THE UNIVERSITY OF MELBOURNE

HILDA PROJECT TECHNICAL PAPER SERIES
No. 2/09, December 2009

[Revised January 2010]

HILDA Imputation Methods

Clinton Hayes and Nicole Watson

The HILDA Project was initiated, and is funded, by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs

MELBOURNE INSTITUTE
of Applied Economic and Social Research

Contents

List of Tables

Introduction

Missing data is a well known and extensively research topic for household surveys. Watson and Wooden (2002) assessed the non-response problem for the Household, Income and Labour Dynamics in Australia (HILDA) Survey with wave 1 data and from this initial research it was established that imputation would be used to deal with missing data in the HILDA Survey.

The HILDA imputation strategy for Release 2 is documented in Watson (2004). Since then considerable changes have been made to the entire imputation process. Starick and Watson (2007) evaluated a range of possible imputation methods and the results have been used to improve the imputation system. While these changes have been documented in the HILDA User Manual (Watson, 2009), this paper details the imputation strategies currently in use for the HILDA Survey.

The most significant change was made in Release 3 with the shift in the primary imputation method. In Release 2, the nearest neighbour regression method was the primary imputation method. With only two waves of data available, the benefit of including data from another wave, although thought to be helpful, was not a key component of the imputation at that stage. From Release 3 onwards, the Little and Su method is the primary imputation method and this capitalizes on the ability to look over an individual’s (or household’s) data series. Other more modest revisions have also been made between Release 4 and 6.

The two main topics requiring imputation are income and wealth. Variables from both these domains experience a higher proportion of missingness than other data and are considered key variables for the HILDA Survey. The wealth module has only been included in the questionnaires in wave 2 and wave 6 though home value has been asked every wave. Additionally, age and employment status have been imputed as these variables are vital inputs to the imputation and weighting processes (Watson, 2004).

At the time of writing, Release 8 data was not yet final, so the Tables in this paper refer to Release 7 data. The scope of the variables imputed in Release 8 has been extended to include a more disaggregated model of benefit income and the expenditure variables primarily collected in the Self-Completion Questionnaire. The User Manual for Release 8 will incorporate information on the changes to the benefit variables and a separate HILDA Technical paper will be released early in 2010 regarding the expenditure imputation.

Individuals who do not provide an interview or do not give an answer to a particular question generally show systematic differences from the rest of the sample. Imputation aims to correct for these differences and improve the usefulness of the data. Ignoring cases with missing values is not appropriate where the missingness is non-random. We recommend the use of data with imputed values in the analysis of income or wealth, or at a minimum, analyses with and without imputation should be compared to identify and understand the differences.

Imputed Variables Provided in Release 7 Datasets

This section lists all variables in the HILDA Survey that have been imputed for Release 7. Generally we have provided users with the pre-imputed variables (i.e. reported by the respondent), the post-imputed variables and a flag indicating which values are reported and which are imputed. While users only need the pre- and post-imputed variables or the post-imputed and the flag variables, we thought the extra flexibility of all three variables would be of assistance to users. The post-imputed variables contain the reported value for cases where no imputation was required and the imputed value for those that do.

An overview of the imputed variables for the responding person file, enumerated person file and the household file is provided in Table 1, Table 2 and Table 3 respectively. The first letter of the variable names in each table (represented as an underscore '_') should be replaced by the letter corresponding to the wave ('a' for wave 1 and 'b' for wave 2, etc.). Wealth data, with the exception of home value, is only available in waves 2 and 6 (the expectation is that the wealth module will be repeated on a 4-year cycle).

Table 1: Imputed variables provided in the Release 7 responding person file

	Pre-Imputed	*Post-Imputed*	*Flag*
Current income			
Wages and salaries – all jobs	_wsce	_wscei	_wscef
Wages and salaries – main job	_wscme	_wscmei	_wscmef
Wages and salaries – other jobs	_wscoe	_wscoei	_wscoef
Benefits	_bncaup	_bncaupi	_bncaupf
Financial year income			
Wages and salaries	_wsfe	_wsfei	_wsfef
Australian govt pensions	_bnfaup	_bnfaupi	_bnfaupf
Foreign govt pensions	_bnffp	_bnffpi	_bnffpf
Business income	_bifn, _bifp	_bifin, _bifip	_biff
Investments	_oifinvn, _oifinvp	_oifinin, _oifinip	_oifinf
Private pensions	_oifpp	_oifppi	_oifppf
Private transfers	_oifpt	_oifpti	_oifptf
Total FY income	Not provided	_tifefn, _tifefp	_tifeff
Windfall income	_oifwfl	_oifwfli	_oifwflf
Assets			
Joint bank accounts	_pwjbank	_pwjbani	_pwjbanf
Own bank accounts	_pwobank	_pwobani	_pwobanf
Superannuation – retirees	_pwsupr	_pwsupri	_pwsuprf
Superannuation – non-retirees	_pwsupwk	_pwsupwi	_pwsupwf
Debts			
HECS debt	_pwhecdt	_pwhecdi	_pwhecdf
Joint credit cards	_pwjccdt	_pwjccdi	_pwjccdf
Own credit cards	_pwoccdt	_pwoccdi	_pwoccdf
Other personal debt	_pwothdt	_pwothdi	_pwothdf
Other			
Age	Not provided	_hgage	_hgagef

Table 2: Imputed variables provided in the Release 7 enumerated person file

	Pre-Imputed	*Post-Imputed*	*Flag*
Current income			
Wages and salaries – all jobs	Not provided	_wscei	_wscef
Wages and salaries – main job	Not provided	_wscmei	_wscmef
Wages and salaries – other jobs	Not provided	_wscoei	_wscoef
Benefits	Not provided	_bncaupi	_bncaupf
Financial year income			
Wages and salaries	Not provided	_wsfei	_wsfef
Australian govt pensions	Not provided	_bnfaupi	_bnfaupf
Foreign govt pensions	Not provided	_bnffpi	_bnffpf
Business income	Not provided	_bifin, _bifip	_biff
Investments	Not provided	_oifinin, _oifinip	_oifinf
Private pensions	Not provided	_oifppi	_oifppf
Private transfers	Not provided	_oifpti	_oifptf
Total FY income	Not provided	_tifefn, _tifefp	_tifeff
Windfall income	Not provided	_oifwfli	_oifwflf
Assets			
Joint bank accounts	Not provided	_pwjbani	_pwjbanf
Own bank accounts	Not provided	_pwobani	_pwobanf
Superannuation – retirees	Not provided	_pwsupri	_pwsuprf
Superannuation – non-retirees	Not provided	_pwsupwi	_pwsupwf
Debts			
HECS debt	Not provided	_pwhecdi	_pwhecdf
Joint credit cards	Not provided	_pwjccdi	_pwjccdf
Own credit cards	Not provided	_pwoccdi	_pwoccdf
Other personal debt	Not provided	_pwothdi	_pwothdf
Other			
Age	Not provided	_hgage	_hgagef
Employment status (wave 2 non-respondents)	Not provided	bhgebi	bhgebf

Table 3: Imputed variables provided in the Release 7 household file

	Pre-Imputed	*Post-Imputed*	*Flag*
Current income			
Wages and salaries – all jobs	Not provided	_hiwscei	_hifwscef
Wages and salaries – main job	Not provided	_hiwscmi	_hifwscmf
Wages and salaries – other jobs	Not provided	_hiwscoi	_hifwscof
Benefits	Not provided	_hicaupi	_hicaupf
Financial year income			
Wages and salaries	Not provided	_hiwsfei	_hifwsfef
Australian govt pensions	Not provided	_hifaupi	_hifaupf
Foreign govt pensions	Not provided	_hiffpi	_hiffpf
Business income	Not provided	_hibifin, _hibifip	_hifbiff
Investments	Not provided	_hifinin, _hifinip	_hifinf
Private pensions	Not provided	_hifppi	_hifppf
Private transfers	Not provided	_hifpti	_hifptf
Total FY income	Not provided	_hifefn, _hifefp	_hifeff
Windfall income	Not provided	_hifwfli	_hifwflf
Assets			
Joint bank accounts	_hwjbank	_hwjbani	_hwjbanf
Own bank accounts	_hwobank	_hwobani	_hwobanf
Children's bank accounts	_hwcbank	_hwcbani	_hwcbanf
Superannuation – retirees	_hwsupr	_hwsupri	_hwsuprf
Superannuation – non-retirees	_hwsupwk	_hwsupwi	_hwsupwf
Business assets	_hwbusva	_hwbusvi	_hwbusvf
Cash investment	_hwcain	_hwcaini	_hwcainf
Equity investment	_hweqinv	_hweqini	_hweqinf
Collectables	_hwcoll	_hwcolli	_hwcollf
Home asset	_hwhmval	_hwhmvai	_hwhmvaf
Home value	_hsvalue	_hsvalui	_hsvaluf
Other property assets	_hwopval	_hwopvai	_hwopvaf
Life insurance	_hwinsur	_hwinsui	_hwinsuf
Trust funds	_hwtrust	_hwtrusi	_hwtrusf
Vehicles value	_hwvech	_hwvechi	_hwvechf
Total household assets	_hwasset	_hwassei	_hwassef
Debts			
HECS debt	_hwhecdt	_hwhecdi	_hwhecdf
Joint credit cards	_hwjccdt	_hwjccdi	_hwjccdf
Own credit cards	_hwoccdt	_hwoccdi	_hwoccdf
Other personal debt	_hwothdt	_hwothdi	_hwothdf
Business debt	_hwbusdt	_hwbusdi	_hwbusdf
Home debt	_hwhmdt	_hwhmdti	_hwhmdtf
Other property debt	_hwopdt	_hwopdti	_hwopdtf
Overdue household bills (w6 only)	_hwobdt	_hwobdti	_hwobdtf
Total household debts	_hwdebt	_hwdebti	_hwdebtf
Net worth	_hwnetwp, _hwnetwn	_hwnwip, _hwnwin	_hwnwf

Missing Data

Missing data in the HILDA Survey is classified into three distinct groups:

- *Item non-response* – Item non-response occurs when a respondent does not provide complete answers to all questions during their interview, either because they do not know or they refuse to provide the answer.
- *Wave non-response* – Wave non-response is where an individual (or household) has failed to provide an interview for that wave of the survey.
- *Unit non-response* – Unit non-response occurs when an individual (or household) has failed to provide an interview every wave.

In the HILDA Survey, imputation is used to complete the missing data for key variables resulting from person- and household-level item non-response. In addition, person-level wave and unit non-response in a household where at least one other person provided an interview is corrected for by imputation of key variables. Household-level wave and unit non-response is corrected for through the survey weighting process.

Table 4 below shows the number of responding persons, enumerated adults and responding households in each wave of the survey. Responding persons are individuals that have completed a personal questionnaire for that wave. Enumerated persons are defined as all individuals who belong to a responding household (which include responding persons, non-responding adults, and children). A responding household is where an individual from the household has completed the household questionnaire and a personal questionnaire. The person level totals in Table 4 exclude children under the age of 15 as they are not required to complete a questionnaire.

Table 4: Number of cases, waves 1 to 7

	Wave						
Variable	1	2	3	4	5	6	7
Responding persons	13,969	13,041	12,728	12,408	12,759	12,905	12,789
Enumerated persons (excl. children)	15,127	14,019	13,601	13,321	13,571	13,698	13,589
Responding households	7,682	7,245	7,096	6,987	7,125	7,139	7,063

The extent of missingness for each imputed variable within the responding person, enumerated person, and responding household groups is outlined below. Both the number and proportion of missingness is provided to give a more detailed picture of the size of the problem.

Persons

Each table below shows the number or proportion of missing values that require imputation for each wave, split by responding and enumerated person groups.

Income

Total financial year income is not imputed directly, but all required components contributing to the total are imputed where necessary. The figures reported for total income highlights the overall extent of missing income data by showing the number of individuals with some component that is missing.

Table 5: Number of cases with missing person-level income data, waves 1 to 7

	Wave						
Variable	1	2	3	4	5	6	7
Responding Persons							
Current income (per week)							
Wages and salaries (main job)	357	228	205	193	177	168	195
Wages and salaries (other jobs)	114	89	84	83	86	67	66
Benefits	136	80	74	66	59	36	56
Financial year income							
Wages and salaries	666	550	434	291	362	381	415
Aust govt pensions	97	95	56	80	58	39	40
Foreign govt pensions	1	6	0	1	5	1	2
Business income	404	366	354	242	270	220	225
Investments							
Interest income	661	596	424	330	355	423	410
Dividends and royalties	584	521	402	291	328	355	353
Rent income	239	180	181	130	130	132	134
Private pensions	59	41	29	35	44	35	40
Private transfers	32	89	72	60	107	58	75
Total FY income	2,071	1,841	1,464	1,130	1,295	1,269	1261
Windfall income							
Windfall income	32	31	39	31	25	53	37
Enumerated Persons (excluding children)							
Current income (per week)							
Wages and salaries (main job)	1,514	1,206	1,078	1,106	989	961	995
Wages and salaries (other jobs)	1,267	1,067	957	996	898	860	866
Benefits	1,294	1,058	947	979	871	829	856
Financial year income							
Wages and salaries	1,824	1,528	1,307	1,204	1,174	1,174	1,215
Aust govt pensions	1,255	1,073	929	993	870	832	840
Foreign govt pensions	1,159	984	873	914	817	794	802
Business income	1,562	1,344	1,227	1,155	1,082	1,013	1,025
Investments							
Interest income	1,819	1,574	1,297	1,243	1,167	1,216	1,210
Dividends and Royalties	1,742	1,499	1,275	1,204	1,140	1,148	1,153
Rent income	1,398	1,158	1,054	1,043	942	925	934
Private pensions	1,217	1,019	902	948	856	828	840
Private transfers	1,190	1,067	945	973	919	851	875
Total FY income	3,230	2,819	2,337	2,043	2,107	2,062	2,061
Windfall income							
Windfall income	1,190	1,009	912	944	837	846	837

Table 6: Proportion of cases with missing person-level income data, waves 1 to 7

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Responding Persons (non-zero cases only)							
Current income (per week)							
Wages and salaries (main job)	4.6	3.1	2.8	2.7	2.4	2.2	2.6
Wages and salaries (other jobs)	15.9	13.9	13.2	13.0	12.9	11.1	10.9
Aust govt pensions	3.2	2.0	2.0	1.8	1.6	1.0	1.6
Financial year income							
Wages and salaries	7.9	6.9	5.5	3.8	4.5	4.6	5.1
Aust govt pensions	2.1	2.1	1.3	2.0	1.4	1.0	1.0
Foreign govt pensions	0.5	2.7	0.0	0.5	2.4	0.5	1.0
Business income	29.1	28.6	27.4	19.4	21.7	18.6	19.8
Investments							
Interest income	19.5	18.6	13.9	11.0	11.3	12.8	11.6
Dividends and royalties	14.6	14.5	11.9	9.2	10.2	11.3	11.3
Rent income	20.3	14.7	14.9	11.3	10.5	10.3	10.2
Private pensions	6.3	4.7	3.3	4.1	4.9	3.9	4.1
Private transfers	8.0	23.1	15.8	14.4	21.2	13.4	18.5
Total FY income	15.7	14.9	12.1	9.6	10.7	10.3	10.4
Windfall income							
Windfall income	4.0	2.8	3.2	2.7	2.1	4.6	3.4
Enumerated Persons (zero and non-zero cases, excluding children)							
Current income (per week)							
Wages and salaries (main job)	10.0	8.6	7.9	8.3	7.3	7.0	7.3
Wages and salaries (other jobs)	8.4	7.6	7.0	7.5	6.6	6.3	6.4
Aust govt pensions	8.6	7.5	7.0	7.3	6.4	6.1	6.3
Financial year income							
Wages and salaries	12.1	10.9	9.6	9.0	8.7	8.6	8.9
Aust govt pensions	8.3	7.7	6.8	7.5	6.4	6.1	6.2
Foreign govt pensions	7.7	7.0	6.4	6.9	6.0	5.8	5.9
Business income	10.3	9.6	9.0	8.7	8.0	7.4	7.5
Investments							
Interest income	12.0	11.2	9.5	9.3	8.6	8.9	8.9
Dividends and Royalties	11.5	10.7	9.4	9.0	8.4	8.4	8.5
Rent income	9.2	8.3	7.7	7.8	6.9	6.8	6.9
Private pensions	8.0	7.3	6.6	7.1	6.3	6.0	6.2
Private transfers	7.9	7.6	6.9	7.3	6.8	6.2	6.4
Total FY income	21.4	20.1	17.2	15.3	15.5	15.1	15.2
Windfall income							
Windfall income	7.9	7.2	6.7	7.1	6.2	6.2	6.2

Wealth

Wealth data has been collected in wave 2 and wave 6 of the HILDA Survey. When considering missing data for wealth variables, it is important to separate out individuals that have provided no data at all from those that have not given a value but responded with an approximate band within which their wealth value lies. In wave 2, the only wealth variable to benefit from a wealth band question was superannuation for those not retired. The wave 6 wealth module saw the introduction of eight extra wealth bands (seven in the Household Questionnaire and one in the Person Questionnaire). Most band questions were safety-net type questions that allowed a respondent that had already passed on giving a value (either because they did not know or did not want to provide the value) to choose a band within which that value is likely to fall. The exception was the superannuation bands for person-level wealth, which asked for the band first and the amount second to try and elicit a point estimate for one of the more difficult wealth questions to answer.

The number and proportion of missing wealth values are provided in Table 7 and 8.

Table 7: Number of cases with missing person-level wealth data including and excluding wealth band responses, waves 2 and 6

	Wave 2		*Wave 6*	
Variable	*No point estimate*	*No point estimate or band*	*No point estimate*	*No point estimate or band*
Responding persons (non-zero cases only)				
Joint bank accounts	598	-	348	-
Own bank accounts	396	-	284	-
Superannuation, retirees	135	-	157	89
Superannuation, not retired	1,404	802	2,348	976
HECS debt	110	-	77	-
Joint credit card debt	91	-	58	-
Own credit card debt	77	-	60	-
Other Debt	70	-	56	-
Enumerated persons (zero and non-zero cases)				
Joint bank accounts	1,576	-	1,136	-
Own bank accounts	1,374	-	1,072	-
Superannuation, retirees	1,113	-	945	877
Superannuation, not retired	2,382	1,780	3,136	1,764
HECS debt	1,088	-	865	-
Joint credit card debt	1,069	-	849	-
Own credit card debt	1,055	-	848	-
Other Debt	1,048	-	844	-

Table 8: Proportion of cases with missing person-level wealth data including and excluding wealth band responses, waves 2 and 6

	Wave 2		*Wave 6*	
Variable	*No point estimate*	*No point estimate or band*	*No point estimate*	*No point estimate or band*
Responding persons (non-zero cases only)				
Joint bank accounts	9.8	-	6.0	-
Own bank accounts	4.6	-	3.3	-
Superannuation, retirees	20.1	-	19.7	12.2
Superannuation, not retired	17.3	10.7	27.5	13.6
HECS debt	10.6	-	7.6	-
Joint credit card debt	10.1	-	7.5	-
Own credit card debt	3.6	-	3.1	-
Other Debt	2.4	-	1.8	-
Enumerated persons (zero and non-zero cases)				
Joint bank accounts	11.3	-	8.3	-
Own bank accounts	9.8	-	7.9	-
Superannuation, retirees	8.0	-	6.9	6.5
Superannuation, not retired	17.1	13.3	23.0	14.4
HECS debt	7.8	-	6.4	-
Joint credit card debt	7.7	-	6.2	-
Own credit card debt	7.6	-	6.2	-
Other Debt	7.5	-	6.2	-

Other

In addition to income and wealth variables, any missing data for age was imputed. Though only a small number of cases are missing age, it is a vital variable in the weighting process and the imputation of other variables.

Table 9: Number and proportion of cases with missing age, waves 1 to 7

	Wave						
Variable	1	2	3	4	5	6	7
Enumerated persons							
Number	5	24	42	36	17	17	15
Proportion	0.0	0.1	0.2	0.2	0.1	0.1	0.1

Further, the labour force status was not collected for 979 non-responding individuals belonging to a responding household in wave 2 (this question was not included on the Household Form in wave 2). As this variable is a key variable in both the weighting and the imputation of other variables, it was imputed for wave 2. This imputation was not required for other waves as the information was collected as part of the questionnaire.

Households

Income

Household-level income is calculated by summing across the income components of all the adults in the household. While the household totals are not imputed directly, the number and proportion of households with missing income data have been provided in Table 10 and Table 11.

Table 10: Number of cases with missing household-level income data, waves 1 to 7

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Households (zero and non-zero cases)							
Current income (per week)							
Wages and salaries (main job)	1,092	894	797	849	778	739	783
Wages and salaries (other jobs)	907	784	710	753	709	663	686
Aust govt pensions	928	769	698	741	683	629	676
Financial year income							
Wages and salaries	1,306	1,137	978	908	913	911	947
Aust govt pensions	894	785	682	751	678	635	662
Foreign govt pensions	813	707	632	684	634	599	627
Business income	1,103	966	897	861	832	760	792
Investments							
Interest income	1,298	1,166	963	949	907	925	944
Dividends and royalties	1,244	1,097	938	909	886	882	890
Rent income	974	820	757	773	727	693	721
Private pensions	867	740	658	715	668	626	662
Private transfers	841	783	696	739	716	647	693
Total FY income	2,256	2,028	1,704	1,526	1,586	1,536	1,559
Windfall income							
Windfall income	838	723	661	710	649	645	655

Table 11: Proportion of cases with missing household-level income data, waves 1 to 7

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Households (zero and non-zero cases)							
Current income (per week)							
Wages and salaries (main job)	14.2	12.3	11.2	12.2	10.9	10.4	11.1
Wages and salaries (other jobs)	11.9	10.8	10.0	10.8	10.0	9.3	9.7
Aust govt pensions	12.1	10.6	9.8	10.6	9.6	8.8	9.6
Financial year income							
Wages and salaries	17.0	15.7	13.8	13.0	12.8	12.8	13.4
Aust govt pensions	11.6	10.8	9.6	10.7	9.5	8.9	9.4
Foreign govt pensions	10.6	9.8	8.9	9.8	8.9	8.4	8.9
Business income	14.4	13.3	12.6	12.3	11.7	10.6	11.2

Table 11 (c'td)

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Investments							
Interest income	16.9	16.1	13.6	13.6	12.7	13.0	13.4
Dividends and royalties	16.2	15.1	13.2	13.0	12.4	12.4	12.6
Rent income	12.7	11.3	10.7	11.1	10.2	9.7	10.2
Private pensions	11.3	10.2	9.3	10.2	9.4	8.8	9.4
Private transfers	10.9	10.8	9.8	10.6	10.0	9.1	9.8
Total FY income	29.4	28.0	24.0	21.8	22.3	21.5	22.1
Windfall income							
Windfall income	10.9	10.0	9.3	10.2	9.1	9.0	9.3

Wealth

Wealth data was also collected and imputed at the household-level. As with person-level wealth, the data has been split to show the number of households where the wealth responses were given as either an estimate or within a band in Tables 12 and 13. Wealth data collected in wave 2 at the household-level did not give respondents an option to answer with an approximate wealth band.

Table 12: Number of cases with missing household-level wealth data including and excluding wealth band responses, waves 2 and 6

	Wave 2		*Wave 6*	
Variable	*No point estimate*	*No point estimate or band*	*No point estimate*	*No point estimate or band*
Household wealth items (non-zero cases only)				
Children's bank accounts	85	-	57	-
Business value	200	-	159	63
Cash investments	29	-	22	12
Equity investments	455	-	359	107
Collectibles	150	-	160	79
Other property value	57	-	8	-
Life insurance	200	-	169	86
Trust funds	123	-	101	66
Vehicles: Value	145	-	93	-
Business debt	105	-	37	25
Home Value	386	-	198	-
Home: All debt	133	-	104	-
Other property: Debt	41	-	42	-
Overdue bills: Debt	-	-	15	-
Household totals (zero and non-zero cases)				
Financial Assets	2,633	2,287	2,902	1,760
Non-Financial Assets	793	-	536	379
Total Assets	2,971	2,652	3,126	1,961
Financial Liabilities	1,096	-	881	874
Net Worth	3,117	2,818	3,207	2,098

Table 13: Proportion of cases with missing household-level wealth data including and excluding wealth band responses, waves 2 and 6

Variable	*Wave 2* *No point estimate*	 *No point estimate or band*	*Wave 6* *No point estimate*	 *No point estimate or band*
Household wealth items (non-zero cases only)				
Children's bank accounts	6.2	-	4.6	-
Business value	20.1	-	17.5	7.8
Cash investments	11.6	-	12.3	7.1
Equity investments	15.3	-	13.3	4.4
Collectibles	14.0	-	15.1	8.1
Other property value	4.6	-	0.5	-
Life insurance	24.9	-	28.5	16.9
Trust funds	35.7	-	35.8	26.7
Vehicles: Value	2.3	-	1.5	-
Business debt	22.9	-	11.6	8.1
Home Value	7.8	-	4.6	-
Home: All debt	5.4	-	4.2	-
Other property: Debt	7.1	-	5.9	-
Overdue bills: Debt	-	-	2.2	-
Household totals (zero and non-zero cases)				
Financial Assets	36.3	31.6	40.6	24.7
Non-Financial Assets	10.9	-	7.5	5.3
Total Assets	41.0	36.6	43.8	27.5
Financial Liabilities	15.1	-	12.3	12.2
Net Worth	43.0	38.9	44.9	29.4

Home value is collected every wave and the level of missingness is reported in Table 14.

Table 14: Number and proportion of households with missing home value data, waves 1 to 7

Imputation Method	*Wave* *1*	 *2*	 *3*	 *4*	 *5*	 *6*	 *7*
Home value							
Number	312	378	269	187	157	196	121
Proportion (non-zero cases only)	5.9	7.6	5.6	4.0	3.3	4.2	2.6

Imputation Methods

The imputation methods used in the HILDA Survey, to varying extents, are:

- Nearest Neighbour Regression Method
- Little and Su Method
- Population Carryover Method
- Hotdeck Method

Most of these methods use the concept of donors and recipients. The record with missing information is called the 'recipient' (i.e., it needs to be imputed). The 'donor' has complete information that is used to impute the recipient's missing value. The methods differ in how a suitable donor is identified and used.

Nearest Neighbour Regression Method

The Nearest Neighbour Regression method (also known as predictive mean matching (Little, 1988)) seeks to identify the 'closest' donor to each record that needs to be imputed via the predicted values from a regression model for the variable to be imputed. The donor's reported value for the variable being imputed replaces the missing value of the recipient.

For each wave and for each variable imputed, log-linear regression models using information from the same wave were constructed. A backwards elimination process in SAS was used to identify the key variables for each variable and wave.

The predicted values from the regression model for the variable being imputed are used to identify the nearest case (donor *d*) whose reported value (Y_d) could be inserted into the case with the missing value ($\hat{Y}_i = Y_d$). Donor *d* has the closest predicted value to the respondent *i*, that is $\left|\hat{\mu}_i - \hat{\mu}_d\right| \leq \left|\hat{\mu}_i - \hat{\mu}_p\right|$ for all respondents *p* (potential donors) where $\hat{\mu}_i$ is the predicted mean of *Y* for individual *i* that needs to be imputed, and Y_d is the observed value of *Y* for respondent *d*.

For some variables, an additional restriction may also be applied to ensure that the donor and recipient match on some broad characteristic (such as age group).

Little and Su Method

The imputation method proposed by Little and Su (1989) incorporates (via a multiplicative model) the trend across waves (column effect), the recipient's departure from the trend in the waves where the income component has been reported (row effect), and a residual effect donated from another respondent with complete income information for that component (residual effect). The model is of the form

$$imputation = (roweffect)\,(columneffect)\,(residualeffect)\ .$$

The column (wave) effects are calculated by $c_j = \frac{\bar{Y}_j}{\bar{Y}}$ where $\bar{Y} = \frac{1}{m}\sum_j \bar{Y}_j$ for each wave *j* = 1, …, *m*. $\bar{Y}_j$ is the sample mean of variable *Y* for wave *j*, based on complete cases and *Y* is the global mean of variable *Y* based on complete cases.

The row (person) effects are calculated by $\bar{Y}^{(i)} = \frac{1}{m}\sum_j \frac{Y_{ij}}{c_j}$ for both complete and incomplete cases. Here, the summation is over recorded waves for case *i*; m_i is the number of recorded waves; Y_{ij} is the variable of interest for case *i*, wave *j*; and c_j is the simple wave correction from the column effect.

The cases are ordered by $\bar{Y}^{(i)}$, and incomplete case *i* is matched to the closest complete case, say *d*.

The missing value Y_{ij} is imputed by

$$\hat{Y}_{ij} = \left(\bar{Y}^{(i)}\right)\left(c_j\right)\left(\frac{Y_{dj}}{\bar{Y}^{(d)}c_j}\right) = Y_{dj}\,\frac{\bar{Y}^{(i)}}{\bar{Y}^{(d)}}$$

where the three terms in brackets represent the row, column, and residual effects. The first two terms estimate the predicted mean, and the last term is the stochastic component of the imputation from the matched case. A worked example of the Little and Su method is provided in Appendix 1.

It is important to note that due to the multiplicative nature of the Little and Su method, a zero individual effect will result in a zero imputed value (Starick and Watson, 2007). However, it is quite valid to have an individual reporting zero income in previous waves and then report that they have income but either don’t know its value or refuse to provide it. The individual’s effect would be zero and any imputed amount via the Little and Su method would also be zero, which we know is not true. Therefore, recipients with zero individual effects are not imputed via the Little and Su method. An additional restriction for this method is that donors must have a non-zero row effect to avoid divisions by zero.

Population Carryover Method

A carryover imputation method imputes missing wave data by utilizing responding information for that case from surrounding waves. Rather than randomly assigning either the preceding wave response or the following wave response, the probability of choosing one or the other of these responses is chosen to reflect the changes in the reported amounts between waves observed in the population. This is known as the ‘population carryover method’ (Williams and Bailey, 1996).

The probability that a value is carried forwards or backwards is calculated in the following way. An indicator variable is created which equals 1 when the reported change between waves *j* and *j+1* is smaller than the reported change between waves *j* and *j-1* for the complete cases; and 0 otherwise. The proportion *p* of the interviewed sample where the change between waves *j* and *j+1* is smaller than the change between waves *j* and *j-1* is then determined. The next value is carried backwards with probability *p* and the last value is carried forwards with probability *1-p*, reflecting the probabilities associated with the occurrence of change between waves found in the complete cases.

Within the context of the HILDA Survey, the Population Carryover method is only used for the identification of zero or non-zero amounts. Where the value is deemed to be non-zero, another imputation is used to impute a non-zero amount.

Hotdeck Method

The hotdeck method randomly matches suitable donors to recipients within imputation classes. The donor's reported value for the variable being imputed replaces the missing value of the recipient.

A number of categorical variables are used to define imputation classes for the variable to be imputed. These variables are assigned an order of priority and when there are not a sufficient number of donors within a class, the imputation classes are sequentially folded back, removing the least important class variable first until a suitable donor is found. When more than one donor can be matched to a recipient *i* within an imputation class *c*, a donor *d* is selected randomly (the class of the donor and the recipient are the same, that is, $c_i = c_d$). The donor's reported value is inserted into the recipient's missing value $\hat{Y}_i = Y_d$. A hotdeck macro (hesimput), written by the Statistical Services Branch of the Australian Bureau of Statistics, was used to run this method for the HILDA Survey.

Income Imputation

The final combination of imputation methods used in the imputation of income was established from the imputation evaluation research study by Starick and Watson (2007). The imputation steps for each income variable are as follows:

1. Carryover zeros: For non-responding persons (in responding households) the population carryover method is used to determine whether the income amount is zero or non-zero prior to any other imputation.
2. Nearest Neighbour Regression imputation: The Nearest Neighbour Regression method (with or without imputation classes) is used to identify donors and impute a value for each income variable for all respondents. For non-respondents, a single donor is identified via the Nearest Neighbour Regression method based on total income only, and all their income components are imputed from the single donor. Zero's imputed for non-respondents in step 1 are not replaced with the imputed values produced in this step and non-zero amounts are imputed for those variables determined to be non-zero in step 1.
3. Little and Su imputation: The Little and Su imputation procedure (with or without imputation classes) is run on all records. Results from the Nearest Neighbour Regression method imputes in step 2 are included as an input in the Little and Su method when calculating a records row and column effects. Where possible all step 2 imputes are replaced. Zero's imputed in step 1 are not overwritten with Little and Su imputes and non-zero amounts are imputed for those determined to be non-zero in step 1.

Step 1: Carryover Zeros

The proportion of zeros imputed for non-respondents via the Population Carryover method for each income variable is shown in Table 15. The table gives an indication of how likely it was that a non-respondent gave a zero response in an abutting wave of the survey. Wave 1 and 7 have a smaller proportion of zeros imputed as both waves have only a single abutting wave to carryover income zeros from.

This step in the imputation did not impute all the zeros possible for non-respondents. In steps 2 and 3 the non-respondent who did not have a zero/non-zero determination from the Population Carryover method could have a zero imputed via the Nearest Neighbour Regression or Little and Su methods.

Step 2: Nearest Neighbour Regression Imputation

The Nearest Neighbour Regression method can be applied so that every record requiring imputation for each variable gets imputed. Both the Population Carryover method used in step 1 and the Little and Su method in step 3 have limitations that restrict them from being able to impute every record. In situations where the other methods are not suitable the Nearest Neighbour Regression method result is used.

For each variable imputed each wave, log-linear regression models were constructed. Over 30 variables were considered for inclusion in the income models covering

Table 15: Proportion of non-respondents with zeros imputed via the population carryover method, waves 1 to 7

	Wave						
Variable	1	2	3	4	5	6	7
Current income							
Wages and salaries – main job	6.6	13.8	11.2	12.4	11.7	10.5	6.8
Wages and salaries – other jobs	17.5	34.7	29.4	33.5	26.4	22.4	17.5
Benefits	15.7	27.3	23.3	27.8	19.5	16.5	13.0
Financial year income							
Wages and salaries	5.3	12.6	9.0	10.0	10.7	9.2	6.1
Australian govt pensions	14.6	26.9	22.2	27.6	19.0	15.9	12.8
Foreign govt pensions	19.2	36.3	30.1	34.7	27.0	23.0	17.6
Business income	16.8	32.9	27.7	31.4	24.4	21.1	15.6
Investments							
Interest income	16.1	30.4	24.9	28.8	22.5	18.9	14.8
Dividends and royalties	14.6	28.0	24.5	26.8	22.5	20.6	15.0
Rent income	17.4	33.9	27.8	31.9	24.5	21.6	16.0
Private pensions	18.6	35.1	28.9	32.9	25.6	21.7	16.8
Private transfers	18.7	36.3	30.0	34.5	26.4	22.2	17.0
Windfall income	17.1	32.2	26.9	30.1	25.2	20.4	16.3

demographic characteristics, employment characteristics, the respondent's partner's characteristics (if the respondent had a partner), and the respondent's partner's income. The variables included in each regression model are listed in Appendix 2. A stepwise elimination process in SAS was used to identify the key variables in the model for each variable and wave.

Table 16 presents the number of separate models constructed for each income variable, along with the variable groups that defined these different models. For instance, financial year wages and salaries had four regression models constructed:

i) individuals who provided current wages and salaries and their household income band was reported (in the Household Questionnaire);

ii) individuals who did not provide current wages and salaries but their household income band was reported;

iii) individuals who provided current wages and salaries but their household income band was not reported;

iv) individuals who did not provide current wages and salaries and their household income band was also not reported.

For respondents, any missing income was imputed separately for each variable. For non-respondents, donors were identified utilizing total income only and the income components were all taken from a single donor to ensure the components were consistent with each other.

Table 16: Income Nearest Neighbour regression models

Variable	*Number of models*	*Model groups (based on availability of each item)*
Current income		
Wages and salaries – main job	4	Financial year main job wages and salaries income (available or unavailable) by household income band (available or unavailable)
Wages and salaries – other jobs	4	Financial year wages and salaries income from other jobs by household income band
Benefits	4	Financial year benefit income by household income band
Financial year income		
Wages and salaries	4	Current wages and salaries income by household income band
Australian govt pensions	4	Current benefit income by household income band
Foreign govt pensions	2	Household income band
Business income	4	Partner business income by household income band
Investments		
Interest income	4	Partner interest income by household income band
Dividends and royalties	4	Partner dividends and royalties income by household income band
Rent income	4	Partner rental income by household income band
Private pensions	2	Household income band
Private transfers	2	Household income band
Windfall income	2	Household income band
Total income	2	Household income band

Each complete record was restricted to being used as a donor twice in the Nearest Neighbour Regression procedure. This limitation avoided the possibility of large or unusual values from being imputed too often.

Imputation Classes

For wages and salaries, government pensions and rental income, an additional restriction that the donor and recipient fall within the same age class (15-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65+) was applied. For interest income, dividends and royalties, windfall income, private or foreign pensions, and private transfers, the age classes the donors and recipients were matched within were (15-24, 25-54, 55+). No age class restrictions were applied for business income. Total income for non-respondents had the more detailed age class restrictions applied.

Step 3: Little and Su Imputation

The Little and Su imputation method has the largest influence on the final imputed income values. Wherever possible the Little and Su method is used instead of the Nearest Neighbour Regression method.

When calculating the row and column effect of a record requiring imputation in the Little and Su process any Nearest Neighbour Regression imputed values were used. In some situations a record to be imputed may only have one wave of non-zero reported data. If only that single wave was used to determine their Little and Su ‘effect’ it could result in the selection of an unsuitable donor if that individual’s situation changes in other waves. The Nearest Neighbour Regression imputes establish a suitable value based on their particular circumstances each wave so gives a better initial view of the record over time. Using the overall Little and Su imputes for all waves to be imputed ensures a more coherent longitudinal imputation.

Table 17 presents the proportion of income imputed by each imputation method. For responding persons, the Nearest Neighbour Regression impute is only used when no other waves of data is available. This occurred more in the end waves due to a larger attrition rate between waves 1 and 2 and new entrants in wave 7 that have not yet had a chance to respond again. Enumerated persons have a much lower rate of imputation from the Little and Su method as many are non-respondents that did not appear in another wave. Zeros from the Population Carryover method were also not overwritten by the Nearest Neighbour Regression or Little and Su results.

Each donor in the Little and Su method was restricted to being used twice for a particular income item to avoid it overly influencing the final results

Table 17: Proportion of missing cases imputed by imputation method (income), waves 1 to 7

	Wave						
Imputation Method	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Responding Persons							
Nearest Neighbour	13.4	4.0	5.3	4.6	4.3	4.7	6.3
Little and Su	86.6	96.0	94.7	95.4	95.7	95.3	93.7
Enumerated Persons							
Carryover	12.5	23.9	20.1	24.1	18.4	15.8	11.9
Nearest Neighbour	53.9	36.8	39.5	39.0	39.9	42.5	47.7
Little and Su	33.6	39.3	40.3	36.9	41.7	41.7	40.4

Imputation Classes

Imputation classes were applied to wages and salaries and government pension income for the Little and Su method. Donors and recipients were matched within longitudinal imputation classes defined by the following age ranges in the latest wave: 15-19, 20-24, 25-34, 35-44, 45-54, 55-64, 65+. The column and row effects are calculated within each imputation class and donors are matched to recipients which share the same imputation class.

Quality of Imputation

A large range of measures and evaluations can be undertaken to assess the quality of imputation. Prior to producing the imputation on the main dataset for HILDA Release 6, the evaluation research work undertaken by Starick and Watson (2007) tested a large set of imputation methods. Their work assessed the outputs from the imputation methods across a range of criteria through a simulation study of income using HILDA data. While an imputation method may not be the ‘best’ available for all applications, their results do provide reassurance that the methods we have adopted are performing well.

The individuals that do not provide some income item or do not provide an interview most likely have some systematic differences from the group that answers every question. Excluding these cases from analysis of the HILDA data can negatively affect the representativeness of the results. Table 18 compare the unweighted distribution of the variables pre- and post-imputation for responding persons in wave 1 (Appendix 3 provides similar tables for waves 2 to 7). The imputation has a relatively small impact on most of the income components, but tends to increase the mean total financial income by 1 to 2 per cent. This is most likely because the people with fewer income sources are more likely to provide all of the relevant details than people with a greater number of income sources. As a result they would contribute to the pre-imputation mean and would be likely to contribute a slightly lesser amount.

Table 19 shows the amount that imputation contributes to wages and salaries income and total income. For households and enumerated persons there is a slight decrease over time in the proportion of the mean that is imputed because of the smaller amount of missing data in the later waves.

Table 18: Wave 1 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	698	600	549	694	600	550
Wages and salaries (other jobs)	205	138	218	207	138	232
Benefits	165	169	79	164	169	80
Financial year income						
Wages and salaries	35,222	30,000	38,045	34,428	29,500	37,560
Aust govt pensions	7,484	8,268	4,085	7,463	8,228	4,097
Foreign govt pensions	22,733	15,000	34,507	20,801	13,000	30,992
Business income	2,787	675	7,807	2,727	613	7,581
Investments	2,224	200	8,434	2,320	200	8,689
Interest income	9,901	4,500	31,232	8,784	4,200	27,177
Dividends and royalties	4,516	3,470	3,719	4,506	3,438	3,711
Rent income	14,212	11,000	13,872	13,989	10,400	13,793
Private pensions	4,774	3,215	5,583	4,702	3,120	5,515
Private transfers	4,195	600	15,660	4,457	700	15,196
Total FY income	29,032	21,000	31,719	29,629	21,054	36,500
Windfall income						
Windfall income	7,554	1,100	22,641	7,584	1,040	22,625

Table 19: Mean financial year income ($) (including imputed values) and proportion of mean income ($) imputed, waves 1 to 7 (weighted)

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Responding persons							
Wages and salaries							
Mean	20,955	21,489	22,145	23,119	24,648	26,607	28,840
Proportion imputed	5.9	4.3	3.6	2.9	3.2	3.3	3.1
Total income							
Mean	27,619	28,730	29,456	31,043	33,111	35,829	38,169
Proportion imputed	7.5	6.6	5.5	4.5	4.6	4.6	4.7
Enumerated persons							
Wages and salaries							
Mean	20,954	21,692	22,471	23,292	24,893	26,704	28,862
Proportion imputed	14.6	15.0	14.6	13.8	12.7	11.8	11.8
Total income							
Mean	27,665	28,924	29,802	31,355	33,510	36,013	38,368
Proportion imputed	15.6	16.3	15.7	15.0	14.3	13.0	13.6
Households							
Wages and salaries							
Mean	42,116	43,477	45,106	46,881	50,052	53,641	58,018
Proportion imputed	14.6	15.0	14.6	13.8	12.7	11.8	11.8
Total household income							
Mean	55,606	57,974	59,820	63,109	67,378	72,339	77,125
Proportion imputed	15.6	16.3	15.7	15.0	14.3	13.0	13.6

Wealth Imputation

The wave 2 wealth imputation for Release 2 was produced by the Reserve Bank of Australia using the Nearest Neighbor Regression imputation method (see Watson, 2004). These imputes continued to be used for wave 2 in Release 3 through 5. In wave 6, the HILDA Survey gained a second wave of wealth data to compliment the wealth module conducted in wave 2. With two waves of data available, longitudinal imputation was possible and the imputation process has been adjusted to incorporate this new benefit.

In addition to items collected in the 4-yearly wealth modules, it was decided to impute home value as it is collected in each wave of the survey and is an important data item.

Wealth data involves longitudinal imputation at both the person- and household- level. At the person-level, longitudinal imputation is analogous to income imputation but at the household-level there are three additional difficulties.

First, as the HILDA Survey does not define households over time through a common identifier, these households need to be linked for any longitudinal imputation to be performed at the household-level.

Second, in many situations it is not clear as to whether or not the individual or household actually has a non-zero amount for the asset or debt. For instance, screening questions determine if an individual had a bank account but that does not imply they have money in the account and hence a missing value could validly be imputed as zero.

Third, it is important to separate out individuals that have provided no data at all from those that have not given a point estimate but responded with an approximate band within which their wealth value lies. Using wealth bands in the questionnaire improves the accuracy of the imputation and can elicit responses from some individuals who may not be willing to provide a precise answer (or may not know). Wealth bands are treated as fixed imputation classes (an imputed value has to lie within the provided wealth band) in all stages of the wealth imputation.

The overall imputation steps for wealth:

1. Create a longitudinal household identifier (household imputation only).
2. Run the Nearest Neighbour Regression imputation process to identify persons and households where zero is a sensible impute (essentially a filter process deciding if the record has the asset or liability).
3. Impute all person- and household-level wealth components via the Nearest Neighbour Regression method for records that haven’t been allocated zero in step 2. Apply appropriate imputation classes, wealth bands and filter variables for groups that have a markedly different distribution than general records.
4. Run the Little and Su imputation process on person- and household-level wealth records.

Step 1: Identifying Longitudinal Households

A longitudinal household identifier was created that linked households in wave 2 to households in wave 6. Households were linked based on how the individuals in their household moved between waves. If a household in wave 2 had common household members with a wave 6 household and any additional household members were children, and/or any missing household members were either children or deceased, then a link was made. An individual under the age of 18 was considered a child for the purposes of linking households. A split or merger of household members across waves resulted in no linking as this was considered to have an unknown effect on household wealth. Of the 7245 wave 2 households in the full dataset, 4306 (or approximately 60%) were linked with a wave 6 household. Unlinked households are unable to be imputed via the Little and Su method and receive an imputed value from the Nearest Neighbour Regression method.

The proportion of households longitudinally linked for home value, across all waves, is presented in Table 20. The *diagonal top half* of the table presents the proportion of linked households across all waves from the start to end wave relative to all households in the *start* wave. The *diagonal bottom half* of the table presents proportions relative to the *end* wave. The proportions tend to be larger for the bottom diagonal as the number of households at later waves is generally smaller. A higher proportion of households are linked when only a gap of one wave is involved.

Table 20: Proportion of linked household for home value imputation

		End Wave						
		1	2	3	4	5	6	7
Start Wave	1	.	77%	62%	53%	48%	41%	37%
	2	82%	.	78%	64%	56%	48%	43%
	3	67%	80%	.	77%	65%	54%	48%
	4	59%	66%	78%	.	77%	62%	54%
	5	51%	57%	65%	75%	.	75%	62%
	6	44%	49%	54%	61%	75%	.	80%
	7	40%	44%	48%	53%	63%	81%	.

A household reference person was identified and used to introduce person-level characteristics into the Nearest Neighbour Regression model for each household. The household reference person was established based on the following prioritised criteria (based on the Australian Bureau of Statistics definition of a household reference person[1]):

- a responding individual;
- a member of a couple or a lone parent;
- the highest income earner;

[1] *Standards for Statistics on the Family*, ABS Catalogue No. 1286.0, p. 16. We have, however, added a further requirement for the HILDA Survey that the household reference person be a responding individual.

- the owner of the home;
- the oldest person.

Approximately 17% of the linked household had a different household reference person in each wave. For these households the household reference person specific to each wave was used in the Nearest Neighbour Regression modeling. As the Nearest Neighbour Regression method implemented is a cross-sectional method, it was appropriate to use the most relevant reference person each wave.

Step 2 and Step 3: Nearest Neighbour Regression Imputation

The Nearest Neighbour Regression imputation process was applied to both the person- and household-level data. Log-linear regression models were produced for each wealth variable in each wave and included both household- and person-level characteristics. For household wealth imputation, the person-level characteristics were those associated with their household reference person for each wave. As with income imputation, a backwards stepwise elimination process in SAS was used to identify the key variables for each wealth item in each wave. The variables initially included in each regression model are listed in Appendix 4. Age groups of 15-19, 20-24, 25-34, 35-44, 45-54, 55-63, 65+ were used as imputation classes.

The wealth imputation also incorporated information from the screening questions and ensured that any imputed amount was within provided wealth bands.

Screening questions

Most wealth variables have screening questions to determine whether or not an individual or household has the asset or debt. Due to the nature of some of these variables, knowing they have the asset or debt in question can be used to restrict the imputed amounts to non-zero amounts. Respondents stating that they do not have the asset or debt have been assigned a zero value before the imputation process begins.

Table 21 shows the wealth variables where information was available to restrict the imputation of some records to non-zero amounts. At the person-level only respondents are included in the table, while all households are included at the household-level. The columns ‘require non-zero imputation’ represents all records to be imputed that we know should receive a non-zero amount due to a screening question. Records that have not answered the screening question can be imputed with any value, including zero.

Business value, trust funds and business debt are all items that, in wave 6, had a question asking for the band their asset/debt fell within. Where a band has been given, only non-zero amounts can be imputed.

Many of the household-level wealth variables (excluding those already mentioned) require all, or nearly all, of their missing values to be imputed with a non-zero amount. For these variables, owning the asset or having the debt implies a non-zero value. The discrepancy between the total and non-zero columns in Table 21 for these variables is due to households that have refused or said they did not know at the screening question.

The wealth variables to be imputed that are not listed in Table 21 did not have a screening question that adequately defined whether or not they have a non-zero wealth value. An example is credit card debt. Owning a credit card, which was asked in a screening question, does not imply having any credit card debt. These assets or debts that can have a zero value are more technically an asset/debt generating item, but for simplicity we will refer to them as assets or debts here.

Table 21: Non-zero restrictions on wealth variables to be imputed

	Wave 2		Wave 6	
	Require imputation	Require non-zero imputation	Require imputation	Require non-zero imputation
Person-level Wealth – Respondents Only				
Superannuation, retirees	135	134	157	154
Superannuation, not retired	1,404	605	2,348	1,377
HECS debt	110	105	77	70
Other debt	70	70	56	448
Household-level Wealth				
Business value	200	0	159	96
Cash investments	29	20	22	15
Equity investments	455	446	359	353
Collectibles	150	126	160	106
Home value	386	383	198	196
Other property value	57	53	8	8
Life insurance	200	191	169	158
Trust funds	123	0	101	35
Vehicle value	145	138	93	87
Business debt	105	0	37	26
Home debt	133	113	104	89
Other property debt	41	35	42	38

When zeros were allowed, given they had the asset, the proportion of zeros is usually much lower than when looking at the entire set of data. Table 22 compares the proportion of zeros for the entire sample against the proportion within the group of people we know to have the asset. Rather than rely on the models in the imputation process to select appropriate number of donors with zero values, the donor pools have been restricted in these situations to those with the asset.

Table 22: Proportion of cases reporting zero value for particular assets or debts

	Wave 2		Wave 6	
	All	Have asset	All	Have asset
Person-level Wealth				
Joint bank accounts	55.4	6.8	56.7	6.6
Own bank accounts	35.4	7.0	33.1	6.3
Joint credit card debt	93.7	71.9	94.5	74.9
Own credit card debt	84.2	60.5	85.5	64.2
Household-level Wealth				
Children's bank accounts	81.9	1.7	83.2	1.7
Business value	88.7	11.4	88.1	9.9
Trust funds	96.9	19.9	97.0	11.9
Business debt	95.1	64.4	95.9	69.8

Selecting Donors

The Nearest Neighbour Regression method incorporated two stages. The first was to determine which cases should be imputed with zero or non-zero amounts (i.e., whether the case had the asset or debt in question). Only the zero amounts from this stage were retained. The second stage determined the non-zero amounts to be imputed for those cases deemed to have non-zero amounts from the first stage.

As a result, the donors were selected in two stages and the regression models were created from different pools of data. The zero selection stage allowed all records to be included while the next stage restricted the cases to a subset of cases with non-zero wealth values.

Step 4: Little and Su Imputation

Applying the Little and Su imputation method with only two waves of wealth data initially caused some problems. The correlations between wave 2 and wave 6, when at least one wave had been imputed, were much higher than the raw reported data (when looking at non-zero data in both waves). The suspected cause of this was the trend adjustment applied to each donor's value in the last stage of the Little and Su process. With only two waves being imputed, the trend of the recipient is calculated on only a single data point and it is also more likely that an imputed amount is close to the reported value in the previous wave. To correct this problem, the Little and Su method was adjusted to calculate the row and column effect of a donor for the wave where the recipient has data available. Home value was imputed across all 7 waves of the survey and did not experience the same initial correlation problem as wealth variables imputed in wave 2 and wave 6 only.

The proportion of missing cases imputed by each imputation method is shown in Table 23. Only a subset of households are linked in the dataset and as a result wealth imputation at the household-level, when compared to individual-level, has a larger proportion of Nearest Neighbour Regression imputes.

Table 23: Proportion of missing cases imputed by imputation method (wealth), waves 2 and 6

	Wave	
Imputation Method	*2*	*6*
Person level wealth items (responding persons)		
Nearest Neighbour	38.1	40.8
Little and Su	61.9	59.2
Person level wealth items (enumerated persons)		
Nearest Neighbour	73.3	67.7
Little and Su	26.7	32.3
Household level wealth items		
Nearest Neighbour	56.4	62.6
Little and Su	43.6	37.4

As shown in Table 24, home value had a much larger proportion of Little and Su imputes. This variable was imputed at the household-level, as with the other household items, but more households were linked from wave to wave as only single wave steps were involved. Household were be linked between wave 2 and 6 for other household wealth items.

Table 24: Proportion of missing cases imputed by imputation method (home value), waves 1 to 7

	Wave						
Imputation Method	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Home value (households)							
Nearest Neighbour	26.0	5.3	15.2	14.4	14.6	12.8	21.5
Little and Su	74.0	94.7	84.8	85.6	85.4	87.3	78.5
Number imputed	312	378	269	187	157	196	121

Imputation Classes

The Little and Su imputation method for both person- and household-level wealth applied age groups of 15-19, 20-24, 25-34, 35-44, 45-54, 55-64, and 65+ as imputation classes. The imputation classes were for a longitudinal situation and were assigned based on date of birth. The age group 15-19 corresponded to people born between 1988 and 1992, age group 20-24 born between 1983 and 1987 etc.

Quality of Imputation

Wealth data typically has a more skewed distribution than income so any problems associated with the imputation affecting the mean or distribution of the reported data can be more pronounced.

The proportion of the mean imputed for the household wealth item totals are reported in Table 25. Financial assets are the most susceptible to imputation as a very large proportion (nearly 19%) is due to imputation in both waves.

Table 25: Mean wealth value ($) (including imputed values) and proportion of mean value imputed, waves 2 and 6 (weighted)

	Wave	
Variable	*2*	*6*
Households		
Financial assets		
Mean	152,070	218,581
Proportion imputed	18.8	18.5
Non-financial assets		
Mean	315,338	506,207
Proportion imputed	7.9	4.4
Total assets		
Mean	467,401	724,788
Proportion imputed	11.4	8.7
Total liabilities		
Mean	65,466	113,578
Proportion imputed	6.1	6.2
Net worth		
Mean	401,927	611,210
Proportion imputed	12.3	9.1

The proportion of the mean imputed for home value (Table 26) is reasonably low across all waves.

Table 26: Mean home value ($) (including imputed values) and proportion of mean value imputed, waves 1 to 7 (weighted)

	Wave						
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Households							
Home Value							
Mean	179,346	205,986	244,735	271,670	285,896	311,191	329,900
Proportion imputed	6.0	7.2	5.4	4.0	3.6	4.5	2.8

Table 27 and Table 28 below give a detailed view of the before and after imputation distribution of wealth data in the HILDA Survey. Most data items are not greatly affected by imputation.

Table 27: Unweighted distribution of wealth data before and after imputation - Wave 2

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Person-Level Wealth (non-zero cases only)						
Joint bank accounts	9,506	1,584	56,501	9,558	1,750	53,977
Own bank accounts	11,615	1,500	39,988	11,574	1,500	39,118
Superannuation, retirees	166,080	100,000	244,397	168,658	100,000	240,642
Superannuation, not retired	62,223	19,250	121,999	59,875	18,000	121,586
HECS debt	8,428	6,635	9,288	8,405	6,500	9,073
Joint credit card debt	1,570	1,000	1,694	1,588	1,000	1,721
Own credit card debt	2,780	1,600	3,536	2,811	1,650	3,557
Other debt	19,949	8,000	50,491	20,091	8,000	50,664
Household-Level Wealth (non-zero cases only)						
Children’s bank accounts	1,206	350	3,920	1,211	385	3,872
Business value	392,901	100,000	1,144,485	393,722	100,000	1,095,432
Cash investments	76,995	30,000	130,607	78,235	30,000	130,912
Equity investments	90,702	16,000	248,903	95,998	18,000	259,310
Collectibles	25,202	10,000	99,888	26,166	10,000	98,903
Home value	297,290	240,000	255,167	294,112	235,000	255,565
Other property value	282,395	200,000	392,547	283,261	200,000	389,927
Life insurance	46,848	15,000	106,841	52,261	15,000	116,075
Trust funds	143,202	15,000	386,329	179,296	19,000	485,530
Vehicle value	20,945	15,000	57,043	21,009	15,000	56,587
Business debt	131,794	44,000	255,941	128,868	40,000	251,332
Home debt	114,097	90,000	98,328	113,771	90,000	97,939
Other property debt	143,291	110,000	125,281	150,706	110,000	143,117

Note: Home value has been imputed across 7 waves, whereas the remaining variables have been imputed across the two waves when the wealth module was included (wave 2 and 6).

Table 28: Unweighted distribution of wealth data before and after imputation - Wave 6

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Person-Level Wealth (non-zero cases only)						
Joint bank accounts	12,365	2,300	52,074	12,736	2,500	52,112
Own bank accounts	15,749	2,000	52,239	15,973	2,000	52,057
Superannuation, retirees	244,168	132,000	332,374	245,128	130,000	346,071
Superannuation, not retired	90,657	32,000	197,399	83,415	30,000	186,269
HECS debt	11,341	10,000	9,447	11,476	10,000	9,699
Joint credit card debt	2,234	1,500	2,659	2,232	1,500	2,661
Own credit card debt	4,344	2,500	5,833	4,380	2,500	5,840
Other debt	33,231	9,970	99,682	33,704	10,000	102,268

Table 28 (c'td)

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Household-Level Wealth (non-zero cases only)						
Children's bank accounts	1,594	500	3,133	1,601	500	3,123
Business value	530,101	112,500	1,245,677	508,610	103,000	1,183,858
Cash investments	74,077	31,143	111,886	78,267	32,285	117,240
Equity investments	147,107	25,000	431,661	147,477	24,000	465,067
Collectibles	29,078	10,000	137,882	27,347	10,000	130,443
Home value	453,317	370,000	369,847	450,829	370,000	367,961
Other property value	577,457	350,000	1,127,891	576,869	350,000	1,126,695
Life insurance	101,553	20,000	258,486	102,964	25,000	248,626
Trust funds	332,623	60,000	905,320	361,008	70,000	881,108
Vehicle value	25,328	16,000	39,521	25,390	16,000	39,462
Business debt	170,078	72,000	288,492	172,152	77,000	283,067
Home debt	172,621	135,000	159,229	172,388	135,000	158,817
Other property debt	253,327	177,000	395,837	271,122	180,000	495,174

Other Imputation

Age

Each wave there is a small number of records that require age to be imputed. A simple Hotdeck imputation method is applied with imputation classes defined by sex, household size, relationship in household, household type, partner age (where applicable) and parent age (where applicable).

The results are manually checked to ensure they are suitable given all information we have on the individual (including data from other waves if available).

If a date of birth is provided at a later wave this is used to overwrite any previous imputation.

Wave 2 Employment Status

The employment status of non-respondents (within responding households) in wave 2 was not collected, though for all other waves it is. This variable is important for the benchmarking and non-response adjustment procedures in the weighting process so it was imputed. Imputation consisted of 2 steps:

1. If the individual responded in wave 3 the response they gave to the labour market activity calendar (which provides their employment status over the 14 to 18 months prior to the date of interview) was used to derive their wave 2 employment status.
2. Remaining records were imputed via a Hotdeck imputation method using the variable categories (in order of importance): age group, wave 1 broad employment status, health status (disabled or not), sex, relationship in household, number of people in household, and state.

Of the 979 non-responding individuals in wave 2, 18 per cent had their broad labour force status derived from the wave 3 calendar. The remaining 82 per cent were imputed via the Hotdeck imputation method.

Concluding Remarks

This paper has documented the current state of play for the imputation methods adopted in the HILDA Survey. The imputation extends to income, wealth, age and wave 2 labour force status variables. From Release 8, this list will also include household expenditure which is primarily collected in the Self Completion Questionnaire (a subsequent technical paper will describe how the existing suite of imputation methods have been applied to these variables).

Users of the HILDA data should be aware that the imputed values can change from one release to the next as more longitudinal data becomes available and are used in the longitudinal imputation methods. The HILDA team will also be exploring and evaluating new imputation methods to ensure the most appropriate methods are used. Any changes to the methods will be documented in the latest HILDA User Manual (available on the HILDA website www.melbourneinstitute.com/hilda/) or subsequent technical papers.

References

Little, R.J.A. (1988), ‘Missing Data Adjustments in Large Surveys’, *Journal of Business and Economic Statistics*, 6, 287-296.

Little, R.J.A., and Su, H.L. (1989), ‘Item Non-response in Panel Surveys’, in *Panel Surveys*, ed. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh, New York: Wiley.

Starick, R, and Watson, N (2007), ‘Evaluation of Alternative Income Imputation Methods for the HILDA Survey’, HILDA Project Discussion Paper Series No. 1/07, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N, and Wooden, M (2002), ‘Assessing the Quality of the HILDA Survey Wave 1 Data’, HILDA Project Technical Paper Series No. 4/02, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N, and Wooden, M (2003), ‘Towards an Imputation Strategy for Wave 1 of the HILDA Survey’, HILDA Project Discussion Paper Series No. 1/03, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N (2004), ‘Income and Wealth Imputation for Waves 1 and 2’, HILDA Project Technical Paper Series No. 3/04, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Watson, N, and Wooden, M (2004), ‘Assessing the Quality of the HILDA Survey Wave 2 Data’, HILDA Project Discussion Paper Series No. 5/04, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Williams, T.R., and Bailey, L. (1996), ‘Compensating for Missing Wave Data in the Survey of Income and Program Participation (SIPP)’, Proceedings of the Survey Research Methods Section, American Statistical Association, 305-310.

Appendix 1: Worked example of Little and Su method

This example was included as an appendix to the HILDA User Manual for Release 3 to 7 and was prepared by Rosslyn Starick.

Suppose we have the following small sample of fictitious responses to current wages and salaries.

All cases

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

From this example, we see that observation 1 did not respond to the current wages and salaries question in wave 1, but provided responses in subsequent waves. Observations 5 and 6 also partially responded and wages and salaries information are not provided in all 3 waves.

The first step in the Little and Su method is to calculate the column effects based on complete cases only. Complete cases were defined as individuals that were interviewed in all 3 waves and responded in all 3 waves for the variable of interest. In this example, the complete cases are:

Complete cases

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
2	675	235	700
3	345	690	800
4	200	480	210
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

The column effects are calculated to be:

Column effects

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200
	0.70	**1.06**	**1.24**

The Little and Su method incorporates trend information into the imputed amounts via the column effects. In this example, the wave 1 column effect of 0.70 indicates that the mean current wages and salaries in wave 1 is 30% lower than the overall mean current wages and salaries, and the means in waves 2 and 3 are 6% and 24% higher than the overall mean, respectively.

Next, the row effects are calculated to be:

Row effects

OBS	Wages & Salaries			
	Wave 1	Wave 2	Wave 3	
1		400	420	**357**
2	675	235	700	**585**
3	345	690	800	**596**
4	200	480	210	**303**
5	200			**287**
6	350	370		**425**
7	400	450	470	**459**
8	0	790	790	**460**
9	360	450	600	**475**
10	135	130	200	**159**
	0.70	**1.06**	**1.24**	

The sample is then ordered by the row effects, and the closest donor is identified.

Sorted by row effects

OBS	Wages & Salaries			
	Wave 1	Wave 2	Wave 3	
10	135	130	200	**159**
5	200			**287**
4	200	480	210	**303**
1		400	420	**357**
6	350	370		**425**
7	400	450	470	**459**
8	0	790	790	**460**
9	360	450	600	**475**
2	675	235	700	**585**
3	345	690	800	**596**

Once the closest donor has been identified, the missing value is imputed by multiplying the actual value for the variable of interest of the donor with the row effect of the recipient divided by the row effect of the donor.

In this example, the imputed current wages and salary amounts using the Little and Su method are highlighted below.

Impute missing values

OBS	Wages & Salaries		
	Wave 1	Wave 2	Wave 3
10	135	130	200
5	200	455	199
4	200	480	210
1	236	400	420
6	350	370	436
7	400	450	470
8	0	790	790
9	360	450	600
2	675	235	700
3	345	690	800

Appendix 2: Variables included in the income regression models

Demographic characteristics

Age
Age squared
Sex
Whether of pension age
Highest level of education
Approximate number of years spent in education
Relationship in household
Whether partnered with child
Whether partnered without child
Marital status
Time spent in Australia
First language spoken was other than English
Whether eldest when growing up
Presence of long term health condition

Employment characteristics

Usual hours worked in all jobs
Occupational status
Occupation - 2 digit (present or most recent)
Industry – 2 digit (present or most recent)
Labour force status
Estimate of hours worked in last year
Tenure with current employer
Whether multiple job holder
Contract type
Proportion of last FY spent in employment
Proportion of last FY spent in full-time study
Proportion of last FY spent in part-time study
Proportion of last FY spent not in labour force
Proportion of last FY spent in unemployment

Household characteristics

SEIFA index of educational disadvantage
SEIFA index of economic resources
SEIFA index of disadvantage
Whether renting, purchasing, owning or other
Household income band

Partners characteristics (if available)

Whether have partner
Partner’s age
Partner’s current wages and salaries
Partner’s current benefits
Partner’s FY wages and salaries
Partner’s FY Aust govt pensions and benefits
Partner’s FY foreign govt pensions and benefits
Partner’s FY business income
Partner’s FY interest
Partner’s FY dividends/royalties
Partner’s FY rent
Partner’s FY private pensions
Partner’s FY private transfers
Partner’s FY total income
Partner’s FY windfall

Appendix 3: Distribution of income data before and after imputation, Waves 2 to 7

Table 29: Wave 2 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	710	619	544	705	612	544
Wages and salaries (other jobs)	222	145	259	222	138	321
Benefits	173	178	79	173	178	80
Financial year income						
Wages and salaries	35,884	31,000	33,219	34,951	30,000	33,210
Aust govt pensions	7,806	8,580	4,237	7,791	8,576	4,241
Foreign govt pensions	26,220	16,046	49,140	23,660	15,000	42,434
Business income	2,265	500	6,438	2,202	500	6,136
Investments	3,053	220	12,661	3,234	250	15,283
Interest income	6,459	4,000	10,105	6,474	4,000	10,303
Dividends and royalties	4,841	3,600	4,751	4,844	3,600	4,701
Rent income	16,037	12,000	16,479	15,920	12,000	16,362
Private pensions	4,907	3,600	5,547	4,688	3,347	5,410
Private transfers	4,014	605	12,527	4,562	760	15,820
Total FY income	30,070	21,568	33,474	30,828	22,000	35,489
Windfall income						
Windfall income	20,802	2,000	73,514	20,644	2,000	73,355

Table 30: Wave 3 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	739	650	562	733	645	561
Wages and salaries (other jobs)	230	145	385	218	138	362
Benefits	177	185	84	177	185	84
Financial year income						
Wages and salaries	36,936	32,000	33,313	36,174	31,300	33,080
Aust govt pensions	8,246	9,100	4,435	8,222	9,100	4,451
Foreign govt pensions	26,658	15,000	51,453	24,076	13,030	46,159
Business income	2,189	534	6,941	2,225	500	6,863
Investments	3,503	281	16,294	3,369	260	15,475
Interest income	7,479	4,000	16,613	7,543	4,000	16,774
Dividends and royalties	4,740	3,612	4,845	4,740	3,612	4,845
Rent income	16,607	12,000	16,595	16,450	12,000	16,546
Private pensions	4,712	3,120	6,002	4,497	2,860	5,926
Private transfers	4,477	700	16,882	4,793	750	17,227
Total FY income	31,470	23,000	35,675	32,095	23,440	36,986
Windfall income						
Windfall income	21,630	2,000	76,557	21,164	1,800	75,410

Table 31: Wave 4 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	768	675	581	762	670	581
Wages and salaries (other jobs)	248	150	497	237	141	467
Benefits	187	195	96	187	195	97
Financial year income						
Wages and salaries	38,188	33,500	32,982	37,659	33,000	32,958
Aust govt pensions	8,772	9,700	4,691	8,740	9,636	4,706
Foreign govt pensions	25,215	16,900	33,147	23,584	15,000	30,716
Business income	2,635	600	8,618	2,684	600	9,249
Investments	4,350	400	17,282	4,394	400	18,243
Interest income	9,004	5,000	21,892	8,517	4,936	20,442
Dividends and royalties	3,968	3,300	3,059	3,967	3,300	3,051
Rent income	15,913	11,000	16,310	15,819	11,000	16,248
Private pensions	5,330	3,640	6,406	4,938	3,209	6,146
Private transfers	5,540	1,000	20,329	5,825	1,000	21,765
Total FY income	33,213	25,000	35,018	33,481	25,000	35,547
Windfall income						
Windfall income	18,541	2,000	54,445	18,225	1,800	53,803

Table 32: Wave 5 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	801	700	609	796	700	609
Wages and salaries (other jobs)	237	150	432	226	145	407
Benefits	190	200	89	190	200	90
Financial year income						
Wages and salaries	39,952	35,000	34,954	39,309	34,716	34,942
Aust govt pensions	8,646	9,854	4,923	8,642	9,826	4,940
Foreign govt pensions	26,962	19,000	32,829	25,047	16,000	30,790
Business income	2,677	610	7,947	2,747	601	8,030
Investments	5,742	500	25,794	5,620	499	24,829
Interest income	10,423	5,000	43,071	9,764	5,000	40,282
Dividends and royalties	4,378	3,250	4,618	4,490	3,174	5,137
Rent income	17,476	12,412	19,205	17,182	12,000	19,004
Private pensions	5,115	3,330	6,880	4,510	2,600	6,379
Private transfers	6,731	1,065	31,354	6,968	1,151	30,795
Total FY income	35,268	26,256	39,347	35,605	26,260	39,571
Windfall income						
Windfall income	20,951	1,500	89,633	20,573	1,500	88,715

Table 33: Wave 6 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	848	744	629	845	740	630
Wages and salaries (other jobs)	260	150	469	266	141	582
Benefits	201	210	108	201	209	108
Financial year income						
Wages and salaries	42,452	36,200	37,217	41,751	35,867	37,313
Aust govt pensions	9,154	10,140	5,101	9,154	10,140	5,101
Foreign govt pensions	30,462	20,000	39,779	27,736	19,083	36,675
Business income	3,103	720	9,707	3,036	700	9,551
Investments	6,767	500	25,356	6,619	500	24,552
Interest income	12,311	5,772	44,630	11,698	5,486	42,078
Dividends and royalties	5,268	3,452	6,695	5,255	3,432	6,679
Rent income	19,794	14,077	22,091	19,644	14,000	21,956
Private pensions	5,203	3,600	5,922	4,850	3,120	5,721
Private transfers	7,923	1,200	33,671	8,090	1,200	32,551
Total FY income	38,223	28,660	43,553	38,453	28,900	43,468
Windfall income						
Windfall income	28,458	1,924	177,057	27,826	1,800	173,174

Table 34: Wave 7 unweighted distribution of income data (responding persons) before and after imputation

	Before Imputation			*After Imputation*		
Variable	*Mean*	*Median*	*Standard Deviation*	*Mean*	*Median*	*Standard Deviation*
Responding Persons (non-zero only)						
Current income (per week)						
Wages and salaries (main job)	893	773	661	888	767	663
Wages and salaries (other jobs)	255	175	358	251	175	345
Benefits	209	220	98	208	220	99
Financial year income						
Wages and salaries	45,815	39,500	48,237	44,788	38,000	47,664
Aust govt pensions	9,457	10,450	5,320	9,450	10,441	5,323
Foreign govt pensions	35,679	20,000	72,195	32,585	20,000	65,346
Business income	2,988	750	8,901	2,976	764	8,702
Investments	7,325	590	28,022	7,207	527	27,267
Interest income	9,232	6,000	18,955	8,827	5,500	17,946
Dividends and royalties	5,396	3,370	8,112	5,397	3,440	8,073
Rent income	19,952	14,172	22,084	19,672	14,000	21,911
Private pensions	6,036	3,600	9,824	5,733	3,000	9,531
Private transfers	7,556	1,300	25,765	7,792	1,400	25,885
Total FY income	40,898	30,200	53,235	40,946	30,130	52,799
Windfall income						
Windfall income	21,701	2,000	78,819	21,720	2,000	78,037

Appendix 4: Variables included in wealth regression models

Person-level

Demographics
Sex
Age
Age squared
Speaks English well
Presence of long term health condition
Marital status
Number of children
Would like more children
Indigenous
Highest level of education
Income unit type

Employment
Employment status
Years retired
Years since school
Years worked
Years worked squared
Years unemployed
Prefer to work more
Prefer to work less
Employment contract
% likelihood of losing job
% likelihood of losing job voluntarily
% likelihood to find a job as good as your first
Receive paid holiday with Job
Receive paid sick leave with Job
Non-government job with for profit company
Non-government job with not for profit company
Government job
Less than 20 employees in company
More than 20 employees in company
Occupation
Member of a trade union

Income
FY wages and salaries
FY Australian Government pensions and benefits
FY interest income

Partner wealth
Partner's equivalent wealth component

History/Parents
Parents ever divorced
Has siblings
Family status when 14
Broad country of birth
Father's employment status when 14
Father's occupation when 14
Father unemployed > 6 months
Mother's employment status when 14
Mother's occupation when 14

Household Characteristics
Number of bedrooms
Household tenure
Household boarder
Household ownership shared
Type of dwelling
Household condition
State
Inner, middle, outer city, rural
Remoteness
Number of adults
Number of children
Number employed
Number of males
Number of females
Number who speak English well
Number born overseas
Number with long term health condition
Average adult age
Average child age

Type of Household Assets Owned
Ever owned bonds
Has life insurance
Has trust fund
Owns all of trust fund
Has investment property loan
Owns shares
Has vehicle
Has recreational vehicles
Has other vehicle

Household-level

HRP demographics

Sex
Age
Age squared
Speaks English well
Presence of long term health condition
Marital status
Number of children
Would like more children
Indigenous
Highest level of education
Income unit type

HRP Employment

Employment status
Years retired
Years since school
Years worked
Years worked squared
Years unemployed
Prefer to work more
Prefer to work less
Employment contract
% likelihood of losing job
% likelihood of losing job voluntarily
% likelihood to find a job as good as your first
Receive paid holiday with Job
Receive paid sick leave with Job
Non-government job with for profit company
Non-government job with not for profit company
Government job
Less than 20 employees in company
More than 20 employees in company
Occupation
Member of a trade union

HRP Income

FY wages and salaries
FY Australian Government pensions and benefits
FY interest income

Household Income

Household wages and salaries
Household government income

HRP History/Parents

Parents ever divorced
Has siblings
Family status when 14
Broad country of birth
Father’s employment status when 14
Father’s occupation when 14
Father unemployed > 6 months
Mother’s employment status when 14
Mother’s occupation when 14

Household Characteristics

Number of bedrooms
Household tenure
Household boarder
Household ownership shared
Type of dwelling
Household condition
State
Inner, middle, outer city, rural
Remoteness
Number of adults
Number of children
Number employed
Number of males
Number of females
Number who speak English well
Number born overseas
Number with long term health condition
Average adult age
Average child age

Type of Household Assets Owned

Ever owned bonds
Has life insurance
Has trust fund
Owns all of trust fund
Has investment property loan
Owns shares
Has vehicle
Has recreational vehicles
Has other vehicle

HRP=Household Reference Person