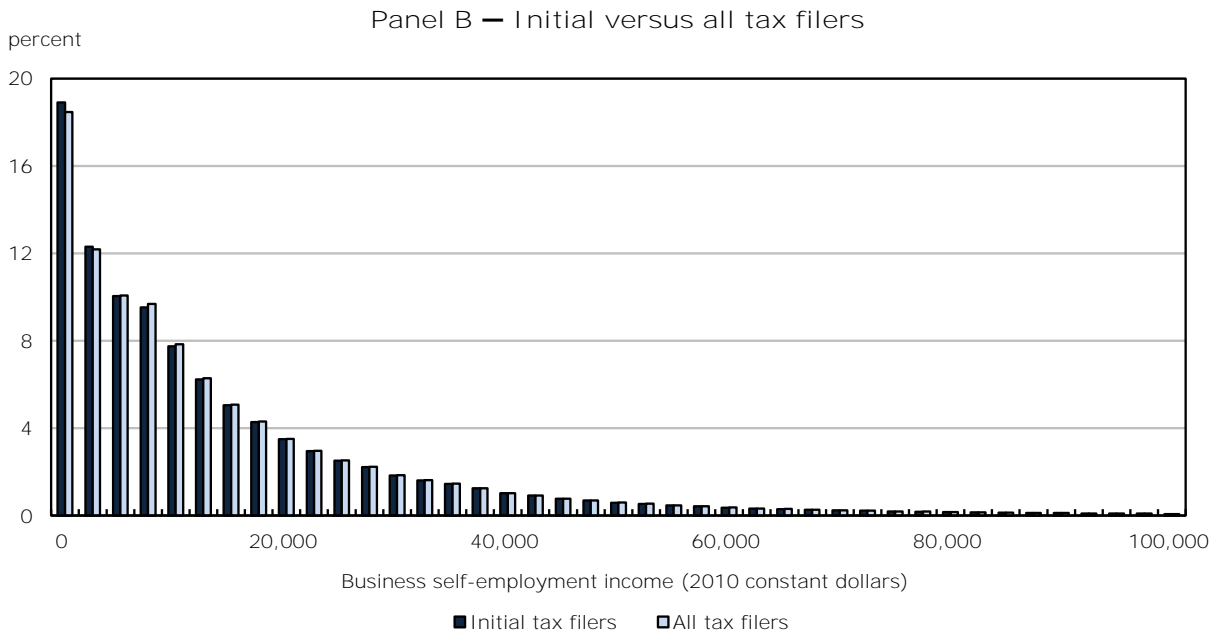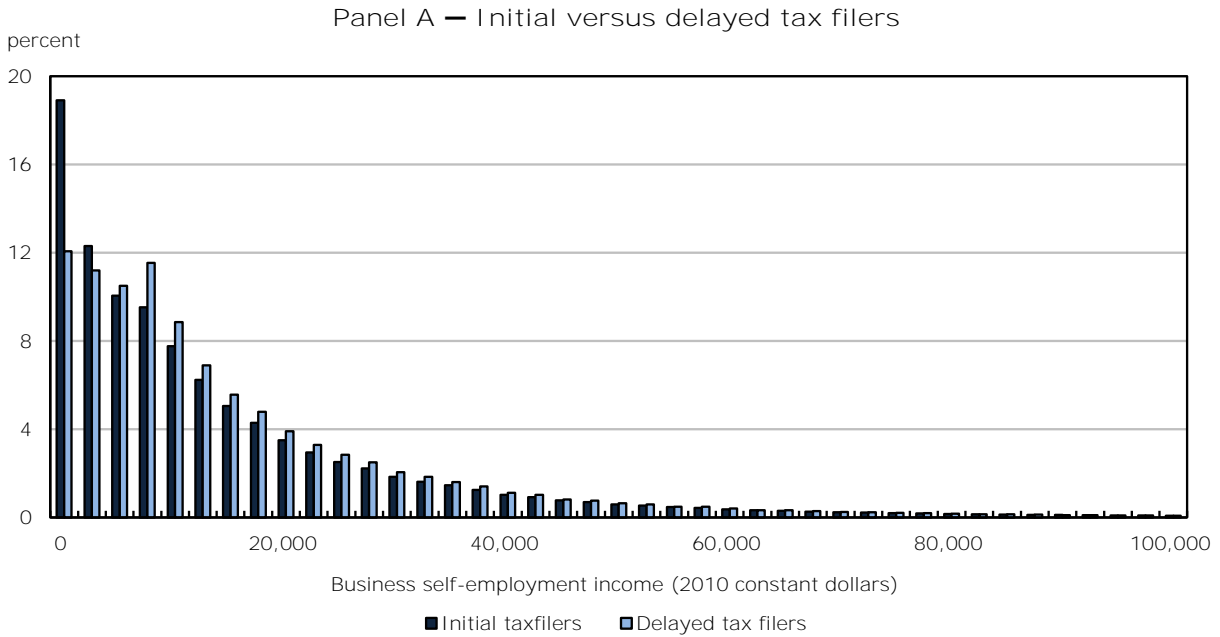
Panel A — Initial versus delayed tax filers

Business self-employment income (2010 constant dollars)

■ Initial taxfilers   □ Delayed tax filers

Panel B — Initial versus all tax filers

Business self-employment income (2010 constant dollars)

■ Initial tax filers   □ All tax filers

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

Last, the distributions of total income for initial and delayed tax filers are presented in Chart 9. The difference between the two groups mirrors that of employment earnings, where delayed tax filers with low income are underrepresented in the T1 PMF data, shown in Panel A; the effect of omitting delayed tax filers to infer the distribution of total income for all tax filers is negligible, as shown in Panel B. The comparability of these results with employment earnings likely arises because the major source of income for many tax filers is the labour market.

Chart 9
Distribution of total income, by type of tax filer, selected years, 1990 to 2010 (combined)
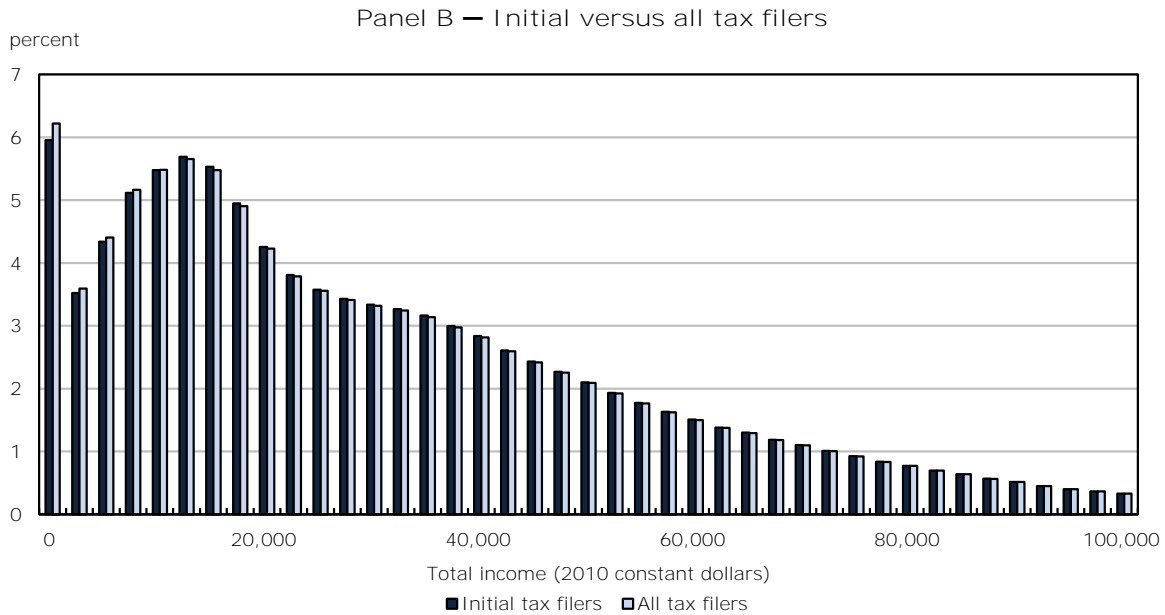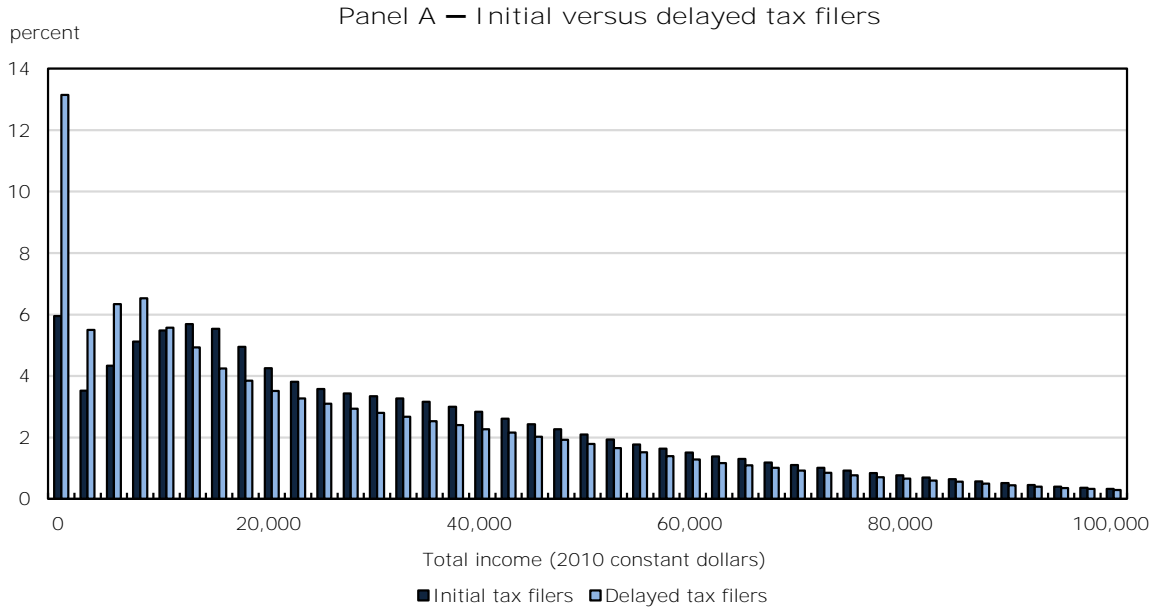
Panel A — Initial versus delayed tax filers

percent

Total income (2010 constant dollars)

■ Initial tax filers ■ Delayed tax filers

Panel B — Initial versus all tax filers

percent

Total income (2010 constant dollars)

■ Initial tax filers ☐ All tax filers

Note: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

4.2   Aggregate income statistics

Although delayed tax filing may have little effect on inferences of the distribution of income, its total effect on aggregate income statistics remains unclear. For example, aggregate income might be misrepresented in T1 PMF data if very high income earners delay filing; this group makes up a small share of all tax filers, but the sum of their incomes could still represent a non-trivial share of the total. This section investigates the extent to which income tax reassessments and delayed tax filing affect aggregate income statistics, centring on the following five types of income: employment, business self-employment, EI, OAS, and total.

Table 4 shows that total employment earnings among initial tax filers ranged from $468.3 billion in 1990 to $681.2 billion in 2010 (2010 constant dollars), based on the T1 PMF data. The corresponding values based on the T1 HPMF data are $468.6 billion and $681.7 billion, respectively, which indicates that the employment earnings of initial tax filers are being adjusted upward between the time that the T1 PMF and T1 HPMF datasets were compiled by the CRA. Across all years, the magnitude of this adjustment ranges from $0.1 billion to $0.9 billion. There are several possible explanations for this result. For instance, tax filers may forget to claim certain income from employment on their tax returns or intentionally under-state income to evade taxes, issues that become corrected as the CRA updates individual tax returns using information submitted by employers and performs random tax audits. Although the behavioural factors behind this result are unclear, the column in Table 4 on the percent of income in the T1 HPMF captured by the T1 PMF shows that under-reporting has little effect on the aggregate estimates; around 99.9% of total employment earnings in the T1 HPMF are also observed in the T1 PMF for initial tax filers.

Table 4
Aggregate income by income source, year and type of tax filer, selected years, 1990 to 2010 (combined)

| | Aggregate income | | | Percentage of income in T1 HPMF captured by T1 PMF | |
	Initial tax filers (T1 PMF)	Initial tax filers (T1 HPMF)	Delayed tax filers (T1 HPMF)	For initial tax filers	For all tax filers
	billions of 2010 constant dollars			percent	
Employment earnings					
1990	468.3	468.6	20.5	99.9	95.7
1995	465.5	465.6	14.0	100.0	97.1
2000	563.0	563.5	25.0	99.9	95.7
2005	621.5	622.4	25.6	99.8	95.9
2010	681.2	681.7	22.2	99.9	96.8
Business self-employment					
1990	11.6	11.9	1.0	97.1	89.9
1995	12.9	13.3	1.0	96.8	90.1
2000	19.0	19.6	2.0	97.0	87.9
2005	22.5	23.1	2.4	97.3	88.2
2010	23.3	23.5	1.8	99.0	91.9
Employment Insurance					
1990	18.3	18.3	1.0	100.0	95.0
1995	16.5	16.4	0.6	100.0	96.3
2000	11.2	11.2	0.6	100.0	95.2
2005	13.4	13.4	0.6	100.0	95.5
2010	18.8	18.8	0.7	100.0	96.2
Old Age Security					
1990	15.1	15.1	0.4	100.0	97.7
1995	18.6	18.6	0.3	100.0	98.6
2000	21.4	21.4	0.2	100.0	99.1
2005	23.6	23.5	0.2	100.0	99.1
2010	26.9	26.9	0.2	100.0	99.2
Total					
1990	674.2	676.3	27.9	99.7	95.7
1995	706.1	707.6	20.5	99.8	97.0
2000	854.2	857.3	34.3	99.6	95.8
2005	932.3	934.8	35.6	99.7	96.1
2010	1,051.4	1,052.7	30.3	99.9	97.1

Note: Results are based on cross-sectional data. An initial tax filer is an individual who appears in the T1 PMF and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File (T1 HPMF) and T1 Personal Master File (T1 PMF).

Table 4 also shows that income tax reassessments among initial tax filers have no effect on EI and OAS statistics. For example, in 1990, the estimated sum of all EI payments to initial tax filers is $18.3 billion in both datasets. More precisely, 100.0% of the aggregate values of EI and OAS incomes of initial tax filers in the T1 HPMF is always observed in the T1 PMF (rounded to one decimal place). In contrast, income under-reporting appears to be somewhat prevalent in the case of business self-employment income: the T1 PMF only captures 96.8% to 99.0% of that observed in the T1 HPMF for the years considered. This result is likely explained, at least in part, by the fact that business self-employment income is generally much more difficult to measure for tax filers, as doing so involves calculating profits from income and expense records, whereas employment income is reported by employers on standard tax forms. This raises the question of whether individuals with business self-employment income are more likely to file income tax adjustments with the CRA than those without such income. This finding is consistent with a growing literature in behavioural public finance, which shows that tax evasion is relatively prevalent for self-employment income because it is more difficult for tax authorities to observe (Clotfelter 1983; Slemrod 1985, 2007; Feinstein 1991; Andreoni, Erard, and Feldstein 1998; Schuetze 2002; Feldman and Slemrod 2007; Hurst, Li, and Pugsley 2014).

The effects of delayed tax filing on the income statistics are presented in the third and fifth columns of data in Table 4. For example, in 1990, aggregate employment earnings among delayed tax filers was $20.5 billion (2010 constant dollars)—the T1 PMF captures 95.7% of total employment earnings observed in the T1 HPMF. Patterns are similar for both EI and OAS income. However, consistent with the previous findings, the T1 PMF only captures 87.9% to 91.9% of all business self-employment income in the T1 HPMF, which suggests that income tax reassessments and delayed tax filing are both somewhat common among recipients of business self-employment income. The causes of this behaviour are outside the scope of this study, and represent an interesting avenue for future research. Combining the results of the first three columns of data permits an assessment of how much of the difference between the T1 PMF and T1 HPMF is due to reassessments versus delayed tax filing. For example, on the basis of the employment earnings statistics for the 2010 tax year, income from delayed tax filers accounted for $22.2 billion (97.8%) of the total difference between the T1 PMF and T1 HPMF of $22.7 billion, whereas the remaining $0.5 billion (2.2%) is the result of reassessments.[5] Similarly, the effects of delayed tax filing versus reassessments for the other sources of income are 90.0% versus 10.0% for business self-employment, 100.0% versus 0.0% for EI and OAS, and 95.9% versus 4.1% for total income, respectively. This assessment continues to find that the effect of reassessments is largest for business self-employment income compared with these other sources.

Last, Table 5 considers how the results vary by province and territory. Because delayed tax filing is relatively prevalent in Ontario, Alberta, British Columbia, and the territories, aggregate income statistics might be disproportionately affected in these regions. The findings suggest that this is not the case. Although some variation exists, the T1 PMF data capture at least 98.4% of the incomes from each source of initial tax filers, across all relevant years. In addition, the finding that business self-employment income is systematically under-represented by the T1 PMF data by a few percentage points is consistent across regions, although this result is most pronounced among non-residents.

---

5. The total difference in employment earnings between the T1 PMF and T1 HPMF ($22.7 billion) is equal to the sum of employment earnings from initial and delayed tax filers in the T1 HPMF ($681.7 billion and $22.2 billion, respectively) minus employment earnings from initial tax filers in the T1 PMF ($681.2 billion).

Table 5-1
Aggregate income by place of residence, income source and type of tax filer, 2010 — Part 1

Place of residence and income source	Aggregate income			Percentage of income in T1 HPMF captured by T1 PMF	
	Initial tax filers (T1 PMF)	Initial tax filers (T1 HPMF)	Delayed tax filers (T1 HPMF)	For initial tax filers	For all tax filers
	millions of 2010 constant dollars			percent	
Newfoundland and Labrador					
Employment earnings	9,397.0	9,402.9	260.7	99.9	97.2
Business self-employment	138.3	140.3	12.6	98.6	90.4
Employment Insurance	932.1	933.1	23.7	99.9	97.4
Old Age Security	470.0	470.0	2.6	100.0	99.5
Total	14,557.1	14,579.3	364.1	99.8	97.4
Prince Edward Island					
Employment earnings	2,246.4	2,247.5	61.0	100.0	97.3
Business self-employment	63.9	64.8	4.7	98.6	91.9
Employment Insurance	217.7	217.8	7.6	100.0	96.6
Old Age Security	130.5	130.5	0.8	100.0	99.4
Total	3,693.1	3,698.8	89.8	99.8	97.5
Nova Scotia					
Employment earnings	15,890.8	15,897.6	561.6	100.0	96.5
Business self-employment	378.8	385.0	39.6	98.4	89.2
Employment Insurance	771.5	772.1	31.6	99.9	96.0
Old Age Security	882.0	881.9	5.8	100.0	99.4
Total	25,966.7	26,002.1	777.7	99.9	97.0
New Brunswick					
Employment earnings	13,040.4	13,047.1	303.0	99.9	97.7
Business self-employment	306.8	311.7	23.5	98.4	91.5
Employment Insurance	836.4	836.3	19.4	100.0	97.7
Old Age Security	707.2	707.1	3.4	100.0	99.5
Total	20,291.5	20,316.0	417.5	99.9	97.9
Quebec					
Employment earnings	143,085.1	143,136.7	2,313.9	100.0	98.4
Business self-employment	4,652.5	4,683.9	207.8	99.3	95.1
Employment Insurance	5,443.9	5,443.6	105.4	100.0	98.1
Old Age Security	7,107.0	7,106.3	37.2	100.0	99.5
Total	227,435.8	227,650.7	3,439.6	99.9	98.4
Ontario					
Employment earnings	265,577.6	265,786.8	10,057.9	99.9	96.3
Business self-employment	9,599.2	9,697.6	788.3	99.0	91.5
Employment Insurance	5,921.3	5,921.4	275.4	100.0	95.6
Old Age Security	9,893.4	9,892.1	90.4	100.0	99.1
Total	410,869.1	411,400.9	13,648.3	99.9	96.7
Manitoba					
Employment earnings	21,806.1	21,818.0	655.7	99.9	97.0
Business self-employment	787.4	790.7	50.8	99.6	93.6
Employment Insurance	476.1	476.0	19.3	100.0	96.1
Old Age Security	990.2	990.1	7.2	100.0	99.3
Total	33,093.6	33,134.9	874.1	99.9	97.3
Saskatchewan					
Employment earnings	20,582.3	20,594.9	452.0	99.9	97.8
Business self-employment	800.5	809.7	49.2	98.9	93.2
Employment Insurance	399.2	399.1	15.4	100.0	96.3
Old Age Security	901.8	901.7	5.3	100.0	99.4
Total	31,734.7	31,776.1	640.9	99.9	97.9

Note: Results are based on cross-sectional data. An initial tax filer is an individual who appears in the T1 PMF and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.

Sources: Statistics Canada, T1 Historical Personal Master File (T1 HPMF) and T1 Personal Master File (T1 PMF).

Table 5-2

Aggregate income by place of residence, income source and type of tax filer, 2010 — Part 2

Place of residence and income source	Aggregate income			Percentage of income in T1 HPMF captured by T1 PMF	
	Initial tax filers (T1 PMF)	Initial tax filers (T1 HPMF)	Delayed tax filers (T1 HPMF)	For initial tax filers	For all tax filers
				percent	
Alberta					
Employment earnings	101,151.4	101,229.2	3,881.1	99.9	96.2
Business self-employment	2,452.7	2,482.8	256.4	98.8	89.5
Employment Insurance	1,589.6	1,590.2	97.3	100.0	94.2
Old Age Security	2,192.0	2,191.7	18.2	100.0	99.2
Total	141,903.8	142,110.7	4,956.0	99.9	96.5
British Columbia					
Employment earnings	83,800.3	83,865.0	3,347.4	99.9	96.1
Business self-employment	3,962.8	4,015.0	343.6	98.7	90.9
Employment Insurance	2,077.6	2,077.8	130.5	100.0	94.1
Old Age Security	3,555.3	3,554.9	34.8	100.0	99.0
Total	135,693.2	135,882.5	4,675.8	99.9	96.5
Territories					
Employment earnings	2,941.6	2,943.5	161.3	99.9	94.7
Business self-employment	73.5	74.4	8.5	98.8	88.6
Employment Insurance	73.1	73.1	4.2	100.0	94.6
Old Age Security	35.3	35.3	0.7	100.0	98.0
Total	3,619.8	3,624.7	196.2	99.9	94.7
Outside Canada					
Employment earnings	1,712.5	1,712.5	132.8	100.0	92.8
Business self-employment	76.8	77.0	16.0	99.7	82.6
Employment Insurance	13.9	14.0	0.3	99.2	97.4
Old Age Security	55.6	55.7	0.3	99.7	99.2
Total	2,579.0	2,571.6	252.4	100.3	91.3

Note: Results are based on cross-sectional data. Results for other years in the repeated cross-sectional are comparable. An initial tax filer is an individual who appears in the T1 PMF and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File (T1 HPMF) and T1 Personal Master File (T1 PMF).

It is important to note that the personal income tax data received by Statistics Canada from the CRA typically exclude 20 to 50 records for the highest income tax filers. This likely has little effect on most of the analysis in this article pertaining to the characteristics of tax filers, income distributions, or income thresholds. However, because these income values are very large, it is possible for certain aggregate statistics to be affected. To some extent, the issue is mitigated by the T1 PMF and T1 HPMF datasets being both affected, making it is less of a concern for the comparative analysis. In addition, the aggregate statistics presented here are very comparable to the final statistics produced directly by the CRA based on approximately 100% of all tax returns, including reassessments. For example, for the 2010 tax year, which includes returns filed up to the cut-off date of June 30, 2012, total income amounted to $1,070.3 billion (CRA 2012), whereas it amounted to $1,083.0 billion for all tax filers in the T1 HPMF based on a cut-off date of December 18, 2012, shown in Appendix Table 1.

4.3   Other income under-reporting

Given that income tax reassessments and delayed tax filing are relatively prevalent for business self-employment income, a related question is whether this also applies for other types of income. This section briefly considers the effects of such behaviour on net income (i.e., income minus expenses) from business self-employment, commissions, farming, fishing, professional, and rental.

Table 6 shows the probabilities of having income from each of these sources, by type of tax filer. For instance, delayed tax filers are nearly twice as likely to have business self-employment income or commission income (13.4% and 1.3%, respectively) than initial tax filers (7.2% and 0.7%, respectively). However, the opposite is true of farming income. The shares of initial and delayed tax filers with fishing or professional income are approximately equal.

Table 6

Possible sources of income and of average conditional income for initial and delayed tax filers, selected years, 1990 to 2010 (combined)

	Initial tax filers	Delayed tax filers
	percent	
Income sources		
Business self-employment	7.2	13.4
Commission	0.7	1.3
Farming	1.9	1.0
Fishing	0.2	0.2
Professional	1.4	1.7
Rental	5.4	4.5
	2010 constant dollars	
Average conditional income sources		
Business self-employment	11,500	12,450
Commission	18,150	17,200
Farming	4,700	-2,150
Fishing	20,400	17,600
Professional	62,500	35,200
Rental	1,900	-1,150

Notes: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010. The average income statistics for each group are also shown, pertain to individuals who had non-zero income reported on their tax returns. An initial tax filer is an individual who appears in the T1 Personal Master File and who submits an income tax return to the Canada Revenue Agency before the cut-off date. A delayed tax filer does not submit an income tax return before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

The results also show that delayed tax filers earn less, on average, than initial tax filers across all of the sources of income except business self-employment. For example, the averages of professional income for initial and delayed tax filers are approximately $62,500 and $35,200 (2010 constant dollars), respectively, conditional on individuals who have professional income or losses to report. Taken together, the characteristics of business self-employment income recipients do not systematically extend to tax filers with these other types of income.

The extent to which income tax reassessments and delayed tax filing affect estimates of the aggregate statistics of these other types of income is presented in Table 7. In the first column of data, the percentage of delayed tax filers is shown among all tax filers with income from each source. For example, 8.1% of individuals with commission income delayed filing over the years considered. Overall, tax filers with business self-employment or commission income appear slightly more likely than the full sample to delay filing; the opposite is true of farming and rental income recipients.

Table 7

Effects of reassessments and delayed tax filing on aggregate statistics, by income source, selected years, 1990 to 2010 (combined)

Income source	Probability of delayed tax filing conditional on having income	Percentage of income in T1 HPMF captured by T1 PMF	
		For initial tax filers	For all tax filers
	percent		
Business self-employment	7.6	97.6	89.6
Commission	8.1	99.4	91.8
Farming	2.3	99.2	100.3
Fishing	4.2	99.5	95.9
Professional	5.1	99.8	96.8
Rental	3.5	97.1	99.2

Notes: Results are based on pooled cross-sectional data for 1990, 1995, 2000, 2005, and 2010, and pertain to tax filers having non-zero income from each source. An initial tax filer is an individual who appears in the T1 PMF and who submits an income tax return to the Canada Revenue Agency before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File (T1 HPMF) and T1 Personal Master File (T1 PMF).

The last two columns show the percentage of other income observed in the T1 HPMF for initial and delayed tax filers captured by the T1 PMF. The results indicate, first, that income tax reassessments have little effect on income from commissions, farming, fishing, and professional; these estimates all exceed 99%. However, the business self-employment and rental incomes of initial tax filers are under-represented in the T1 PMF, the estimates being only 97.6% and 97.1%, respectively.

In contrast, the effect of delayed tax filing are most prevalent for the cases of business self-employment income and commission income, which is likely explained by the fact that delayed tax filing is the most common for individuals with income from these sources. The shares of T1 HPMF income captured by the T1 PMF are only 89.6% and 91.8%, respectively. As mentioned, explaining the factors that underpin these results is outside the scope of this study, but would constitute an interesting topic of future research.

An important caveat of these findings is that the analysis only uses personal income tax data from T1 tax records. Whether similar reassessment and delayed tax filing patterns would be observed on the basis of an analysis of T2 tax records is unclear. In general, this behaviour is expected to be different for businesses compared with individuals.

4.4 Estimating income inequality

In a recent study, "Who are Canada's top 1 percent?", published in *Income Inequality: The Canadian Story* by the Institute for Research on Public Policy, Thomas Lemieux and W. Craig Riddell (2015) analyze the evolution of top incomes and the top 1 percent of income earners between 1981 and 2011. They show that top incomes rose since the 1980s, largely as a result of increasing inequality in the financial, business services, and oil and gas sectors.

Although their study was primarily based on census master files, the authors also considered how using tax records from the LAD changed the results. They showed that "the cut-offs for the 95th and 99th percentiles in the two data sources are remarkably similar, with those from the LAD slightly higher than those from the census, but in most cases the gap is less than 5 percent" (Lemieux and Riddell 2015, p. 115–116).[6] However, cut-offs for the 99.9th percentile are larger in the LAD by as much as 25%, consistent with other evidence that high-income earners tend to underreport income (Bound and Krueger 1991).

---

6. The census and LAD estimates of the top income cut-offs are initially reported in Milligan (2013) and Veall (2012), respectively.

Because the LAD is constructed using the T1FF, which is derived from T1 PMF data for initial tax filers, this raises the question of whether, and to what extent, the top-income cut-off estimates from these data are affected by income tax reassessments and delayed tax filing. The finding from Subsection 4.1 that such behaviours are most prevalent among individuals at the bottom of the income distribution means that top-income cut-offs are biased upward by excluding these individuals from the calculations, and the magnitude of this bias would become the most pronounced for very high income cut-offs.

The degree to which this bias occurs and plays a role in explaining the discrepancy in the 99.9th percentile cut-off estimates from the administrative versus survey data is addressed in this section. Table 8 reports the top-income cut-offs across employment earnings, business self-employment income, and total income for the 95th, 99th, and 99.9th percentiles, based on the populations of initial and all tax filers. The estimates for initial tax filers use the T1 PMF, and the estimates for all tax filers use the T1 HPMF. Thus, differences between the two groups stem from both income tax reassessments and delayed tax filing.

Table 8

Top income percentile cut-offs for employment earnings, business self-employment income and total income of initial and all tax filers, selected years, 1990 to 2010

	1990	1995	2000	2005	2010
	2010 constant dollars				
Panel A – Employment earnings					
95th percentile					
Initial tax filers	79,200	77,850	83,700	87,250	92,450
All tax filers	79,150	77,600	83,600	87,150	92,400
99th percentile					
Initial tax filers	119,800	118,400	140,200	150,000	160,050
All tax filers	119,800	117,950	139,800	149,500	159,850
99.9th percentile					
Initial tax filers	300,300	315,100	483,550	524,050	480,100
All tax filers	300,000	310,550	476,650	517,550	477,550
Panel B – Business self-employment income					
95th percentile					
Initial tax filers	0	0	1,050	1,450	1,850
All tax filers	0	50	1,900	2,500	2,550
99th percentile					
Initial tax filers	24,350	23,850	28,250	28,800	26,750
All tax filers	25,200	24,600	29,700	30,450	27,800
99.9th percentile					
Initial tax filers	75,900	75,400	89,950	96,400	92,700
All tax filers	77,850	77,100	92,400	99,400	94,300
Panel C – Total income					
95th percentile					
Initial tax filers	91,200	87,950	97,700	101,000	107,850
All tax filers	91,200	87,700	97,600	100,900	107,650
99th percentile					
Initial tax filers	168,200	160,950	200,400	206,600	217,400
All tax filers	168,600	160,250	199,200	205,450	216,150
99.9th percentile					
Initial tax filers	521,100	498,950	771,450	795,900	771,150
All tax filers	528,600	498,000	765,700	788,900	763,300

Note: Results are based on cross-sectional data. An initial tax filer is an individual who appears in the T1 Personal Master File and who submits an income tax return to the Canada Revenue Agency before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

The analysis indicates, for example, that the 99.9th percentile of employment income in 2010 was $480,100 among initial tax filers, and $477,550 among all tax filers. Therefore, omission of delayed tax filers means this cut-off is overstated by approximately $2,550. In general, the direction of the bias is consistent with expectations: the top-income cut-offs for employment earnings and total income are slightly larger from the T1 PMF than from the T1 HPMF. The opposite is true for business self-employment income, although this is to be expected because delayed tax filers tend to earn more from this source than do initial tax filers, as shown in Table 2.

However, although income tax reassessments and delayed tax filing have the expected effect on estimates of top-income cut-offs, the magnitude of this bias is very small. For example, the 99.9th percentiles of total income in 2010 were $771,150 and $763,300 based on the T1 PMF and T1 HPMF data, respectively—a difference of only $7,850. Table 9 shows that this difference is the largest of any discrepancy observed between the two datasets, yet it only represents 1.03% of the value of this cut-off. Discrepancies across other income percentiles, sources of income, and years are typically much smaller. Therefore, estimates of top-income cut-offs derived from administrative data based on the approximately 100% sample of initial tax filers are reasonable approximations of the values that would have been obtained based on estimates from the population of all tax filers; income tax reassessments and delayed tax filing do not explain the discrepancies between the 99.9th percentile estimates documented by Lemieux and Riddell from the census and LAD data.

Table 9
Differences between the top-income cut-offs of initial and all taxfilers, by type of income and percentile, selected years, 1990 to 2010

Type of income and percentile	1990	1995	2000	2005	2010
	2010 constant dollars				
Panel A – Employment earnings					
95th percentile	-50	-250	-100	-100	-50
99th percentile	0	-450	-400	-500	-200
99.9th percentile	-300	-4,550	-6,900	-6,500	-2,550
Panel B – Business self-employment income					
95th percentile	0	50	850	1,050	700
99th percentile	850	750	1,450	1,650	1,050
99.9th percentile	1,950	1,700	2,450	3,000	1,600
Panel C – Total income					
95th percentile	0	-250	-100	-100	-200
99th percentile	400	-700	-1,200	-1,150	-1,250
99.9th percentile	7,500	-950	-5,750	-7,000	-7,850

Notes: Results are based on cross-sectional data and show differences in top income cut-offs in Table 8 based on whether the dataset for initial or for all tax filers is used in the calculations. An initial tax filer is an individual who appears in the T1 Personal Master File and who submits an income tax return to the Canada Revenue Agency before the cut-off date.
Sources: Statistics Canada, T1 Historical Personal Master File and T1 Personal Master File.

5   Conclusion

Amid an increasing reliance on administrative tax records in economic analysis for evidence-based policy decision-making and academic research (Chetty 2012; Einav and Levin 2014), the extent to which big tax datasets are confounded by income tax reassessments and delayed tax filing is an important, underexplored empirical issue. Using population datasets for initial and delayed Canadian tax filers, from 1990 to 2010, this study provides novel insight into this issue and discusses the resulting implications for economic analysis that uses big tax data.

The results of this study show that, on average, only around 3.5% to 4.8% of individuals do not submit their tax returns to the Canada Revenue Agency in time each year to be included in conventional datasets derived from population files of initial Canadian tax filers. This behaviour is the most prevalent among younger tax filers; residents of Ontario, Alberta, British Columbia, and the territories; non-residents; emigrants; and very low income earners. Consistent with expectations, individuals with final income tax balances due (or refunds) close to zero tend to be the most likely to delay filing, as the incentives to file promptly can be the weakest in this case.

The implications of income tax reassessments and delayed tax filing for economic analysis that uses big tax data are generally favourable: on balance, the effects of such behaviour are small. For most income sources—including employment, Employment Insurance, and Old Age Security—the distributions of income for initial and all tax filers closely overlap over normal income ranges, and aggregate statistics are nearly identical. The only noteworthy exception is business self-employment income, for which reassessments are a bit more common. Explaining the factors behind this result constitutes an interesting avenue for future research in behavioural public finance.

Last, this study assesses the consequences of estimating top-incomes using big tax data derived only from initial tax filers. This issue is particularly relevant for studies on income inequality in Canada, given that delayed tax filing is shown to be the most prevalent among individuals with very low incomes. In theory, this should have the effect of biasing top-income cut-off estimates upward, especially for very high income thresholds. However, the results of this analysis are again favourable, and indicate that these cut-off estimates are not significantly skewed in practice by omitting delayed tax filers from the calculations.

6 Appendix

Appendix Table 1
Latest filing and assessment dates in the PMF and HPMF, by tax year

| | T1 PMF | | T1 HPMF | |
	Filing	Assessment	Filing	Assessment
1990	November 28, 1991	December 20, 1991	January 24, 1997	December 28, 2005
1995	December 30, 1996	January 13, 1997	January 23, 2002	December 28, 2005
2000	December 21, 2001	January 10, 2002	December 12, 2005	December 28, 2005
2005	December 7, 2006	December 18, 2006	December 19, 2008	January 6, 2009
2006	December 13, 2007	December 28, 2007	August 18, 2009	August 27, 2009
2007	December 10, 2008	December 22, 2008	January 28, 2011	February 4, 2011
2008	December 22, 2009	January 6, 2010	December 22, 2011	January 6, 2012
2009	December 22, 2010	January 6, 2011	December 20, 2011	January 6, 2012
2010	December 22, 2011	January 6, 2012	December 18, 2012	January 7, 2013

Notes: The latest filing and assessment dates observed in each dataset, for all of the relevant tax years analyzed in this study, are shown here.
Sources: Statistics Canada, T1 Historical Personal Master File (T1 HPMF) and T1 Personal Master File (T1 PMF).

References

Andreoni, J., B. Erard, and J. Feldstein. 1998. "Tax compliance." *Journal of Economic Literature* 36 (2): 818–860.

Bound, J., and A.B. Krueger. 1991. "The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?" *Journal of Labor Economics* 9 (1): 1–24.

Chetty, R. 2012. *Time Trends in the Use of Administrative Data for Empirical Research*. Presentation slides. Available at: http://rajchetty.com/chettyfiles/admin_data_trends.pdf.

Citro, C.F. 2014. "From multiple modes for surveys to multiple data sources for estimates." *Survey Methodology* 40 (2): 137–161. Statistics Canada Catalogue no. 12-001-X.

Clotfelter, C. 1983. "Tax evasion and tax rates: An analysis of individual returns." *Review of Economics and Statistics* 65 (3): 363–373.

CRA (Canada Revenue Agency). 2012. "Final table 1: General statement by province and territory of taxation." In *Final Income Statistics 2012 (2010 Tax Year)*. Last updated June 21, 2013. Available at: http://www.canada.ca/en/revenue-agency/programs/about-canada-revenue-agency-cra/income-statistics-gst-hst-statistics/t1-final-statistics-2012-edition-2010-tax-year/final-table-1-general-statement-province-territory-taxation-2010-tax-year.html (accessed August 24, 2017).

Duncan, J.W., and W.C. Shelton. 1978. *Revolution in United States Government Statistics 1926–1976*. Office of Federal Statistical Policy and Standards, U.S. Department of Commerce. Washington, D.C.: Government Printing Office.

Einav, L., and J.D. Levin. 2013. *The Data Revolution and Economic Analysis*. NBER Working Paper Series, no. 19035. Cambridge, Massachusetts: National Bureau of Economic Research.

Einav, L., and J.D. Levin. 2014. "Economics in the age of big data." *Science* 346 (6210): 715–721.

Feinstein, J. 1991. "An econometric analysis of income tax evasion and its detection." *RAND Journal of Economics* 22 (1): 14–35.

Feldman, N.E. 2010. "Mental accounting effects of income tax shifting." *Review of Economics and Statistics* 92 (1): 70–86.

Feldman, N.E., and J. Slemrod. 2007. "Estimating tax noncompliance with evidence from unaudited tax returns." *The Economic Journal* 117 (518): 327–352.

Finnie, R., D. Gray, and Y. Zhang. 2016. "A longitudinal analysis of GIS entries and exits." *Canadian Public Policy* 42 (3): 287–307.

Harris-Kojetin, B. 2012. "Federal household surveys." In *Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey, Second Edition*, ed. M.J. Anderson, C.F. Citro, and J.J. Salvo, p. 226–234. Washington, D.C.: CQ Press.

Heffetz, O., and D.B. Reeves. 2016. *Difficulty to reach respondents and nonresponse bias: Evidence from large government surveys*. NBER Working Paper Series, no. 22333. Cambridge, Massachusetts: National Bureau of Economic Research.

Hurst, E., G. Li, and B. Pugsley. 2014. "Are household surveys like tax forms? Evidence from income underreporting of the self-employed." *The Review of Economics and Statistics* 96 (1): 19–33.

Jarmin, R.S., and A.B. O'Hara. 2016. "Big data and the transformation of public policy analysis." *Journal of Policy Analysis and Management* 35 (3): 715–721.

Lane, J. 2016. "Big data for public policy: The quadruple helix." *Journal of Policy Analysis and Management* 35 (3): 708–715.

LaRochelle-Côté, S., J. Myles, and G. Picot. 2012. "Income replacement rates among Canadian seniors: The effect of widowhood and divorce." *Canadian Public Policy* 38 (4): 471–495.

Lemieux, T., and W.C. Riddell. 2015. "Who are Canada's top 1 percent?" In *Income Inequality: The Canadian Story*, ed. D.A. Green, W.C. Riddell, and F. St-Hilaire, Volume 5, p. 103–155. Montreal: Institute for Research on Public Policy.

Meyer, B.D., W.K.C. Mok, and J.X. Sullivan. 2015. *Household surveys in crisis*. NBER Working Paper Series, no. 21399. Cambridge, Massachusetts: National Bureau of Economic Research.

Milligan, K. 2013. *Income inequality and income taxation in Canada: Trends in the census 1980–2005.* The School of Public Policy Research Paper Series, Volume 6, Issue 24. Calgary: University of Calgary.

Rees-Jones, A. 2014. *Loss Aversion Motivates Tax Sheltering: Evidence From U.S. Tax Returns*. Philadelphia, Pennsylvania: The Wharton School, University of Pennsylvania. Unpublished manuscript.

Schuetze, H.J. 2002. "Profiles of tax non-compliance among the self-employed in Canada: 1969 to 1992*." Canadian Public Policy* 28 (2): 219–238.

Slemrod, J. 1985. "An empirical test for tax evasion." *The Review of Economics and Statistics* 67 (2): 232–238.

Slemrod, J. 2007. "Cheating ourselves: The economics of tax evasion." *The Journal of Economic Perspectives* 21 (1): 25–48.

Varian, H.R. 2014. "Big data: New tricks for econometrics." *The Journal of Economic Perspectives* 28 (2): 3–27.

Veall, M.R. 2012. "Top income shares in Canada: Recent trends and policy implications." *The Canadian Journal of Economics* 45 (4): 1247–1272.