2. Data and definitions

Key information on data used

- The analysis for this report relies on data extracted from Australian tax returns for the period 1990–91 to 2016–17, which represents 10 percent of individuals who had a tax file number at some point during this period.
- An important advantage of the dataset is that it captures the same individuals over an extended period, allowing for the comparison of their incomes before and after adverse events such as earnings shocks.
- For this study, we focus on individuals aged 25-54 who report positive labour income and whose income is observable for at least three consecutive years.
- We define earnings shock as an event in which earnings fall by at least 40 percent accompanied with a comparable drop in total income

2.1

Sample creation

The analysis for this report relies on data extracted from Australian tax returns for the period 1990–91 to 2016–17.[4] The dataset is known as ALife, the ATO Longitudinal Information Files. ALife consists of a random sample of 10 percent of individuals in the ATO client register. The client register is constructed from tax returns lodged since 1980, as well as other means by which the ATO becomes aware of the existence of an individual, such as an employer or Centrelink lodging a payment summary for that individual. Most individuals are longitudinally linked via their tax filer number, a unique individual identifier.

The dataset captures information from lodged tax returns for all years. Information on earnings and government benefits is available for non-lodgers from 2002 onwards.[5] Once a tax filer is selected to be in the sample, we can observe the information from their annual tax returns from the beginning of the sample period or the year in which they first start filing (whichever occurs later) until the last year of data collection or the year in which they stop filing (e.g., due to death or emigration).

The ALife dataset presents several advantages for studying issues related to the entry into or exit from poverty. First, the data permit the exploration of income across several components. For this report we focus on labour earnings, which we define broadly to include self-employment and business income. Second, given the liability associated with failing to report and/or misreporting information on one's income tax return, we assume that information provided on the tax return is reasonably accurate.[6] Third, we are able to track annual earnings for a large group (10 percent of tax filers) of individuals residing in Australia. The depth and extent of the data coverage permits us to undertake analyses that focus on different parts of the income distribution, age groups and other dimensions such as geographic location.

There are, however, some disadvantages in using ALife. Compared to surveys, we observe a limited amount of demographic information. We can observe gender, age and residential location. We can derive some information about marital status (in more recent years) and having children (based on child benefits received and self-reports in the

4 Hereafter, we refer to tax years by the year in which it ends. For example, 2017 refers to the 2016–17 tax year.

5 Polidano et al. (2020) show that combining lodgers' with non-lodgers' data results in a good representation of the Australian resident population aged 20 and older.

6 In surveys, earnings may be misreported when the payslip is not available to the interviewer, and respondents make mistakes in recollecting annual earnings. Researchers have highlighted that non-response rates for earnings in surveys is higher at the bottom and the top of the earnings distribution (Bollinger et al., 2019) and that there is often an under-reporting of welfare benefits in survey data (Meyer et al., 2015).

tax returns). While we observe annual labour earnings, we cannot observe the number of hours or days worked in that year.

As this report analyses earnings shocks, we have refined the dataset to focus on those individuals in the sample whose tax returns are observed when they are between the ages of 23 and 57. We identify earnings shocks for those aged between 25 and 54. For those aged 25, we compare earnings at age 25 with earnings reported in the two prior years. Those aged 54 are followed for the three subsequent years to capture information on the recovery from an earnings shock experienced at age 54 or earlier. The age range was selected to capture the core years one would be expected to work. Individuals who are under 23 are likely to be engaged in a training program and/or post-high school education. Individuals who are over 57 may be exiting the workforce by retiring completely or by reducing hours. The year-to-year variation in earnings for those under 23 or over 57 are expected to be noisy, which could lead to an overcounting of earnings shocks as one transitions from training to work and from work to retirement. We note that there are a range of issues that could be studied for those under 23 and those over 57, but we leave the study of these other issues to a future report.

In Table 2.1 we report the number of tax filers and/or observations excluded from the dataset used for this report based on a set of rules. We exclude 779,696 individuals from the sample because their tax information is captured only for ages that are outside of the age range used for this report. This leaves a total of 1,769,008 tax filers who could be studied. As explained in more detail below, we identify an individual as experiencing an earnings shock based on the earnings received in the previous two years. Thus, this requires that the individual has reported earnings for three consecutive years. Moreover, we include a requirement that reported earnings are at least 25 percent of the annual earnings of a full-year full-time worker paid the contemporary adult federal minimum wage (approximately $8,900 in 2017).[7] We further exclude individuals as follows.

1. Individuals who never report earnings that exceed the minimum threshold. This results in the exclusion of 176,813 individuals.
2. Individuals for whom we do not observe at least three consecutive years of tax information. This results in the exclusion of 76,751 individuals.
3. Individuals for whom we cannot calculate whether they have experienced an earnings shock because we can never observe pre-shock year earnings that are above the threshold. This results in the exclusion of 102,009 individuals.

Our final sample consists of 1,413,435 individuals, of which 53 percent are males. Of the male tax filers, 82 percent of the possible individuals that could be studied remain in the sample. Of the female tax filers, 78 percent of the possible individuals that could be studied remain in the sample. The primary reason for a greater exclusion of female tax filers is the rule that excludes tax filers whose earnings never exceed the minimum threshold of approximately $8,000.

The resulting dataset captures individuals born between 1938 and 1991. Those born in 1938 would be aged 53 in 1991, the first year of our data. And those born in 1991 would be aged 26 in 2017, the last year of our data. Appendix A provides further detail on the implications of the sample restrictions we implemented for the purposes of this study.

7 As the focus of the study is earnings shocks, we set a minimum earnings threshold to ensure that tax filers in the sample have a significant attachment to the labour market. Persons who earn less than one quarter of the annual full-time minimum wage for successive years are likely reliant on welfare benefits or other family members and, although a group that deserves attention, are out of the scope of this study.

Table 2.1. Development of the working dataset. ALife data, 1991—2017

	Number of persons		
Total number of persons in the dataset	2,548,704		
Persons with gender/age not reported	191,735		
Persons that are observed before the age of 25 or after 54 and not in between	587,961		
Starting sample for analysis	1,769,008		
	Males (1)	Females (2)	Total (3)
Starting sample for analysis	919,891	849,117	1,769,008
Persons whose earnings never exceeded the minimum threshold for measuring an earnings shock (~$8,900 in 2017)	72,683	104,130	176,813
Persons who are never observed for at least three consecutive years	44,144	32,607	76,751
Persons whose earnings for the two consecutive years used to identify an earnings shock are always less than the minimum threshold (~$8,900 in 2017)	50,638	51,371	102,009
Number of persons studied	752,426	661,009	1,413,435

Notes: See Polidano et al. (2020) for more information on ALife data. For the definition of earnings shock see chapter 2, section 2.

Earnings distributions for lodgers and non-lodgers

ALife contains data from all tax returns lodged by tax filers and, from 2002, data for non-lodgers. Correspondingly, earnings and incomes of individuals who lodge a tax return are observed in every year, while for individuals who do not lodge a return, information is only available from 2002 onwards, and even then the income data are restricted to earnings and taxable government benefits (thus excluding business and investment income). Prior to 2002, whenever a tax return was not lodged[8] our analysis assumes that earnings and income of the individual were zero in that year.[9] While the pre-2002 data give us no option other than to take this approach, is this a reasonable assumption? A second related question in respect of non-lodgers is whether we should be concerned about potential bias when studying individuals with low earnings if these individuals do not lodge tax returns.

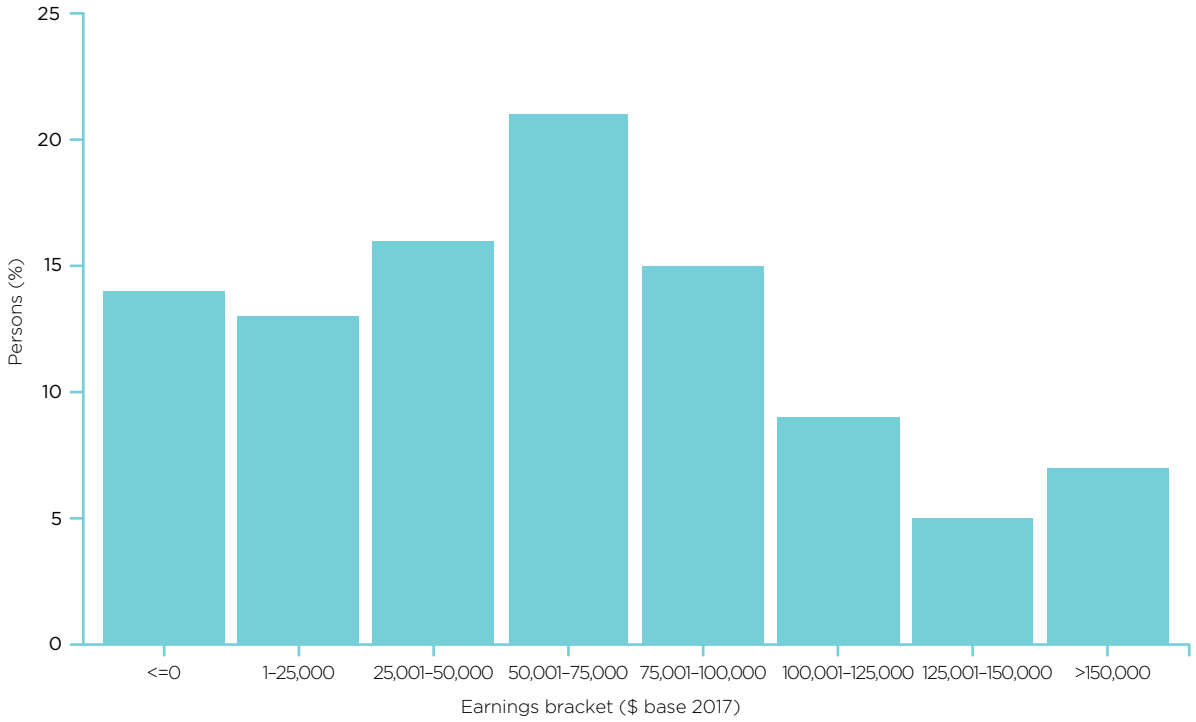
To explore these questions, we focus on the information for lodgers and non-lodgers in the most recent year, 2017. For this year we can capture earnings for both groups. Figure 2.1 depicts the distribution of reported earnings for males. The earnings are broken into eight groups. The first group captures individuals whose reported earnings are zero or negative. One can have negative earnings if self-employment or business income is negative—that is, the business reports a loss.[10] The second group captures individuals with earnings that range between $1 and $25,000. Less than 15 percent of the sample falls into each of these lower threshold groups. Most males (Figure 2.1) report earnings ranging between $25,000 and $125,000.

8 After 2002, information on earnings and income is missing if the individual does not work and does not receive taxable government benefits and/or if the individual is not residing in Australia. Although this may raise concerns about ALife's representativeness, Polidano et al. (2020) show that lodgers and non-lodgers approximate the Australian resident population aged 20 and older.

9 We impute zero earnings and income only if we observe earnings and income of those individuals in successive years.

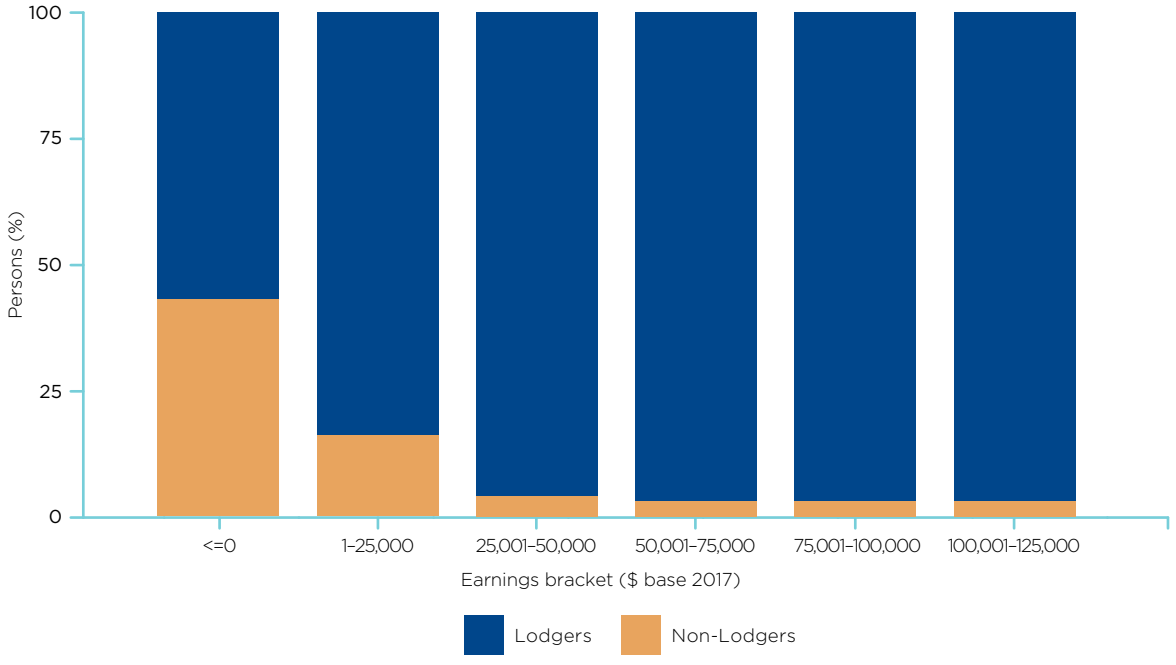
10 Given the uptake of "gig economy" jobs, such as Uber driving, which are treated as self-employment, we wanted to include this type of income in our measure of earnings.

Figure 2.1. Share of persons by earnings bracket, 2017—Males



Notes: Captures persons who file a tax return (lodgers) or for who we have non-lodger information. For definition of lodger and non-lodger see chapter 2, section2. All dollars used in this report are converted to nominal dollars, with 2017 as the base year.

Figure 2.2. Proportions of lodgers and non-lodgers, by earnings bracket, 2017—Males



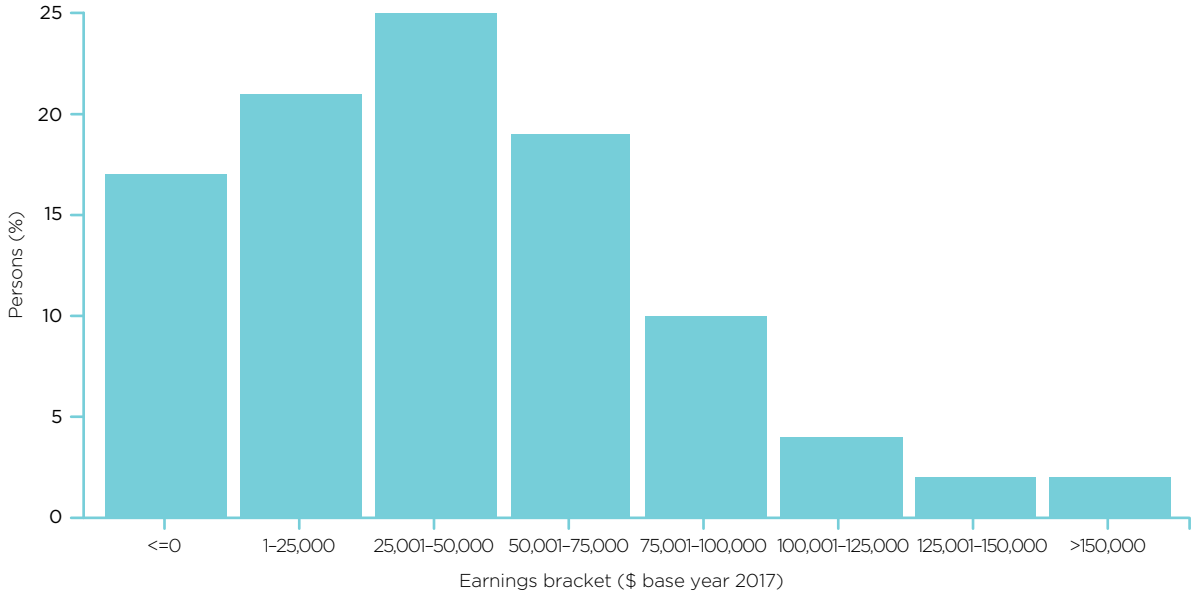
Notes: For definition of lodger and non-lodger see chapter 2, section2.

Figure 2.2 depicts the proportion of males within each earnings group based on whether the individual is a lodger or non-lodger. Most of the non-lodgers fall into the group of individuals with zero or negative earnings. The second largest group of non-lodgers has earnings that are less than $25,000. These figures illustrate that a challenge for the period before 2002 (when no non-lodger data are available) is that an individual could be classified as experiencing a negative earnings shock if they did not lodge a return and we assume zero earnings for the year of the shock. This will mean that for the period before 2002 we likely are overstating the share of tax filers with an earnings shock. As explored further in Appendix C however, we believe that any overstatements will be slight given the methods used to identify a shock and the fact that most of the non-lodgers for whom we can observe earnings report zero earnings.

Figures 2.3 and 2.4 depict the earnings distribution and the proportion of non-lodgers by earnings group for females. A higher share of females report earnings that are less than $25,000, and the majority earn between $25,000 and $100,000. Like males, most non-lodgers have earnings that are equal to zero or less. Compared to males, however, a lower proportion of female non-lodgers report earnings that are between $1 and $25,000.

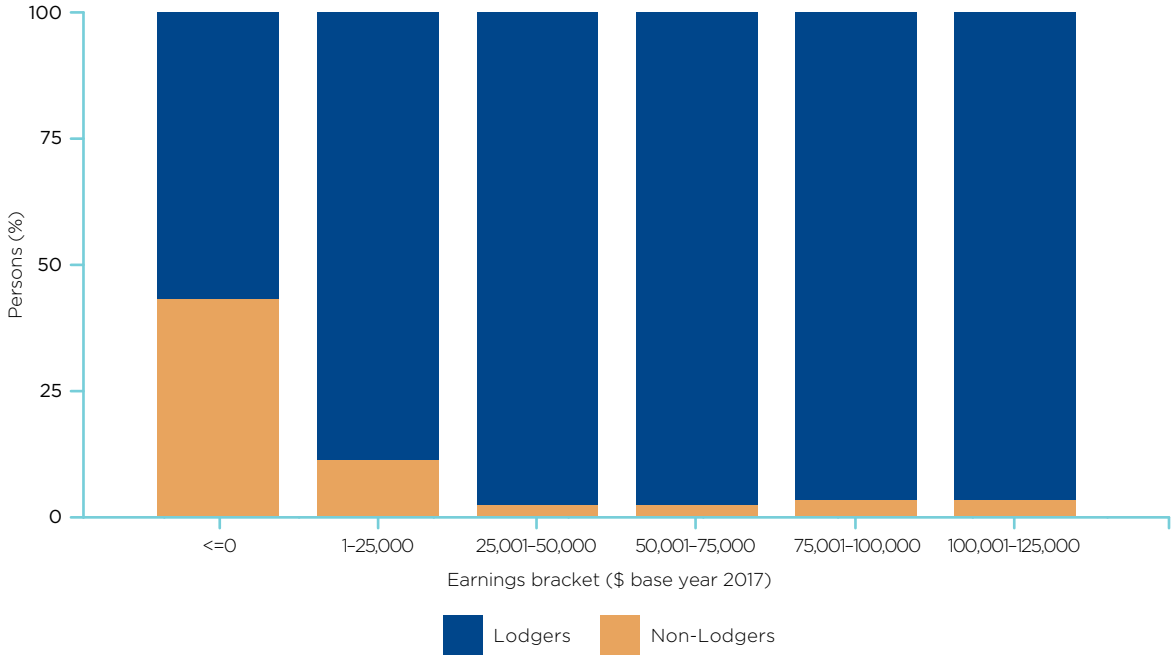
In sum, not having information for non-lodgers leads to a loss of information on earnings for a tax filer for the period before 2002. Our assumption that these non-lodgers have earned zero earnings in the year for which a tax return was not lodged, however, will understate earnings for a small proportion of the tax filers that are studied.

Figure 2.3. Share of persons by earnings bracket, 2017—Females



Notes: Captures persons who file a tax return (lodgers) or for who we have non-lodger information. For definition of lodger and non-lodger see chapter 2, section 2. All dollars used in this report are converted to real dollars, with 2017 as the base year.

Figure 2.4. Proportions of lodgers and non-lodgers, by earnings bracket, 2017—Females



Notes: For definition of lodger and non-lodger see chapter 2, section 2.

2.2

Defining earnings shocks

What constitutes a relevant drop in earnings to constitute a negative shock? Our definition relies on three key features: (a) the definition of earnings; (b) the period used to identify a shock; and (c) the minimum percentage loss in earnings used to identify a shock, which includes a consideration of the change in overall income (excluding government benefits) relative to the change in earnings. We address each of these features separately.

Definition of an earnings shock

To be classified as having experienced an earnings shock we compare current earnings against the minimum earnings in the previous two years. If current earnings have fallen by more than 40 percent, then an individual is identified as experiencing a shock. We define earnings as: the sum of wages, gross business income and gross self-employment income.

Defining earnings

Critical to this report is to explore significant changes in one's finances, especially for those individuals who are at the lower end of the income distribution and/or those who are at risk of falling into the lower end of the income distribution if they are unable to recover from the decline in their finances.

The Australian tax return captures the following income components:

- salaries/wages;
- additional payments from one's employers (lump sum payments,[11] termination payments,[12] allowances,[13] tips and gratuities, consultation fees);
- business income including personal services income;
- Australian government pensions and allowances such as Newstart Allowance, Parenting Payment Single and the Disability Support Pension;

11 Lump sum payments for unused annual leave and unused long service leave.
12 Lump sum payments given to employees when they resign, retire or paid to an estate in the event of an employee's death.
13 Allowances are expenses reimbursed by the employer.

- Australian annuities and superannuation income streams;
- interest, dividends and other capital gains; and
- foreign income (and other revenue sources).

We concentrate on the core earnings that are associated with working, either as an employee or as a business owner. The measures used to capture earnings are salaries/wages and earnings from self-employment and business income. We focus on this measure of earnings on the assumption that most individuals will cover necessities from sources of income tied to wages or self-employment income.

Period used to measure a negative earnings shock

The richness of the data permits us to capture year to year variations in earnings. One might, however, experience a temporary increase in earnings in one year that is representative of an anomaly rather than a trend for that individual. For example, one might work overtime due to a crisis at work that would temporarily increase one's earnings during the period of the crisis. And if this were the case, an individual who reports the same level of earnings for the first and third year over a 3-year period, but experiences an increase in earnings in the middle year, could be incorrectly classified as experiencing a negative shock when, in fact, the individual has simply returned to a level of earnings approximately the same as they were receiving previously. For this reason, we adopt a 3-year period for capturing a negative shock to earnings. To assess whether earnings for the year under study have declined sufficiently to be classified as a negative shock, we compare the earnings reported for the year under study relative to the minimum earnings reported in each of the two previous years.[14]

Minimum percentage loss in earnings and income to be classified as experiencing a negative earnings shock

An individual is classified as having experienced negative earnings shock each year if:

(a) their earnings in each of the previous two years exceeded 25 percent of the annualised earnings if one worked full-time and earned the minimum wage (approximately $8,900 in 2017);

(b) their earnings that year are less than 40 percent of the minimum earnings received in each of the previous two years; and

(c) their total income, net of taxable government benefits, has also declined by 40 percent.[15]

We include this latter requirement to ensure that we exclude individuals with other sources of income that would exceed earnings in a way that the non-earnings-related income offsets any substantial earnings decline.

14 The zero earnings imputation for those with missing information does not affect the pool of individuals at risk of falling into shock in the successive two years as we require earnings to be greater than one-quarter of the annualised full-time minimum wage in each of the two years prior to the shock.

15 Individuals with missing information have earnings and income imputed to zero. If they earn more than the minimum threshold in each of the two previous years, they are classified as experiencing an earnings shock. The prevalence of earnings shocks is overstated as some of those individuals will not, in fact, have experienced a shock (that is, earnings will not have declined by at least 40 percent). In Appendix C we report the prevalence of earnings shocks without using imputation. Although the estimated prevalence of shocks decreases slightly, trends and patterns of earnings shocks are not affected.

Table 2.2. Relation between changes in earnings and changes in total income

		Change in total income		
	Number of observations (1)	Decrease 40—100% (2)	Decrease <40% (3)	Increase or no change (4)
Panel A: Males				
A. Decrease in earnings of 40–100%	677,914	78.01%	11.91%	10.08%
B. Decrease in earnings of <40%	1,825,117	2.12%	80.66%	17.22%
C. Increase or no decrease in earnings	5,420,766	0.29%	5.18%	94.53%
Panel B: Females				
A. Decrease in earnings of 40–100%	662,589	82.40%	9.71%	7.90%
B. Decrease in earnings of <40%	1,417,024	2.50%	80.54%	16.96%
C. Increase or no decrease in earnings	4,304,724	0.38%	5.15%	94.47%

Notes: Changes are calculated as the percentage change between the current year and the minimum value in the two previous years. Government benefits are excluded from total income.

In Table 2.2, we present statistics that depict the correlation between observed changes in earnings versus changes in total income. Panel A captures the information for males and Panel B captures the information for females. In the first column we report the number of observations based on a classification of the change in earnings. The first column reports the number of observations in each of three categories for earnings changes: a decrease in earnings of at least 40 percent; a decrease in earnings that is less than 40 percent; and no decrease in earnings. The following three columns then report the percentage of observations for each of these groups in analogous categories for changes in total income.

Focusing on the group that could be classified as experiencing a negative earnings shock (fall of more than 40 percent of earnings), 78 percent of males and 80 percent of females will be classified as experiencing a shock under our definition. Close to 12 percent of males and 10 percent of females experience a drop in total income, but the net drop is less than 40 percent. For both genders, a further 10 percent experience no change or an overall increase in income.

For this latter group, intuitively we likely would not want to classify these individuals as experiencing an earnings shock given it appears that they are drawing income from other sources to cover any drop in earnings. The more challenging issue, however, is whether to classify an individual whose earnings drop by more than 40 percent but whose total income drops by less than 40 percent as experiencing an earnings shock. Our analysis will not treat these individuals as experiencing a shock.[16]

We explore the robustness of the analysis in the Appendices by testing the sensitivity of our judgement calls on sample development and classification as experiencing an earnings shock. Appendix B shows that our analysis is robust to changes in the definition of total income. Appendix C tests the sensitivity of the analysis to the exclusion of non-lodger data and the zero-earnings assumption for missing values. Appendix D shows the robustness of the analysis to changes in the thresholds of the definition of earnings shock. While there will be differences in the shares of those identified as experiencing a shock, the trends over time and across age and other characteristics follow similar patterns as those discussed in this report.

16 It is worth pointing out that individuals with high spousal incomes are not over-represented among those who experience shock according to this definition. Median spousal income of those who experience a shock is close to the median income in the full sample, and the share of people with spouses in the top 1 percent of total annual income among those who experience shock is less than 0.5 percent.