Australian Government
Department of Social Services
Australian Institute of Family Studies

Growing Up in Australia:
The Longitudinal Study of Australian Children (LSAC)

LSAC Technical Paper No. 14

The longitudinal study of Australian children

Imputing income in the Longitudinal Study of Australian Children

Killian Mullan[1], Galina Daraganova and Kalyca Baker

Australian Institute of Family Studies

January 2015

[1] The analyses were conducted while the first author was on staff at the Australian Institute of Family Studies. Killian Mullan is now with the Centre for Time Use Research, Department of Sociology, University of Oxford.

Acknowledgements

This report makes use of data from *Growing Up in Australia: the Longitudinal Study of Australian Children* (LSAC). LSAC is conducted in partnership between the Department of Social Services (DSS), the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS), with advice provided by a consortium of leading researchers. Findings and views expressed in this publication are those of the individual authors and may not reflect those of AIFS, DSS or the ABS.

For more information, write to:

Research Publications Unit
Policy Strategy Branch
Australian Government Department of Social Services
PO Box 7576
Canberra Business Centre ACT 2610

Phone: (02) 6146 8061

Fax: (02) 6293 3289

Email: publications.research@dss.gov.au

Contents

List of tables

List of figures

1 Introduction

Income is one of the most important pieces of economic information collected in *Growing Up in Australia*: the Longitudinal Study of Australian Children (LSAC). Income is an important variable as it serves as an indicator of a family's ability to financially invest in children's development and access services that might support a child's wellbeing (Bradbury, 2007; Katz & Redmond, 2009). Useful on its own, income can also be combined with other socio-economic data to provide a comprehensive overview of the socio-economic contexts within which children live and develop (Hauser, 1994).

However, income tends to suffer from item non-response across all surveys (Yan, Curtin & Jans, 2010). This may occur because individuals do not know or cannot accurately state their (or their family/partner's) income, or do not wish to disclose this type of personal information (Giusti & Little, 2011). The result is that levels of missing income data tend to be higher than other variables and can pose difficulties for analyses that incorporate income.

A common approach to deal with missing items is to apply listwise or casewise deletion. However, this might lead to two main issues. First, exclusion of missing cases lowers the number of overall cases, which in turn reduces the power of the statistical analysis (Cohen, 1977). Second, generalisability of the interpretation of the analysis can be limited because missing data can be unevenly distributed across sample populations and can be related to the target variable (Little & Su, 1989). For example, those with very low or very high earnings may be more hesitant to report their income, thereby decreasing the overall variability of the income data and underrepresenting earners at the extremes of the distribution (Schräpler, 2006). In order to overcome this, researchers responsible for large national household surveys increasingly impute values through reliable and robust methods rather than simply excluding missing responses from the analysis (Frick & Grabka, 2007).

Therefore, this report aims to impute missing income data for Parent 1 and Parent 2 using the approach developed for another longitudinal study—the Household, Income and Labour Dynamics in Australia survey (HILDA). The HILDA method (Hayes & Watson, 2010) is a well regarded and widely accepted method based on decades of previous research (see for example Starick & Watson, 2007). At the completion of the second wave of HILDA, the methods team investigated the Little and Su method, among other methods, for imputing component income fields used to derive household income. The method that was adopted as a result of these investigations was an augmented Little and Su method, and it is this method that has been migrated to the LSAC. Unlike in HILDA, the role of income in LSAC is primarily a marker of the socio-economic status of primary parents. Rather than being a central purpose as it is in HILDA (Household, Income and Labour Dynamics), in LSAC the criticality of income is in relation to fluctuating resources as parents balance child caring and work, which shifts as the child grows up in a more or less predictable way over the income range.

The report gives an argument for imputing missing income fields in LSAC after four waves were completed, describes the selection of method and justifies this selection, and gives some summary information on the impact of imputation on distribution. The report is structured as follows. Section 2 explains the income variable in detail, including how it was asked and derived. Section 3 displays the percentage of the total sample that had missing data for Parent 1 and Parent 2 income in both the B and K cohorts. Section 4 outlines the methodology of the imputation used, explaining 'Nearest Neighbour' and 'Little and Su' methods. Finally, the report provides descriptive data on imputed income derived as a result of applying this methodology in Section 5 and concludes with the impact of these transformations on the income variable in Section 6.

This report uses the LSAC In Confidence data from Waves 1 through to 4. The In Confidence data was chosen because it includes more detailed information than the General Release data, as fewer confidentialisation techniques are applied to the raw data (e.g., top coding). At the time of the report writing, Wave 5 data had not been released and, therefore, were not included in this report.

2 Income variables

Currently in LSAC, information on individual and family income is not collected from every adult member of a household/family but rather is provided by the main respondent.[2] The main respondent (later referred to as Parent 1) is defined as the child's primary caregiver, or the parent who knows the child best in Wave 1. In the majority of instances, this is the child's biological mother, but is sometimes the father or another guardian.

Across all waves, Parent 1 was asked the question regarding her/his individual income and then her/his partner's individual income (where applicable). The answers were recorded in open ended, numeric format. In Wave 1, during the face-to-face interview, Parent 1 was asked: 'Before income tax is taken out, how much do you usually receive from all sources in total?' The same question was asked in relation to Parent 2 (where applicable). Immediately preceding this question a checklist of sources and a checklist of government support sources were shown (see Mullan & Redmond, 2011 for the items on the checklists). Income questions in subsequent waves (Waves 2, 3 and 4) followed a similar sequence with slight differences to the wording of the income question, the list of sources, and whether Parent 1 was prompted for her/his best estimate.

In Wave 2, rather than 'all sources' the wording was changed to 'this source/these sources'. Also, the list of sources was shortened, and the list of government sources was shown after the total income question. Finally, an instruction for interviewers was included that read 'If respondent is unable to answer please prompt for best estimate'. The prompt was given in all following waves. In Wave 3, the list of sources remained the same as in Wave 2, but the question changed to 'Before income tax, *salary sacrifice or anything else is taken out*, how much do you usually receive from all sources in total?' For Wave 4 the question and sources remained the same as in Wave 3. The changes are not substantial, however readers and data users should keep these changes in mind as they may nonetheless have consequences for longitudinal comparisons. The changes and consequences are discussed in details in Technical Paper No. 7 (Mullan & Redmond, 2011).

In all waves the income amount question was followed by 'what period does this cover?', with the options 'daily, weekly, fortnightly, monthly, annually, other'. Using these three questions (amount, source and period covered) the individual gross income was derived for Parent 1 and Parent 2. Specifically, the income variable was derived from the provided amount as well as the period, converted into a weekly amount for each parent to get the total weekly gross income variables for Parent 1 and Parent 2. Other information (such as income tax paid) from other informants (such as the Australian Taxation Office) was not acquired. If the amount was reported to be a weekly amount, the value remained the same. In order to make the income values comparable, if the income amount was reported in any other timeframe, it was converted to a weekly amount. For example, if income was reported by calendar month the value was divided by 4.35 (number of weeks in an average month). If income was reported yearly, then the value was divided by 52.18 weeks (365.25 days, which is the average number of days in a year, accounting for leap years, divided by 7 days a week).

As a result, the following income variables were derived (with * referring to a wave indicator):

- *fn09a – gross individual weekly income of Parent 1;
- *fn09b – gross individual weekly income of Parent 2.

[2] The income questions, respondents and the definition of main respondent is subject to change in future waves or reports.

The derived income was a continuous measure with extra possible response codes:

'–4' – when Parent 1 refused to answer all questions regarding income;

'–3' – when Parent 1 refused to answer the specific question regarding income;

'–2' – when Parent 1 did not know the amount;

'–99' – when Parent 1 reported a loss. Although a loss was classified as a legitimate response, there was no recorded information about the specific dollar amount of any negative income;

' . ' – nonsensical responses were set to missing (for more information see LSAC Issue Paper No. 5). Specifically, respondents who reported a zero or negative income, but also reported receiving government payments, had their income variable set to missing. Other unusual responses were queried on a case-by-case basis. These were reports of: government payments as the only source of income, but a dollar amount of over $750 a week; no salary, but large amounts of profit or loss (>$200,000 per year); and large incomes (>$260,000 per year). If responses were consistent with responses provided for other, related variables, income was unchanged. If responses were inconsistent, income was modified where possible, or set to missing. Finally, some participants' total weekly income could not be calculated because they had specified an amount, but no timeframe, or an ambiguous timeframe. For example, income of individuals who were reporting an hourly rate of pay, but no weekly working hours, was set to missing.

It is worth noting that starting from Wave 2, Parent 1 was asked more precise information about the income of other adult members of the household. Therefore a household income variable called 'hinc' is available in LSAC data sets from Wave 2 onwards. However, as there is not enough information on who, in particular, earns the money (e.g., we do not know if it was a parent), the focus of this report is on imputation of the individual incomes of Parent 1 and Parent 2. Though the original combined household income variable 'hinc' is not discussed in this report, or imputed, it is re-calculated for use in future LSAC data sets using imputed Parent 1 and Parent 2 income variables and renamed 'hinci', as described later in this paper.

It is important to remember that all surveys are susceptible to respondent errors. An attempt is made in LSAC collection to ensure that data are as accurate as possible, for example through interviewer training, and by not allowing outside of range values to be entered in computerised surveys. However, missing cases are a known inaccuracy in the data. Types of missing data are described next.

3 Missing income data in LSAC

Missing Parent 1 and/or Parent 2 income data in LSAC can be due to the following reasons:

- Parent 1 did not complete the income questions during the interview either because she/he did not know or refused to answer (item non-response);
- Parent 1 did not participate in that wave (wave non-response);
- Parent 1 reported a loss;
- Income was set to a missing value during the data cleaning process due to nonsensical responses (system missing).

With the exception of wave non-response, all cases with missing values were imputed. Note that values of zero were considered legitimate responses. For the Parent 2 income variable only, a legitimate 'not asked' due to 'not applicable' was also possible, in the case of sole parents. Table 1 shows the percentage of responses with missing data on Parent 1 and Parent 2 income variables for both cohorts B and K, displayed separately for each wave.

The extent of missingness is similar between cohorts and Parents 1 and 2; however, it varies between waves. The highest level of missing responses was observed for Wave 1 responses. Previous research on the Wave 1, K cohort indicated that higher rates of missing data on income tended to be associated with other demographic variables of age, employment status, relationship status and education (Mullan & Redmond, 2011). Generally non-response was associated with older ages, unemployed or self-employed status (of Parent 1 and/or Parent 2), coupled parents (for Parent 1) and different statuses between parents (e.g., Parent 2 was employed or more educated than Parent 1, who was the primary respondent) (Mullan & Redmond, 2011). This provides evidence that the missing responses are not missing completely at random, and it is important to impute missing data using a reliable and robust method rather than simply ignore missing responses.

Table 1: Proportions of missing/non-missing data for income variable, by wave, cohort and parent

	Wave 1	Wave 2	Wave 3	Wave 4
Cohort B				
		Parent 1 (%)		
Usual gross weekly income reported	89.8	96.9	96.6	96.0
Usual gross weekly income missing	10.2	3.1	3.4	4.0
Total	**100.0**	**100.0**	**100.0**	**100.0**
Number of observations, N	**5,107**	**4,606**	**4,386**	**4,242**
		Parent 2 (%)		
Usual gross weekly income reported	86.0	96.4	95.4	93.9
Usual gross weekly income missing	14.1	3.6	4.8	6.2
Total	**100.0**	**100.0**	**100.0**	**100.0**
Number of observations, N	**4,630**	**4,099**	**3,900**	**3,706**

Table 1: Proportions of missing/non-missing data for income variable, by wave, cohort and parent

	Wave 1	Wave 2	Wave 3	Wave 4
Cohort K				
		Parent 1 (%)		
Usual gross weekly income reported	88.5	97.0	96.2	94.9
Usual gross weekly income missing	11.5	3.0	3.8	5.1
Total	**100.0**	**100.0**	**100.0**	**100.0**
Number of observations, N	**4,983**	**4,464**	**4,331**	**4,164**
		Parent 2 (%)		
Usual gross weekly income reported	84.3	95.2	94.2	90.8
Usual gross weekly income missing	15.7	4.8	5.8	9.2
Total	**100.0**	**100.0**	**100.0**	**100.0**
Number of observations, N	**4,286**	**3,804**	**3,708**	**3,512**

Note: The total sample size is smaller for the Parent 2 variable than the Parent 1 variable because sole parents are removed as eligible respondents for the Parent 2 variable. Percentages may not total 100% due to rounding.

Source: LSAC, Waves 1–4, Cohorts B and K.

4 Imputation methodology

To impute the missing individual income data, the following imputation methods were employed in LSAC:

- Nearest Neighbour (NN) method
- Little and Su method.

Based on earlier work evaluating different methodologies for imputing income, these methods were found to produce the most robust and reliable estimates of missing income data (Starick & Watson, 2007) and have been implemented in the HILDA survey (Hayes & Watson, 2010). Both of these methods use the concept of 'recipient' and 'donor'. The 'recipient' is the record with missing income data, whereas the 'donor' is the record with complete income data. In both methods the donor's data are used to impute the recipient's missing data, but how the income value is imputed varies across methods.

4.1 Nearest neighbour method

The Nearest Neighbour (NN) method (Little, 1988) imputes a donor (participant with data) value to the recipient (participant with missing data) based on a Poisson regression model. We estimated a Poisson regression because there is a substantial proportion of mothers in the sample with zero income, which raises non-ignorable problems for log-linear models that are often used when modelling income (Nichols, 2010; see also Gould, 2011).

Specifically, the NN method uses the predicted income from multivariate regressions to pair each case with missing income data (the recipient) to a case with reported income (the donor) based on the similarity of the respective income predicted by the regression model. Once a 'nearest neighbour' has been identified for each case with missing data, the donor's reported income is imputed to the recipient. The advantage of the NN method is its ability to impute every record within a single wave.

To impute individual income of Parent 1 and Parent 2 at each wave, regression models were fitted for each wave and each cohort separately. The log of income was used as the dependent variable and a number of key variables of interest were used as predictor variables. The variables were chosen based on the previous research and the exploratory analysis (Mullan & Redmond, 2011). These independent variables are presented in Table 2. Note that Parent 1 and Parent 2 income values were imputed simultaneously, i.e. in the same regression model.

Table 2: Independent variables in the linear regression model for predicted income (log)

Metric variable label (abbreviated)	Values
Age*	Number
Age squared	Number
Home—SEIFA Disadvantage**	Number
Home—SEIFA Economic Resources**	Number
SEIFA Education & Occupation**	Number
Number of same-age siblings in the household	Number
Number of younger siblings in the household	Number
Number of older siblings in the household	Number
Average weekly working hours*	Number

Table 2: Independent variables in the linear regression model for predicted income (log)

Binary/categorical variable (abbreviated)	Values
Gender	0 – Male 1 – Female
Partner income	0 – Sole parents*** 1 – Bottom 25% 2 – Middle 50% 3 – Top 25% 4 – Unspecified (missing)
Parent has a degree	0 – No degree 1 – Degree
Parent finished year 12	0 – No year 12 1 – Year 12 2 – Unspecified (missing)
Parent main language spoken at home	0 – Other 1 – English
Parent employed type	1 – Not in labour force 2 – Unemployed 3 – Maternity leave 4 – Full-time permanent/fixed 5 – Full-time casual 6 – Full-time unspecified 7 – Part-time permanent/fixed 8 – Part-time casual 9 – Part-time unspecified 10 – Unspecified (missing)
Parent occupation	1 – Managers 2 – Professionals 3 – Associate professionals 4 – Tradespersons 5 – Advanced clerical and service providers 6 – Intermediate clerical and service providers 7 – Intermediate production and transport 8 – Elementary clerical, sales and service 9 – Labourers 10 – Unspecified (missing)
Housing type	1 – Buyer 2 – Renter 3 – Owner 4 – Other 5 – Unspecified (missing)
Housing region	1 – Metropolitan 2 – Regional

Table 2: Independent variables in the linear regression model for predicted income (log)

Parent main source of income	1 – Wages or salary 2 – Business 3 – Rental property 4 – Dividends or interest 5 – Government support 6 – Child support 7 – Superannuation 8 – Workers' compensation 9 – Other 10 – Unspecified (missing)
Parent has an ongoing medical condition	0 – No 1 – Yes
Parent is Indigenous	0 – No 1 – Yes

Note: * Missing values were replaced with mean. Only a small number of respondents had missing age (less than 0.5%);
** Missing values were replaced with the corresponding values from previous wave.
*** Included in the P1 models only (sole parent is the reference category);
Partner's characteristics were only included in the model if Parent 1 had a partner.

Sources: LSAC, Waves 1–4, Cohorts B and K.

The predicted log values from the regression were then transformed to the dollar value and used as the recipient's estimated income value. Of the complete responses from the remaining sample, the respondent with the predicted value that most closely matched the recipient's estimated value was then chosen as the donor. The donor's observed value was inserted for the recipient's income.

The most closely matched donor was chosen as follows (Hayes & Watson, 2009):

First, donors (*d*) and recipients (*i*) were restricted to fall in the same age group. Three age classes were created (based on population thirds) and the donor was required to come from within the appropriate class where possible.

Second, the absolute difference between the predicted value of the recipient and the predicted value of the potential donor was calculated:

$$| \hat{\mu}_i - \hat{\mu}_{d_j} | \quad (1)$$

where $\hat{\mu}_i$ is the predicted mean of Y (income value) for recipient i;

$\hat{\mu}_{d_j}$ is the predicted mean of Y_d (income value) for donor j.

Thirdly, the closest donor was chosen based on the smallest value of the absolute difference calculated in equation (1):

$$|\hat{\mu}_i - \hat{\mu}_{d_c}| \leq |\hat{\mu}_i - \hat{\mu}_{d_{j-c}}| \quad (2)$$

where d_c refers to the donor who has the closest predicted value to the recipient i;

d_{j-c} refers to all potential donors except donor c;

$\hat{\mu}_i$ is the predicted mean of Y (income value) for recipient i;

$\hat{\mu}_{d_c}$ is the predicted mean of Y_{d_c} (income value) for the closest donor c;

$\hat{\mu}_{d_{j-c}}$ is the predicted mean of $Y_{d_{j-c}}$ (income value) for all potential donors, excluding donor c.

Finally, when the closest donor was identified, the observed value of donor ***d*** replaced missing income for the recipient ***i***:

$$\hat{Y}_i = Y_d \quad (3)$$

where $\hat{Y}_i$ refers to the imputed income of the recipient ***i***;

Y_d refers to the observed income of donor ***d***.

4.2 Little and Su method

The Little and Su method is considered superior for longitudinal data because the imputation incorporates information about variation in income across waves. However, this method is not feasible in all instances and it is thus supplemented with a 'nearest neighbour' method (NN). Specifically, the Little and Su method cannot be used if:

- Parent 1 or Parent 2 reported non-zero income only at one wave of data collection;
- Parent 1 or Parent 2 participated only at one wave of data collection.

The Little and Su method (1989) calculates trend values based on data reported across multiple waves for respondents with complete data. This is referred to as the column effect. For respondents where income data are missing (recipient) for one wave, information from other waves is usually available. Using the previous income responses, the recipient's departure from the overall trend is calculated. This is referred to as the row effect. Finally, the method incorporates the value from the closest respondent with complete data (donor). This is referred to as the residual effect. The model is calculated as follows:

imputation = (roweffect) (columneffect) (residualeffect) (4)

where:

(a) Column (wave) effect is calculated for complete responses as follows:

$$c_j = \frac{\bar{Y}_J}{\frac{1}{k}\sum_j \bar{Y}_J} \quad (5)$$

here ***k*** refers to a number of waves (in our case ***k***=4);

j refers to a particular wave such that ***j*** = 1, … , ***k*** (in our case ***k***=4);

refers to the sample mean income amount for wave ***j*** based on complete responses.

(b) Adjusted row (respondent) effect is calculated for both complete and incomplete responses as follows:

$$\bar{Y}^{(i)} = \frac{1}{k}\sum_j \frac{Y_{ij}}{c_j} \quad (6)$$

where ***k*** refers to a number of waves (in our case ***k***=4);

j refers to a particular wave such that ***j*** = 1, … , ***k*** (in our case ***k***=4);

i refers to a respondent ***i***;

Y_{ij} refers to the income variable for the respondent ***i*** of wave ***j***;

c_j refers to the column effect calculated using equation (5).

(c) Residual effect (stochastic component) is calculated only after column and row effects are calculated. Firstly, responses are ordered by row effects (). Secondly, incomplete case ***i*** (the recipient) is matched to the closest complete case ***d*** (donor). Note, that the closest donor is identified in the same way as in the NN method. Then, the residual effect is calculated as:

$$R_i = \frac{Y_{dj}}{\bar{Y}^{(d)} c_j} \quad (7)$$

where Y_{dj} refers to the income of the closest donor ***d*** at wave ***j***;

$\bar{Y}^{(d)}$ refers to the row effect of the donor ***d*** calculated using equation (6);

c_j refers to the column effect calculated using equation (5).

Substituting column, row and residual effect calculations in equation (4), the missing income value for person ***i*** at wave ***j*** $(\hat{Y}_{ij})$ is imputed as follows:

$$\hat{Y}_{ij} = [c_j][\bar{Y}^{(i)}]\,[R_i] =$$

$$= [\mathbf{c_j}][\bar{\mathbf{Y}}^{(\mathbf{i})}]\left[\frac{\mathbf{Y_{dj}}}{\bar{\mathbf{Y}}^{(\mathbf{d})}\mathbf{c_j}}\right] = \frac{\bar{\mathbf{Y}}^{(\mathbf{i})}\mathbf{Y_{dj}}}{\bar{\mathbf{Y}}^{(\mathbf{d})}} \quad (8)$$

Those values that were not possible to impute with the Little and Su method were imputed using the Nearest Neighbour method. As in the Nearest Neighbour method, donors and recipients were matched within longitudinal imputation classes defined by three age classes (based on population thirds).

4.3 Combination of methods used in LSAC

To impute missing income for Parent 1 and Parent 2 we used the combination of these two methods:

Step 1—The Nearest Neighbour method (within age class imputation) was calculated to identify a donor for each recipient. The values from the closest donors were imputed for all missing responses at each wave.

Step 2—The Little and Su method (within age class imputation) was then calculated and the Little and Su value imputed to all possible recipients. The longitudinal approach of the Little and Su method makes it preferable to (or more accurate than) the Nearest Neighbour method, and therefore values from Step 2 replace values from Step 1 where possible. Therefore at the end of the process there are recipients with values imputed from both methods.

5 Income imputation

In LSAC, income imputation was used to complete the missing values for Parent 1 and Parent 2 individual incomes. Note, that no household income was imputed.

There were a number of imputation rules applied:

- The Nearest Neighbour method was used to impute the Parent 1/Parent 2 income data regardless of whether Parent 1/Parent 2 swapped their roles or a new household member became Parent 1 or Parent 2.
- The Little and Su method was used to impute the income data for Parent 1 and Parent 2 who were identified at Wave 1 as the household member 2 and the household member 3. The individual income was imputed only for these Parent 1 (household member 2) and Parent 2s (household member 3) across all waves, regardless of whether they swapped roles or not. However, if a new household member became Parent 1 or Parent 2 (not household members 2 or 3) at any other wave of data collection, her/his income was not imputed with the Little and Su method.
- For households with a Parent 2 present, only respondents where Parent 1 and Parent 2 were a couple were used (Parent 2 was Parent 1's spouse or de-facto partner).

In all imputations we used In Confidence data. After the individual income was imputed the extreme responses were top or bottom coded using a case-by-case approach.

5.1 Distribution of imputation

The distribution of the income variable (for both Parent 1 and Parent 2) before and after imputation with both methods is presented separately for the B and K cohorts in Figures 1 and 2, respectively. For the B cohort, the distributions were similar for pre- and post-imputed values across all waves with more differences observed for the K cohort. At Waves 1 and 4, the Little and Su method produced similar distributions compared to the Nearest Neighbour method.

The evaluation research by Starick and Watson (2007) tested a number of different imputation methods and provided strong evidence that the combination of the Nearest Neighbour and the Little and Su methods performs very well in the longitudinal context. Our results suggest that the combination of Little and Su and Nearest Neighbour methods is the preferable imputation approach, rather than just use of Nearest Neighbour method in replicating the distributional properties of the income data.

Table 3 compares the unweighted distributions of the confidentialised income variable pre- and post-imputation using the combined imputation method for Parent 1 and Parent 2, by wave and cohort. The imputation has a negligible impact on the distribution of the income for Parent 1 and Parent 2.

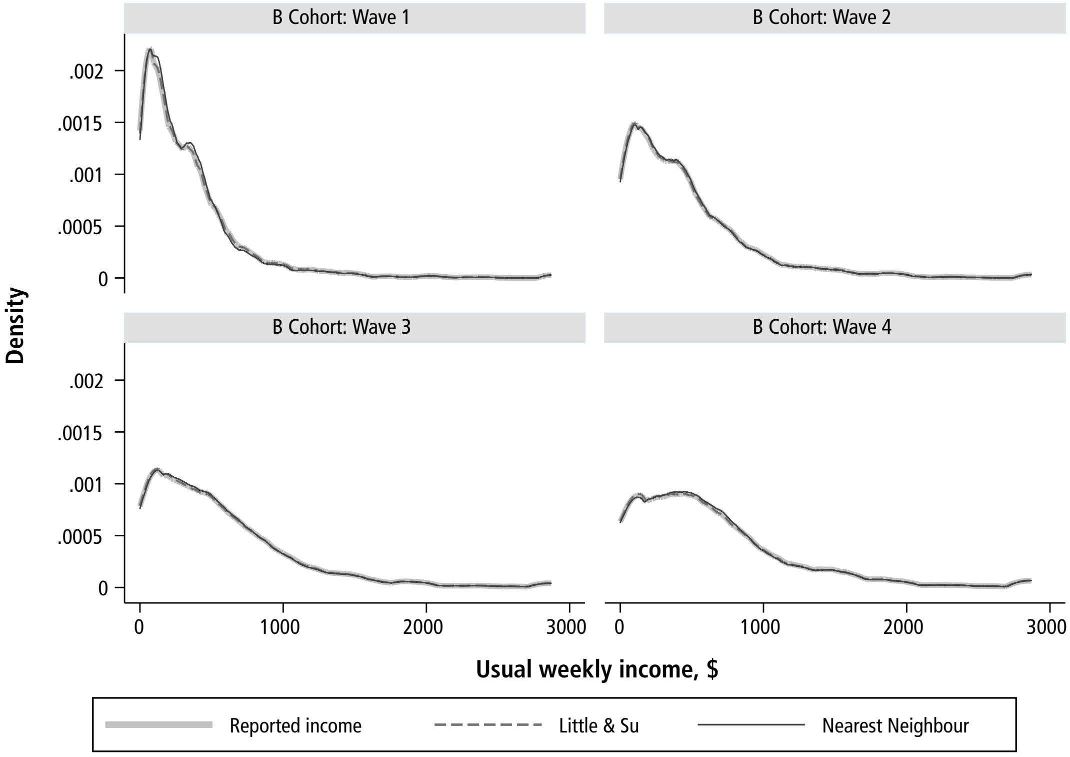

Figure 1: Distribution of income variable before and after imputation with both methods for the B cohort, using Kernel density graph

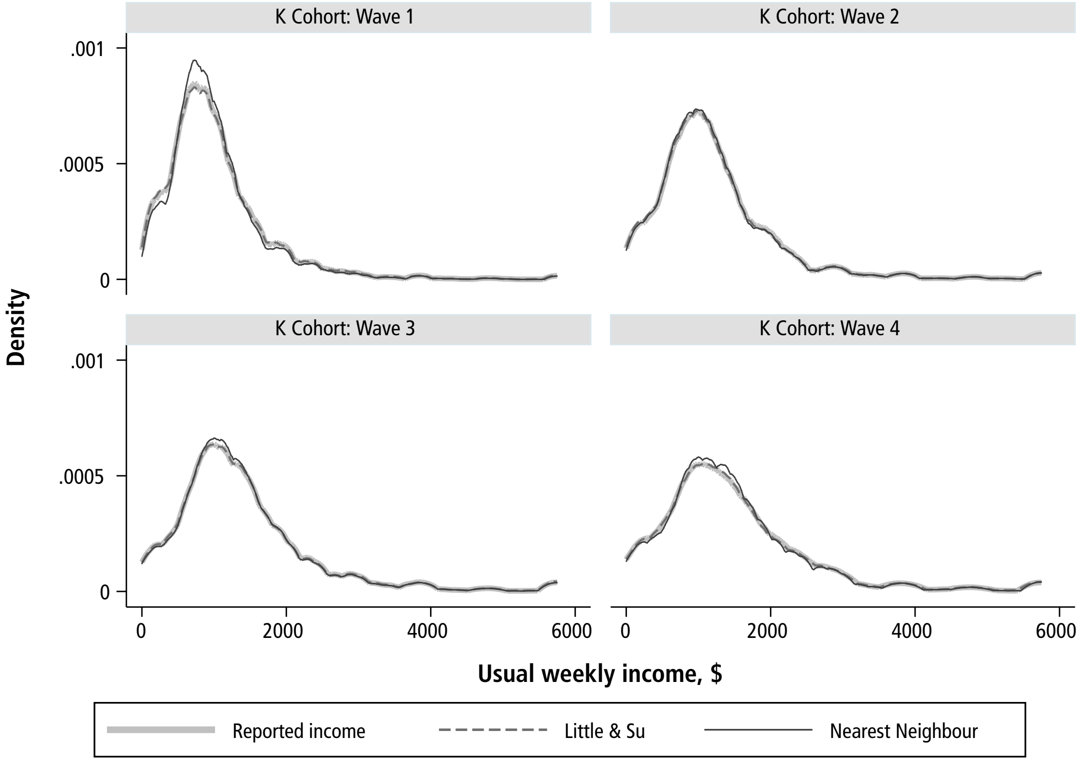

Figure 2: Distribution of income variable before and after imputation with both methods for the K cohort, using Kernel density graph

Table 3: Descriptive statistics of the unweighted distribution of weekly gross income ($) before and after imputation, by wave, cohort and parent

	Before imputation			After imputation		
	Mean	Median	Standard deviation	Mean	Median	Standard deviation
B cohort						
			Parent 1			
Wave 1	331.8	239.8	357.9	334.3	240.0	360.5
Wave 2	450.8	350.0	430.2	450.1	350.0	431.1
Wave 3	547.2	430.0	495.2	545.5	425.0	495.4
Wave 4	630.6	500.0	540.8	627.7	500.0	542.4
			Parent 2			
Wave 1	984.7	850.0	720.6	988.9	850.0	726.2
Wave 2	1,209.8	1,050.0	861.1	1,201.1	1,034.0	858.9
Wave 3	1,409.7	1,200.0	987.7	1,406.3	1,200.0	992.4
Wave 4	1,477.7	1,322.3	1,011.3	1,481.0	1,322.3	1,021.8
K cohort						
			Parent 1			
Wave 1	431.7	350.0	374.2	432.8	350.0	377.5
Wave 2	530.0	421.6	455.2	526.3	421.6	541.9
Wave 3	639.2	631.8	504.5	635.3	625.0	503.1
Wave 4	718.9	600.0	543.2	715.5	600.0	542.9
			Parent 2			
Wave 1	1,019.5	881.6	720.7	1,028.7	898.2	747.9
Wave 2	1,228.7	1,054.0	865.8	1,220.4	1,054.0	867.0
Wave 3	1,396.0	1,200.0	979.7	1,398.4	1,200.0	988.7
Wave 4	1,473.6	1,264.9	1,046.7	1,472.1	1,264.6	1,050.3

Source: LSAC, Waves 1–4, Cohorts B and K.

Table 4 presents the proportion of responses imputed by the Nearest Neighbour and Little and Su methods in the combined imputation. It can be seen that the vast majority of missing income responses were imputed using the Little and Su method.

Table 4: Proportion of missing responses imputed by the Nearest Neighbour and the Little and Su method in combined imputation approach, by wave, cohort and parent

Method	Wave 1		Wave 2		Wave 3		Wave 4	
	P1	P2	P1	P2	P1	P2	P1	P2
Cohort B	%							
Nearest Neighbour method	14.0	19.0	11.3	25.7	7.4	20.0	4.7	18.3
Little and Su method	86.0	81.0	88.7	74.3	92.6	80.0	95.3	81.7
Total number of missing responses, N	523	648	142	148	149	180	170	224
Cohort K	%							
Nearest Neighbour method	15.0	18.7	16.3	29.3	12.3	26.9	15.3	23.1
Little and Su method	85.0	81.3	83.7	70.7	87.7	73.1	84.7	76.9
Total number of missing responses, N	574	669	135	181	163	321	216	321

Source: LSAC, Waves 1–4, Cohorts B and K.

5.2 Imputed variables provided in datasets

Table 5 provides a list of variables that have been imputed for release to data users. There are three individual income variables for Parent 1/Parent 2: pre-imputed income variable, post-imputed variable and a flag-indicator variable. The pre-imputed income variable for Parent 1/Parent 2 is the variable that has been provided since Wave 1 and includes all the possible responses for income ('–99' – loss; '–9' – not applicable; '–4' – section refused; '–3' – question refused; '–2' – do not know; ' . ' – missing responses and the amount). The post-imputed income variable for Parent 1/Parent 2 contains the reported income value for responses where imputation was not required and the imputed value where imputation was required. The flag indicator variable has value '0' if the corresponding value was not imputed, '1' if it was imputed using the Nearest Neighbour method and '2' if it was imputed using the Little and Su method. The combined household income was also re-calculated using the imputed Parent 1 and Parent 2 income data where it was applicable.

Table 5: Imputed variables provided

	Pre-imputed	Post-imputed	Flag
Parent 1 individual income (member 2)	*fn09a	*fn09ai	*fn09aif
Parent 2 individual income (member 3)	*fn09b	*fn09bi	*fn09bif
Household income	*hinc	*hinci	–

Note: * refers to the letter corresponding to the child's age ('a' for 0–1 years old, 'c' for 4–5 years old, etc.). '–' = not applicable.

Household income was derived by adding P1, P2 (where applicable) and remaining household members' income (see Mullan & Redmond, 2011 for more information). For this reason, if income for other household members was missing then it was impossible to calculate the household income. Table 6 presents the proportion of missing responses pre- and post re-calculation for the household income variable. For more than 50 per cent of missing responses the household income was re-calculated using the imputed individual income.

Table 6: Proportion of missing responses for the original and re-calculated household income variables, by wave and cohort

	Wave 1	Wave 2	Wave 3	Wave 4
Cohort B		%		
Missing responses in *hinc	–	8.1	8.7	10.3
Missing responses in *hinci	–	3.7	3.3	4.4
Total number of responses, N	–	4,606	4,386	4,242
Cohort K		%		
Missing responses in *hinc	–	8.5	9.5	13.5
Missing responses in *hinci	–	3.7	4.0	5.9
Total number of responses, N	–	4,464	4,331	4,164

Note: Household income was not calculated for Wave 1 as insufficient information was collected in Wave 1 to derive this variable.

6 Concluding remarks

This paper has documented how the Little and Su and Nearest Neighbour methods have been used in combination to impute income for the LSAC data. The income variables (weekly gross income for Parent 1 and Parent 2) are important for analysis and the paper discusses the imputation method. The modified variables contain more responses, potentially increasing the power of analyses conducted with these variables. Further, the modified variables are considered to be more representative of the overall sample and therefore implications based on analysis using these variables are potentially more generalisable. In conclusion, although all imputation methods are subject to error and bias, the modified LSAC income variables are considered more appropriate for analysis than excluding missing responses altogether. Users may nonetheless want to perform analysis using the imputed variables as well as the non-imputed variables and compare obtained results to ensure imputed income is not having an undue influence on their particular topic of research.

Users of LSAC data should be aware that the imputed values may be revised as subsequent waves of data become available, either by incorporating new data into original models, or by revising the models used. This practice ensures latest releases represent the best available information. Changes will be reported on the LSAC website or in future technical papers.

References

Bradbury, B. (2007). *Child Outcomes and Family Socio-Economic Characteristics: Final Report of the Project LSAC Outcomes and the Family Environment*, May 2007, Social Policy Research Centre, University of NSW.

Cleaning of Income data. *LSAC Data Issue Paper* No. 5. AIFS, 2006. **http://www.growingupinaustralia.gov.au/pubs/issues/ip5.pdf**

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Frick, J.R. & Grabka, M.M. (2007). *Item Non-Response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective*. Institute for the Study of Labor Discussion Paper No. 3043, September 2007, Institute for the Study of Labor, Bonn, Germany.

Giusti, C. & Little, R.J.A. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics*, *27*(2), 211–229.

Gould, W. 2011. "Use poisson rather than regress; tell a friend". The Stata Blog: **http://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/**

Hauser, R.M. (1994). Measuring socioeconomic status in studies of child development. *Child Development*, *65*(6), 1541–1545.

Hayes, C. & Watson, N. (2009). *HILDA imputation methods*. HILDA Project Technical Paper Series, No. 2/09, December 2009 [revised January 2010]. Melbourne: The Melbourne Institute of Economic and Social Research.

Katz, I. & Redmond, G. (2009). Family Income as a Protective Factor for Child Outcomes. In K. Rummery, I. Greener & C. Holden (eds.), *Social Policy Review 21* (pp. 167–197). Policy Press, Bristol.

Little, R.J.A. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*, 287–296.

Little, R.J.A., & Su, H.L. (1989). *Item non-response in panel surveys*, in *Panel surveys*, ed. D. Kasprzyk, G.J. Duncan, G. Kalton & M.P. Singh, New York: Wiley.

Mullan, K. & Redmond, G. (2011). *Validating Income in the Longitudinal Study of Australian Children*. LSAC Technical Paper No. 7, October 2011, Australian Institute of Family Studies, Melbourne. **http://www.growingupinaustralia.gov.au/pubs/technical/tp7.pdf**

Nichols, A. 2010. Regression for non-negative skewed dependent variables. Paper presented at the Stata Conference, Boston 2010.

Schräpler, J-P. (2006). Explaining income nonresponse—A case study by means of the British Household Panel Study (BHPS). *Quality and Quantity*, *40*(6), 1013–1036.

Starick, R. & Watson, N. (2007). *Evaluation of alternative income imputation methods for the HILDA survey*. HILDA Project Technical Paper Series, No. 1/07, June 2007. Melbourne: The Melbourne Institute of Economic and Social Research.

Yan, T., Curtin, R. & Jans, M. (2010). Trends in income nonresponse over two decades. *Journal of Official Statistics*, 26(1), 145–164.