THE UNIVERSITY OF MELBOURNE

HILDA PROJECT TECHNICAL PAPER SERIES
No. 5/04, July 2004

Assessing the Quality of the HILDA Survey Wave 2 Data

Nicole Watson and Mark Wooden

The HILDA Project was initiated, and is funded, by the Australian Government Department of Family and Community Services

MELBOURNE INSTITUTE of Applied Economic and Social Research

ACER

Australian Institute of Family Studies

Contents

INTRODUCTION 1

ATTRITION BIAS 2

CORRELATES OF ATTRITION 2

Wave 2 Sample Distributions by Wave 1 Characteristics 2

Attrition Logits 5

EFFECT OF ATTRITION ON POPULATION ESTIMATES 12

MISSING DATA AND IMPUTATION 15

MISSING INCOME DATA 15

IMPUTED INCOME ESTIMATES 17

MISSING WEALTH DATA 21

IMPUTED WEALTH ESTIMATES 21

OTHER DATA ISSUES 25

QUALITY OF PERSON MATCHES 25

CROSS-FORM MATCHING 25

LONGITUDINAL INCONSISTENCIES 27

DERIVED VARIABLES 29

Tax Derivation 29

Treatment of Owner Managers of Incorporated Businesses 29

Geographic Identifiers 29

Other Issues 30

QUESTIONNAIRE DESIGN PROBLEMS 30

REFERENCES 31

APPENDIX 1: ITEM NON-RESPONSE RATES, WAVES 2 32

Introduction

Following the release of data from the first wave of the Household, Income and Labour Dynamics in Australia Survey, in 2002, a paper was released which discussed the quality of the data (Watson and Wooden 2002). This paper repeats that exercise, but with a focus on the data collected during the second wave.

The paper commences with an analysis of sample attrition and its consequences. We find that the characteristics of those who attrit and those who do not are quite different. Nevertheless, our judgment is that any bias imparted by the selectiveness of attrition is, at this stage at least, likely to be quite small and unlikely to have significant consequences for analyses of most outcome variables.

We then discuss the issue of missing data. As for wave 1, this is only a serious problem for items seeking monetary values (i.e., income and wealth). Unlike the wave 1 release (release 1.0), however, the wave 2 data release (release 2.0) included a set of income and wealth variables which provide imputed values for all missing values. From a cross-section perspective, these imputed estimates seem to be quite sensible. There is, however, much more reason to be concerned about estimates of change based on these imputed data.

The paper also provides a discussion of a range of other data issues including the quality of person matches across waves, the quality of person matches across survey instruments, the longitudinal consistency of the data and problems with the construction of some derived variables included in data release 1.0. For the most part, the data appear to behave as expected. Nevertheless, it is very clear that there are some inconsistencies across waves. Such problems are virtually unavoidable in panel survey collections – conceivably they could be eliminated through much more frequent data collection, but this would be very costly, extremely burdensome on respondents and difficult to administer. Fortunately, with one exception, we do not believe that these inconsistencies will have serious ramifications for most uses of the data. Users interested in the annual activity calendar, however, will inevitably be confronted with the difficulty of how to reconcile inconsistencies that occur at the calendar 'seams'.

Finally, it needs to be recognised that like the data collection process itself, data management is a dynamic activity. There are numerous weaknesses with, and problems in, the publicly released data file that can and will be rectified in future data releases. Indeed, changes between wave 1 and wave 2 in the way some of our derived variables have been constructed (notably, after-tax income variables) attest to this.

Attrition Bias

In our previous paper examining the quality of the wave 1 data, considerable attention was devoted to the issue of sample representativeness. It was concluded that despite achieving a household response rate of only 66 per cent, the extent of bias arising out of non-response in the first wave of the HILDA Survey was likely to be relatively small. Moreover, the sources of bias that appeared to be of greatest importance – differences in rates of response across both sex and location – could be relatively easily corrected through the application of population weights. That said, it does need to be recognised that the design of the HILDA Survey ensures that the sample will become less representative over time. This is because the only way immigrants who arrive in Australia after mid-2001 can join the sample is through joining a household containing a sample member. Fortunately, it will take quite a long time before the magnitude of this effect is large enough to create any significant concern.[1]

Of far greater concern is the potential for bias created by non-random attrition. As outlined elsewhere (Watson and Wooden 2004), attrition between wave 1 and wave 2 was relatively high, with only 86.8 per cent of respondents in wave 1 successfully re-interviewed in wave 2.[2] Nevertheless, high rates of attrition per se are not necessarily a serious problem. Obviously attrition can lead to declining sample size and thus gradually reduce the efficiency of panel data estimates. This problem, however, is largely non-existent in panel studies that employ an infinite life design where the sample is constantly being augmented by new sample members (through the use of predetermined following rules). Far more problematic is attrition that is non-random. If the persons and households that drop out of the panel have characteristics that are systematically different from those who remain, then analyses of these data that do not account in some way for the selective nature of the attrition will most likely lead to biased inferences.[3]

Correlates of Attrition

In this section we consider the question of whether or not the persons who did not respond in wave 2 are systematically different from those who did. We begin by comparing the distribution of the responding wave 2 sample with the non-responding sample on selected characteristics. We then report results from the estimation of binary choice models for the probability of response in wave 2.

Wave 2 Sample Distributions by Wave 1 Characteristics

Tables 1 and 2 show, for selected sample characteristics, measured at wave 1, both the composition of the sample eligible for interview in wave 2 (all wave 1 respondents less deaths and movers overseas) and the attrition rate. Further, we also report the results of a simple non-parametric test for significance of difference in attrition for each characteristic.

[1] For example, exclusion from the population of all immigrants who arrived in Australia in the five years prior to the 2001 Census, has only a very small impact on the basic demographic profile (age and sex).

[2] Note that this rate of attrition still compares favourably with other international panel surveys. For example, in the British Household Panel Survey the proportion of wave 1 respondents who provided interviews in wave 2 was 87.6 per cent (after excluding proxy interviews).

[3] For a formal statistical model of attrition bias, see Fitzgerald et al. (1998).

Table 1: Wave 2 Attrition Rates by Selected Wave 1 Demographic Characteristics

Wave 1 characteristic	*Wave 2 eligible (%)*	*Attrition rate (%)*	*P-value*	*Wave 1 characteristic*	*Wave 2 eligible (%)*	*Attrition rate (%)*	*P-value*
Area			<0.001	No. of adults in household[a]			<0.001
Sydney	17.2	17.6		One adult	16.3	8.7	
Rest of NSW	13.7	11.8		Two adults	54.5	11.3	
Melbourne	18.0	14.6		Three adults	16.2	17.2	
Rest of Victoria	7.7	11.9		Four or more adults	13.0	21.8	
Brisbane	8.4	12.5		No. of children in household[a]			<0.001
Rest of Qld	10.7	10.8		No children	64.2	12.6	
Adelaide	6.3	10.8		One child	14.4	16.7	
Rest of SA	3.1	11.3		Two children	13.7	13.0	
Perth	7.2	11.4		Three or more	7.7	11.9	
Rest of WA	2.7	13.8		Country of birth			<0.001
Tasmania	3.0	14.2		Australia	74.7	11.9	
Northern Territory	0.5	4.6		O/S: English-spkg	10.8	12.9	
ACT	1.7	9.1		O/S: Other	14.4	20.1	
Sex			0.117	Indigenous status			0.002
Male	47.5	13.7		Indigenous	1.9	19.8	
Female	52.5	12.8		Non-indigenous	98.1	13.1	
Age group (years)			<0.001	Education attainment			<0.001
15-19	8.9	20.3		Year 11 or below	36.6	15.1	
20-24	7.1	23.4		Year 12	11.8	16.6	
25-34	18.7	14.2		Certificate	26.1	13.1	
35-44	21.6	11.2		Diploma	8.1	10.5	
45-54	17.4	11.5		Degree or higher	17.3	8.3	
55-64	12.0	10.4		Dwelling type			<0.001
65-74	8.6	8.0		House	82.1	12.8	
75+	5.6	12.8		Semi-detached	8.1	12.8	
Marital status			<0.001	Flat, unit, apartment	9.0	17.7	
Married	54.1	11.4		Other	0.8	10.6	
De facto	9.7	16.5		Index of disadvantage[b]			0.082
Separated	3.1	11.8		Lowest quintile	18.8	12.7	
Divorced	5.4	10.8		2nd lowest quintile	20.0	14.6	
Widowed	4.8	7.9		Middle quintile	20.1	13.2	
Single	22.9	17.9		2nd highest quintile	20.9	13.3	
Relationship in h'hold			<0.001	Highest quintile	20.2	12.1	
Married couple	53.6	11.4					
De facto couple	9.5	16.7					
Lone parent	5.6	14.4					
Child/relative	14.4	20.2					
Lone person	13.9	8.0					
Unrelated	3.0	23.8					

Notes: a An adult is defined here as anyone aged 15 years or over. A child is therefore anyone under the age of 15 years.

b The index of disadvantage used here is the index of relative socio-economic disadvantage constructed by the ABS (ABS cat. no. 2039.0) and is based on the place where people live.

Table 1 presents results for key demographic variables. Specifically, for each characteristic we report the distribution of the population eligible for interview at wave 2 and the attrition rate. Thus, if we consider the area in which people reside (based on the address recorded at the wave 1 interview), we can see that attrition rates vary from just 4.6 per cent in the Northern Territory (though the sample size involved here is quite small) up to 17.6 per cent for residents of Sydney. Further, since Sydney residents represent a sizeable proportion of the sample (17.2 per cent), it suggests that this relatively high rate of attrition cannot be easily ignored.

In general, the most striking feature of Table 1 is how different attritors and non-attritors are on most characteristics, as reflected in the reported P-values. Compared with non-attritors, attritors are more likely to have been, at wave 1: living in Sydney, young, single or alternatively living in a de facto relationship, born overseas but not in one of the main English-speaking countries, an Aboriginal or Torres Strait Islander, without a post-school qualification and living in a flat, unit or apartment. Of the variables considered, only sex and an index of socio-economic disadvantage (based on the CD in which people reside) were not significantly related to attrition.

Table 2 is similar to Table 1, but all of the characteristics listed here are better thought of as outcome variables. Again, attritors and non-attritors varied significantly on most characteristics. Attritors were more likely to have been unemployed when interviewed at wave 1, were more likely to have reported relatively low levels of life satisfaction, were less likely to own or be buying their own home and, as would be expected, were much more likely to have changed address between waves. There are also significant differences by income, with attritors more likely to have relatively low levels of personal income, but in

Table 2: Wave 2 Attrition Rates by Selected Wave 1 Outcome Variables

Wave 1 characteristic	*Wave 2 eligible (%)*	*Attrition rate (%)*	*P-value*	*Wave 1 characteristic*	*Wave 2 eligible (%)*	*Attrition rate (%)*	*P-value*
Labour force status			<0.001	Annual household income			<0.001
Employed full-time	41.6	13.3		<$20,000	16.8	11.0	
Employed part-time	19.6	12.6		$20,000 - $39,999	19.9	13.2	
Unemployed	4.4	18.9		$40,000 - $59,999	19.4	14.5	
Not in labour force	34.5	12.7		$60,000 - $79,999	16.0	13.8	
Life satisfaction			0.001	$80,000 - $119,999	17.6	11.8	
Low (0-4)	3.3	17.8		$120,000+	10.3	15.9	
Medium (5-7)	28.4	14.0		Annual personal income			<0.001
High (8-10)	68.3	12.6		Negative/nil	4.6	15.8	
Household tenure			<0.001	$1-$9,999	24.4	15.1	
Own/purchasing	72.4	11.8		$10,000 – $14,999	12.7	11.0	
Rent	25.2	17.9		$15,000 - $29,999	21.4	14.3	
Other	2.5	7.8		$30,000 - $59,999	26.4	12.4	
Benefit recipient status			0.907	$60,000+	10.4	10.0	
Benefit recipient	33.1	13.2		Moved b/w wave 1 and 2			<0.001
Not benefit recipient	66.9	13.2		Moved	17.6	21.4	
				Did not move	82.4	11.4	

Note: The income variables are for the financial year preceding the wave 1 interview (2000-01) and include imputed values for missing cases.

general were not from low-income households. This latter finding no doubt reflects, at least in part, the difficulty securing interviews with teenagers and adult children in many households. The only outcome variable where there was no marked difference between attritors and non-attritors was benefit status. Indeed, the rate of attrition among recipients of government benefits and pensions was identical to that of non-recipients. That said, there are almost certainly differences depending on the type of benefit claimed. Most obviously, the higher rate of attrition among the unemployed points to relatively high rates of attrition among unemployment benefit recipients.

Attrition Logits

We now turn to a consideration of the determinants of attrition within a multivariate framework. Specifically, we estimated logit equations for the probability of response at wave 2. We also estimated a two-part model which distinguished between the two major stages of response – making contact and obtaining an interview.

The results from the single-equation model predicting response are provided in Table 3. Three different specifications are reported. Specification 1 only includes basic personal and demographic characteristics. The list of characteristics is the same as those considered in Table 1 and all are specified in binary form, except age and the number of adults and children in the household, which are specified as continuous variables. Specification 2 augments this equation with the outcome variables considered in Table 2, but with labour force status interacted with hours of work for those in paid employment. Finally, specification 3 adds to this model an array of variables describing the interview situation. With one exception, all explanatory variables are derived from data collected at wave 1. The exception is the mobility variable, which is based on the observed movement of sample members between waves 1 and 2.[4]

Looking first at specification 1, it is immediately apparent that the probability of response at wave 2 does vary significantly with numerous individual and household characteristics. In particular, it rises with age (but at a declining rate) and educational attainment, falls with the number of adults in the household, is relatively low among indigenous Australians and persons born overseas and in households living in flats, units or apartments, and varies with place of residence. Further, the magnitudes of the effect of many of these variables are arguably quite large.[5] The value of the pseudo R-squared term, however, is very low which, while only providing a crude measure of goodness of fit, suggests that most of the variation in attrition probabilities is either due to other factors or is random. This is a positive finding, implying that weights based on this equation would not greatly affect analyses of outcomes (Fitzgerald et al. 1998, p. 276).

[4] We have assumed that all households lost in tracking have moved. It is possible, however, that contact might not have been made with a household and then that household subsequently be deemed untraceable, when in fact it had actually not changed address. We believe that the number of such cases is likely to be very few, especially given the wide range of tracking measures employed.

[5] A guide to the size of these effects is provided by looking at the estimated odds ratio for each variable, which is simply the inverse log of the coefficient. Thus if we take the variable ‘Other NSW’, the coefficient of 0.312 gives an odds ratio of 1.366, meaning that the odds of people living in New South Wales but not in Sydney at wave 1 responding in wave 2 were 36.5 per cent greater than the odds of persons living in Sydney (the base group) responding.

Table 3: Wave 2 Response, Logit Results (n = 13,817)

	(1)		*(2)*		*(3)*	
	Coeff.	*z*	*Coeff.*	*z*	*Coeff.*	*z*
Constant term	1.136	4.16	1.252	2.81	0.973	2.13
Area of residence (base = Sydney)						
Other NSW	0.312	3.23	0.301	3.07	0.126	1.24
Melbourne	0.205	2.49	0.172	2.07	0.057	0.66
Other Vic	0.271	2.30	0.269	2.26	0.064	0.52
Brisbane	0.288	2.63	0.338	3.05	0.164	1.43
Other Qld	0.485	4.53	0.542	4.97	0.440	3.91
Adelaide	0.548	4.31	0.509	3.97	0.316	2.40
Other SA	0.391	2.29	0.353	2.04	0.187	1.05
Perth	0.411	3.50	0.403	3.40	0.257	2.11
Other WA	0.153	0.92	0.151	0.90	0.007	0.04
Tasmania	0.124	0.78	0.075	0.47	-0.126	-0.76
Northern Territory	1.592	2.65	1.742	2.87	1.713	2.78
ACT	0.597	2.48	0.578	2.39	0.267	1.09
Sex (Female=1)	0.080	1.49	-0.016	-0.27	0.024	0.40
Age	0.055	5.67	0.060	5.89	0.075	7.20
Age squared	-0.0005	-5.02	-0.0006	-5.76	-0.0007	-6.96
Marital status (base = Married)						
De facto	0.676	1.23	0.715	1.27	0.793	1.37
Separated	0.323	0.90	0.352	0.96	0.351	0.93
Divorced	0.312	0.90	0.295	0.84	0.261	0.72
Widowed	0.697	1.94	0.663	1.82	0.660	1.76
Single	0.405	1.19	0.350	1.01	0.336	0.94
Relationship in household (base = Married couple)						
De facto couple	-1.051	-1.90	-0.962	-1.70	-0.990	-1.70
Lone parent	-0.694	-2.09	-0.562	-1.65	-0.396	-1.12
Child / relative	-0.268	-0.77	-0.347	-0.99	-0.308	-0.85
Lone person	-0.352	-1.04	-0.295	-0.79	-0.172	-0.45
Unrelated	-1.016	-2.87	-0.729	-2.02	-0.573	-1.54
Number of adults in HH	-0.287	-9.19	-0.313	-9.21	-0.218	-5.08
Number of children in HH	-0.024	-0.90	-0.034	-1.25	-0.057	-2.02
Country of birth (base = Australia)						
Overseas: Main English-speaking	-0.290	-3.34	-0.261	-2.98	-0.249	-2.78
Overseas: Other	-0.618	-8.67	-0.618	-8.42	-0.342	-3.87
Indigenous	-0.399	-2.39	-0.308	-1.81	-0.169	-0.97
Education (base = Year 11 and below)						
Year 12	0.175	2.12	0.212	2.52	0.212	2.45
Certificate	0.137	2.02	0.195	2.81	0.183	2.57
Diploma	0.371	3.39	0.432	3.87	0.349	3.08
Bachelor or higher	0.732	8.02	0.800	8.43	0.754	7.76
Dwelling type (base = Separate house)						
Semi-detached	-0.091	-0.91	0.022	0.21	0.061	0.57
Unit / apartment / flat	-0.396	-4.29	-0.259	-2.70	-0.214	-2.16
Other dwelling	0.041	0.13	0.194	0.59	0.322	0.96

Table 3 (cont'd)

	(1)		*(2)*		*(3)*	
	Coeff.	*z*	*Coeff.*	*z*	*Coeff.*	*z*
SEIFA disadvantage (base = Lowest quintile)						
2nd lowest quintile	-0.242	-2.91	-0.258	-3.06	-0.259	-3.00
Middle quintile	-0.069	-0.80	-0.091	-1.04	-0.092	-1.02
2nd highest quintile	-0.097	-1.12	-0.131	-1.49	-0.127	-1.41
Highest quintile	-0.007	-0.07	-0.062	-0.66	-0.074	-0.76
Employment and LF status (base = Not in Labour Force)						
Unemployed			-0.149	-1.20	-0.111	-0.87
Employed PT (1-24hrs)			0.145	1.51	0.130	1.31
Employed PT (25-34hrs)			-0.133	-1.05	-0.161	-1.25
Employed FT (35-44hrs)			-0.153	-1.58	-0.157	-1.58
Employed FT (45-54hrs)			0.011	0.09	0.018	0.15
Employed FT (55+hrs)			-0.236	-1.89	-0.212	-1.65
Employed FT (hrs unknown)			-0.374	-0.34	-0.005	0.00
Housing tenure (base = Own/purchasing)						
Rent			-0.172	-2.37	-0.151	-2.02
Rent/buy, rent-free			0.461	2.21	0.435	2.06
Negative or zero personal income			-0.033	-0.12	-0.022	-0.08
Log personal income			-0.041	-1.43	-0.042	-1.43
Negative or zero income of others in hh			0.510	1.62	0.750	2.34
Log income of others in household			0.048	1.67	0.082	2.83
Benefit recipient			0.054	0.73	0.036	0.48
High life satisfaction (8+)			0.092	1.62	0.055	0.94
Moved between w1 and w2			-0.564	-8.38	-0.601	-8.66
Partially cooperating household					-1.059	11.51
Number of calls made to h'hold					-0.071	-6.82
Same interviewer in both waves					0.021	0.38
Interview time in h'hold					0.000	0.33
Interview time in h'hold unknown					0.422	2.81
Assistance required during ivw (base = No assistance required)						
Due to English difficulties					-0.407	-2.26
Due to sickness/disability					-0.690	-2.33
Due to other reason					0.478	1.49
English difficulties experienced					-0.288	-2.04
Other language difficulties					-0.340	-1.26
Non-cooperative in interview[a]					-0.828	-5.73
Suspicious of study[b]					-0.563	-5.14
Log likelihood	-5090.2		-5019.5		-4804.0	
Chi-squared	602.1		743.5		1174.5	
Pseudo R-squared	0.056		0.069		0.109	

Notes: The pseudo R-squared equals 1 minus the ratio of the log likelihood of the fitted function to the starting value for the log likelihood (a function with only an intercept).

a Equals 1 if the respondent's cooperation was described as fair, poor or very poor and 0 if described as excellent or good.

b Equals 1 if the interviewer reported that the respondent was suspicious (either somewhat or very suspicious) of the study after the interview was completed.

The additional outcome variables included in specification 2 raise the explanatory power of the model, but not by much. Indeed, of the variables considered, only mobility and housing tenure have effects that are statistically significant at conventional levels. Again, this is a positive finding. While most of the outcome variables considered here co-vary significantly with attrition probabilities, once we condition on a broad range of covariates these effects decline to insignificance. Furthermore, easily the most important variable for attrition – changing address – is one where, a priori, we would expect very large effects. Most obviously, households that change address between waves will simply be harder to find. Indeed, as revealed below, in Table 4, once movers are located, there is no evidence that they are any more likely to refuse to participate.

In specification 3 we include a range of variables which describe the interview situation at wave 1. A priori, we expected that all of these variables would be significantly related to attrition. Laurie, Smith and Scott (1999), for example, in their analysis of attrition over the first four waves of the British Household Panel Survey (BHPS), found that response was significantly linked to interviewer observations recorded in wave 1 about the level of respondent cooperation and the presence of health and language problems that affected the interview, to whether the respondent was from a household where other household members did not cooperate in wave 1 and to interviewer continuity. We were able to include controls for all of these variables in this analysis. In addition, we also included the total number of calls that had to be made to the household during wave 1, which we argue is a good measure of how difficult it will be to reach a household in wave 2.[6] We also included a measure of total time spent by the interviewer in the household in wave 1 given it is typically assumed that interview lengths can have an influence on survey participation.[7] Finally, we included an interviewer-assessed measure of the degree of suspicion of the study exhibited by the respondent.

As a group, these variables are clearly of considerable importance. The overall explanatory power of the model is markedly enhanced, and most of the new variables exhibit large and statistically significant effects in the expected direction. Thus, in line with the results reported by Laurie et al. (1999) for the BHPS, we see that coming from a partially responding household is a major risk factor for non-participation at the next wave. Indeed, the estimates suggest that the mean predicted probability of a sample member from a partially cooperating household responding in wave 2 was about 84 per cent that of sample members from fully responding households.[8] Also as expected, households that were more difficult to reach in wave 1 were much more likely to be non-respondents in wave 2. Similarly, interviewer assessments collected in wave 1 about the degree of respondent cooperation and suspicion were found to be good predictors of wave 2 non-response. Respondents whose interviews had to be assisted by others because of health problems or physical incapacity or because of English language difficulties in wave 1 were also found to be more likely to be non-respondents. Finally, respondents with poor English language skills, even though they did not need interpreters to complete the interview, were also less likely to respond in wave 2.

[6] This variable will also be a function of the size of the household.

[7] In general, evidence from telephone and personal surveys support the hypothesised negative relationship between interview length and response rates, though the magnitude of the effect is arguably quite small (Frankel and Sharp 1981, Collins et al. 1988).

[8] The mean predicted probability of a person from a fully responding household responding in wave 2 was 0.88. This compares with a mean predicted probability of 0.74 for a person from a partially responding household.

Not all of the results here, however, were in accord with expectations. Laurie et al. (1999) placed great emphasis on their finding in the BHPS data that respondents who were assigned the same interviewer each year were more likely to respond. The HILDA Survey also pursued a policy of maintaining interviewer continuity wherever possible, and indeed about 44 per cent of all households in wave 2 were assigned the same interviewer from wave 1. Nevertheless, we find no evidence that interviewer continuity matters.

The insignificance of the interview time variable might also surprise some. Interview time, however, is not a direct measure of instrument length. Indeed, it is a product of both instrument length and respondent interest in the survey. That is, the respondents who most enjoy the survey experience are also likely to take longer to interview. Given this, its insignificance should not be surprising.

Of course, from a data user perspective, the more important question is not what influences attrition, but whether these influences are correlated with variables of interest (such as the outcome variables listed in Table 2). The results presented in Table 3 suggest that, for the most part, it is reasonable to conclude that there is little correlation between the variables of interest and the interview situation variables. The only outcome variable greatly affected by the inclusion of these interview situation variables is the income of other household members. This is a direct reflection of the correlation between this variable and the variable for partially cooperating households which, in turn, simply reflects the fact that single-person households cannot be partially cooperating households.

Our final set of analyses of attrition involved estimating logit equations that distinguished between two key stages in the response process – establishing successful contact and then obtaining a successful interview. The results, using the full specification, are presented in Table 4.

These results add insights into the response process. Most obviously, a number of the explanatory variables are of much greater importance in explaining variations in the probability of making contact while others are of much greater importance in explaining the variation in response probabilities once contact has been established. The clearest example of the importance of this distinction is the variable identifying those who move house. As noted earlier, the relocation of a household has a marked impact on the likelihood of finding the members of that household. Indeed, based on the parameter estimates the mean predicted probability of making contact with a mover is 92 per cent. While seemingly high, this is well below the predicted mean probability of making contact with a non-mover – almost 99 per cent. Movement, however, has no influence on the likelihood of obtaining an interview once contact is established.

Other variables which have markedly different impacts at the two separate stages of the response process include the following:

- sex – females are easier to make contact with but are no more or less likely than males to agree to an interview;
- household type – married couple households are easier to find than other household types, but again refusal probabilities do not vary with household type;
- number of persons in the household – a greater number of adults in the households is associated with lower response probabilities but has no bearing on the likelihood of making contact, while the number of children has the opposite effects;

Table 4: Two-equation Wave 2 Response Model, Logit Results

Variable	*Contact*		*Response* │ *contact*	
	Coeff.	*Z*	*Coeff.*	*Z*
Constant term	1.237	1.33	1.745	3.42
Area of residence (base = Sydney)				
Other NSW	0.667	2.97	0.000	0.00
Melbourne	0.439	2.29	-0.026	-0.27
Other Vic	0.615	2.12	-0.040	-0.30
Brisbane	0.296	1.32	0.103	0.81
Other Qld	0.242	1.22	0.553	4.13
Adelaide	0.869	2.90	0.179	1.25
Other SA	0.113	0.34	0.188	0.92
Perth	0.061	0.27	0.332	2.36
Other WA	-0.410	-1.35	0.187	0.89
Tasmania	1.642	2.69	-0.390	-2.26
Northern Territory	0.418	0.54	2.505	2.45
ACT	-0.079	-0.18	0.391	1.34
Sex (Female=1)	0.262	2.09	-0.030	-0.45
Age	0.069	2.96	0.073	6.21
Age squared	-0.0004	-1.65	-0.0008	-6.52
Marital status (base = Married)				
De facto	2.311	1.92	0.398	0.61
Separated	0.477	0.80	0.423	0.92
Divorced	0.431	0.73	0.218	0.50
Widowed	1.066	1.46	0.524	1.19
Single	0.795	1.39	0.277	0.64
Relationship in household (base = Married couple)				
De facto couple	-2.831	-2.36	-0.528	-0.80
Lone parent	-1.203	-2.08	-0.221	-0.52
Child / relative	-1.073	-1.83	-0.251	-0.58
Lone person	-1.366	-2.15	0.053	0.12
Unrelated	-1.562	-2.63	-0.335	-0.74
Number of adults in HH	-0.025	-0.28	-0.236	-5.12
Number of children in HH	-0.145	-2.72	-0.040	-1.25
Country of birth (base = Australia)				
Overseas: Main English-speaking	-0.437	-2.42	-0.186	-1.85
Overseas: Other	-0.275	-1.43	-0.323	-3.35
Indigenous	-0.642	-2.57	0.308	1.25
Education (base = Year 11 and below)				
Year 12	0.324	1.84	0.134	1.41
Certificate	0.027	0.18	0.192	2.42
Diploma	0.314	1.25	0.329	2.65
Bachelor or higher	0.467	2.32	0.771	7.13
Dwelling type (base = Separate house)				
Semi-detached	-0.017	-0.09	0.051	0.41
Unit / apartment / flat	-0.278	-1.65	-0.126	-1.07
Other dwelling	-0.417	-0.89	0.782	1.64

Table 4 (cont'd)

Variable	*Contact*		*Response | contact*	
	Coeff.	*z*	*Coeff.*	*z*
SEIFA disadvantage (base = Lowest quintile)				
2nd lowest quintile	-0.171	-1.06	-0.295	-2.97
Middle quintile	0.070	0.40	-0.164	-1.62
2nd highest quintile	0.049	0.27	-0.174	-1.71
Highest quintile	0.235	1.18	-0.160	-1.49
Employment and LF status (base = Not in Labour Force)				
Unemployed	-0.026	-0.13	-0.044	-0.28
Employed PT (1-24hrs)	0.264	1.38	0.072	0.65
Employed PT (25-34hrs)	0.657	2.24	-0.365	-2.58
Employed FT (35-44hrs)	0.616	3.13	-0.370	-3.31
Employed FT (45-54hrs)	0.758	2.98	-0.198	-1.45
Employed FT (55+hrs)	0.189	0.74	-0.329	-2.28
Employed FT (hrs unknown)			-0.375	-0.33
Housing tenure (base = Own/purchasing)				
Rent	-0.198	-1.39	-0.110	-1.29
Rent/buy, rent-free	1.052	1.74	0.323	1.45
Negative or zero personal income	0.727	1.24	-0.181	-0.57
Log personal income	-0.024	-0.40	-0.044	-1.34
Negative or zero income of others in hh	1.642	2.59	0.557	1.55
Log income of others in household	0.158	2.66	0.059	1.86
Benefit recipient	0.190	1.30	-0.006	-0.07
High life satisfaction (8+)	0.213	1.87	-0.001	-0.01
Moved between w1 and w2	-2.044	15.67	0.002	0.03
Partially cooperating household	-0.552	-2.69	-1.102	-11.27
Number of calls made to h'hold	-0.054	-2.59	-0.075	-6.45
Same interviewer in both waves	-0.192	-1.61	0.065	1.05
Interview time in h'hold	0.000	-0.09	0.000	0.36
Interview time in h'hold unknown	0.525	1.61	0.375	2.30
Assistance required during ivw (base = No assistance required)				
Due to English difficulties	-0.136	-0.41	-0.491	-2.47
Due to sickness/disability	-0.489	-0.79	-0.628	-1.92
Due to other reason	0.766	0.98	0.397	1.15
English difficulties experienced	-0.825	-3.17	-0.177	-1.11
Other language difficulties	-0.987	-2.43	-0.070	-0.21
Non-cooperative in interview[a]	-0.933	-3.50	-0.723	-4.43
Suspicious of study[b]	0.023	0.09	-0.656	-5.59
Log likelihood	-1363.1		-4074.2	
Chi-squared	951.5		895.1	
Pseudo R-squared	0.259		0.099	
N	13810		13409	

Notes: See Table 3.

- indigenous status – the lower response probabilities of indigenous Australians are entirely due to greater difficulties making contact; and
- employment status – workers (but not those working very long hours) are relatively easy to make contact with but, compared with non-workers, are more likely to refuse to participate.

Effect of Attrition on Population Estimates

We now briefly consider the impact that attrition has on population estimates. Table 5 provides two population estimates for selected sample characteristics measured at wave 2. The first population estimate is simply based on population weights carried forward from wave 1, whereas the second population estimate uses weights that have made adjustments for the non-random attrition (based on the estimation of a model similar to that reported in Table 3).[9] This is the standard technique for dealing with possible attrition.

As can be seen from this table, for a number of the characteristics the two estimates are quite different. The largest differences occur when the attrition has been high for a particular group and the population size is reasonably large. For example, we would understate the population estimates for the proportion of people who live in Sydney, are single, are relatively young, were born in a mainly non-English speaking country, are currently renting, or who have moved since wave 1. Conversely, we would overstate the proportion of people who are married, born in Australia, have high education levels, or have high levels of life satisfaction.

In areas where the attrition has been reasonably differential but the proportion of the population affected is relatively small, the effect on the population estimates is less apparent. This can be seen in the estimates for indigenous status, the proportion of people living in the Northern Territory or ACT, those that are widowed, the unemployed and those reporting low levels of life satisfaction.

In general though, it is clear that the estimates most affected by attrition are demographic characteristics. Consistent with our earlier multivariate analysis, weighting for attrition between wave 1 and wave 2 appears to make very little difference to the estimated distribution of our selected outcome variables.

[9] The actual specification was similar to specification 3 reported in Table 3, but included some additional variables.

Table 5: Wave 2 Characteristics Using Unadjusted Weights and Weights Adjusted for Attrition (%)

Wave 2 characteristic	*Unadjusted weights*[a]	*Adjusted weights*[b]	*Wave 2 characteristic*	*Unadjusted weights*[a]	*Adjusted weights*[b]
Area			Indigenous status		
Sydney	19.9	21.2	Indigenous	1.6	1.7
Rest of NSW	12.8	12.5	Non-indigenous	98.4	98.3
Melbourne	18.2	18.4	Education attainment		
Rest of Victoria	7.1	7.0	Year 11 or below	33.1	33.4
Brisbane	8.8	8.7	Year 12	12.3	13.0
Rest of Qld	10.1	9.8	Certificate	26.7	26.6
Adelaide	6.1	5.8	Diploma	8.5	8.4
Rest of SA	2.1	2.1	Degree of higher	19.3	18.5
Perth	7.5	7.2	Dwelling type[d]		
Rest of WA	2.6	2.6	House	80.1	79.3
Tasmania	2.5	2.5	Semi-detached	7.3	7.4
Northern Territory	0.7	0.7	Flat, unit, apartment	9.8	10.4
ACT	1.7	1.6	Other	1.7	1.7
Sex			Index of disadvantage		
Male	49.2	49.5	Lowest quintile	18.2	18.8
Female	50.8	50.5	2nd lowest quintile	22.0	21.9
Age group (years)			Middle quintile	18.7	18.4
15-19	7.0	7.6	2nd highest quintile	18.2	18.2
20-24	7.7	8.6	Highest quintile	23.0	22.7
25-34	18.6	19.0	Labour force status		
35-44	19.7	19.3	Employed full-time	43.2	43.3
45-54	18.2	17.8	Employed part-time	18.7	18.6
55-64	13.1	12.5	Unemployed	3.6	3.9
65-74	9.1	8.6	Not in labour force	34.4	34.1
75+	6.6	6.5	Life satisfaction		
Marital status			Low (0-4)	3.2	3.3
Married	54.7	53.5	Medium (5-7)	30.9	31.3
De facto	9.2	9.6	High (8-10)	65.9	65.4
Separated	3.2	3.1	Household tenure		
Divorced	5.3	5.0	Own / purchasing	73.0	71.5
Widowed	5.5	5.3	Rent	24.2	25.7
Single	22.2	23.5	Other	2.7	2.8
Relationship in h'hold			Benefit recipient status		
Married couple	54.3	53.1	Benefit recipient	34.3	34.4
De facto couple	9.0	9.4	Not benefit recipient	65.6	65.6
Lone parent	5.4	5.4	Annual household income		
Child/relative	13.9	14.9	<$20,000	15.6	15.4
Lone person	15.3	14.8	$20,000 - $39,999	19.6	19.6
Unrelated	2.1	2.3	$40,000 - $59,999	18.3	18.4
No. of adults in household[c]			$60,000 - $79,999	15.6	15.7
One adult	17.6	17.1	$80,000 - $119,999	19.4	19.4
Two adults	52.6	51.3	$120,000+	11.5	11.5
Three adults	17.4	18.1	Annual personal income		
Four or more adults	12.5	13.5	Negative / nil	3.0	3.3
No. of children in househol			$1 - $9,999	22.4	22.0
No children	69.8	69.8	$10,000 - $14,999	13.5	13.3
One child	12.3	12.6	$15,000 - $29,999	20.7	20.9
Two children	11.8	11.5	$30,000 - $59,000	28.5	28.2
Three or more	6.1	6.1	$60,000+	12.2	11.9
Country of birth			Moved b/w wave 1 and 2		
Australia	73.8	72.2	Moved	16.3	18.0
O/S: Main English-spkg	10.9	10.9	Did not move	83.7	82.0
O/S: Other	15.2	16.9			

Table 5 (cont'd)

Notes:

a	The unadjusted weights are the wave 1 weights for responding wave 2 individuals.
a	The adjusted (for attrition) weights are calculated by: i) multiplying the wave 1 responding person weights by the inverse of the modeled probability of response to wave 2 (given an interview provided in wave 1); and ii) benchmarking these weights for the wave 2 respondents and out-of-scopes to wave 1 characteristics such as age, sex, State, part of State, and labour force status.
c	An adult is defined here as anyone aged 15 years or over. A child is therefore anyone under the age of 15 years.
d	Not reported are a small proportion of cases where the dwelling types was not reported.

Missing Data and Imputation

Another potential source of response bias is item non-response. That is, while a member of a selected household may agree to an interview, they may then subsequently either refuse or be unable to answer some of the questions asked. Frequency counts from both the wave 1 and wave 2 interview data (as distinct from the data collected as part of the self-completion questionnaire – or SCQ), however, indicated that, with the exception of questions requiring respondents to provide a monetary value, missing data is generally not a large problem.[10] A summary of all items where item non-response exceeds two per cent is provided in an Appendix. As should be apparent, the interview items with high item non-response were generally the monetary items. There were of course exceptions, but in most cases these exceptions can be explained. For example, the apparently high rate of item non-response for the questions on difficulties experienced with childcare is largely a function of the fact that not all of these questions would apply to households with children, and we do not distinguish don't know responses from cases where the question does not apply. Thus the main reason 78 per cent of households with children did not provide an answer to the question about difficulties finding care for special needs children is that they do not have children with 'special needs'. Indeed, while not reported in the Appendix, for the majority of variables there were no missing cases at all, and where there were missing cases, the incidence was generally relatively low – less than 2 per cent.

Item non-response rates for the SCQ, however, were higher, averaging 2.5 per cent per item in wave 1 and 2.8 per cent in wave 2.[11] This higher rate of non-response is to be expected given the self-completion nature of this instrument. The slightly higher rate of item non-response in wave 2, on the other hand, was not expected, and possibly suggests a degree of survey fatigue.

On balance though, it is only the high incidence of missing data for monetary item which we believe is of any serious concern. In particular, the relatively high level of 'missingness' has the potential to undermine our efforts to collect data on both income and, in wave 2 at least, on wealth.[12] As a result, it was deemed necessary to impute values for key income and wealth components that were missing (see Watson and Wooden 2003). In what follows, we focus exclusively on the issue of missing income and wealth data. Specifically, we present data on the extent of the problem, and then present estimates which provide at least some indication of how successful the imputation process was.

Missing Income Data

As previously observed with respect to wave 1 data (Watson and Wooden 2002), the income section is affected by relatively high rates of item non-response. There are two main sources of missing data. First, some respondents are either unable or unwilling to indicate how much income they derive from a particular source. Second, not all eligible adult members (persons

[10] Missing data is defined here to include both refusals to provide an answer and the inability to provide an answer (i.e., a response of 'don't know').

[11] The figure of 2.5 per cent for wave 1 is slightly higher than that previously reported (Watson and Wooden 2002). This mainly reflects a difference in the way we have treated questions providing explicit don't know options.

[12] Unlike income, and with the exception of the value of the primary residence, we do not intend to collect data on household assets and debts each year. Instead, the hope is that wealth data will be collected every 3 or 4 years.

aged 15 years or older) of cooperating households agreed or were able to be interviewed. The latter is obviously a more serious problem than the former. In the former type of case, we usually have information on at least some of their income components. Further, we have a lot of other information about these individuals. Taken together, these two things will mean imputations are likely to be reasonably reliable. In contrast, for those cases where no interview is obtained we have no information on either the amount of income received or from which components income is derived. Furthermore, while we have a lot of information about the households these non-respondents come from, we often know very little about the individuals themselves.[13] Table 6 provides summary data on the extent of the missing income data problem. The table consists of three panels providing data on responding individuals, enumerated persons (which thus includes members of responding households who were not interviewed) and households, respectively. Focusing first on responding individuals, it can be seen that that the levels of missing data associated with current income (and the HILDA Survey only collects current data on wages and salaries and government benefits) are relatively modest.[14] When we look at financial year income, with the exception of government pensions and benefits, most components are affected by relatively high rates of missing data. This is especially true of business income and investment income. In the case of business income, for example, in both waves around 25 to 26 per cent of respondents who identified as being an owner of an unincorporated business were unable or unwilling to provide an estimate of their share of profits (or losses) from those businesses. Table 6 also reveals that for most items the incidence of missing data changed only slightly between wave 1 and wave 2. The notable exception is windfall income, but we believe this difference is largely (if not entirely) the result of modifications to the questionnaire introduced in wave 2. Specifically, two additional categories of income previously included under ‘other sources’ – redundancy and severance payments and inheritances / bequests – were separately identified, both of which are treated here as windfall income.

Overall, in wave 1 there were just over 2000 individual cases where at least one component of income was unknown. These cases represent 14.7 per cent of the total sample. In wave 2 this proportion fell slightly to 13.9 per cent.

As noted above, not all eligible adult members of cooperating households agreed, or were able, to be interviewed. In wave 1 there were 810 incomplete households, which represent 10.5 per cent of the household sample. In wave 2 there were 704 such cases, representing 8.5 per cent of the sample of responding households. This means the problem of missing data will be magnified when dealing with variables constructed by combining the different responses of household members. The most obvious example here is household income. While we are able to derive a gross financial year income estimate for 85 to 86 per cent of all individual sample members without the need for any imputation, an estimate of total household gross income can only be derived in 71 per cent of cases in wave 1 and 72 per cent in wave 2.

[13] The notable exception here is where the non-respondent was a respondent in wave 1.

[14] That said, Watson and Wooden (2002) reported even lower numbers. We suspect this difference was mainly the result of the earlier analysis not appropriately taking account of multiple job holders.

Table 6: Missing Income Data by Component, Waves 1 and 2

	Wave 1		*Wave 2*	
	n	*% of valid N*[a]	*n*	*% of valid N*[a]
RESPONDING PERSONS				
Current income				
Wages and salaries	462	6.0	310	4.2
Benefits	136	3.2	81	2.1
Financial year income				
Wages and salaries	666	7.9	550	6.9
Australian government pensions	67	1.6	52	1.7
Foreign government pensions	1	0.5	3	1.4
Business income	404	25.5	366	26.1
Investments				
Interest	661	19.5	596	18.6
Dividends and royalties	584	14.6	521	14.5
Rent	240	18.3	189	13.9
Private pensions	59	6.2	41	4.6
Private transfers	28	7.1	89	23.1
Total financial year income	2054	15.6 (14.7)	1817	14.7 (13.9)
Windfall income	32	4.1 (0.2)	31	2.9 (0.2)
ENUMERATED PERSONS				
Total financial year income	3212	21.2	2795	19.9
Windfall income	1190	7.9	1009	7.2
HOUSEHOLDS				
Total financial year income	2243	29.2	2009	27.7
Windfall income	838	10.9	723	10.0

Note: a The valid N for enumerated persons and households is all cases. In contrast, the valid N for responding persons is generally all cases where the expected value is non-zero. The exceptions to this are business income and rental income where both zero and negative values are possible. For total financial year income and windfall we report percentages with and without zero cases (the figures in brackets include zero cases).

Imputed Income Estimates

The foregoing clearly suggests that missing data on income and its components is problematic. As a result, it was decided to impute the missing values for most of the major income components. Imputation essentially involves making use of the data that are collected to make informed guesses about the missing values. A detailed description of the imputation process, including a discussion of potential weaknesses, is provided in Watson (2004), but essentially the imputation method chosen involved identifying for every respondent with missing values a ‘nearest neighbour’ using parametric regression techniques. The responses provided by these ‘neighbours’ were then assumed to provide unbiased estimates of the income components of those persons who failed to provide an answer.

Table 7 provides a summary of both how the imputed income estimates compare with ABS income estimates from the Survey of Income and Housing Costs (SIHC) and the impact that imputation had on the estimates from the HILDA Survey. Unfortunately, the most recent ABS estimates available only provide information for the 1999/2000 financial year.[15] As a result, to enable comparisons to be made with the HILDA data, the ABS estimates have to be adjusted on the basis of other information (e.g., the Consumer Price Index or male average weekly earnings).

The first point to be noted from this table is that imputation does not appear to have had much impact on the mean estimates of both wage and salary income and benefit income. For example, the difference in the imputed and non-imputed estimates of income from wages and salaries is, in both waves, less than one per cent.

Second, the impact of imputation on other income components, and especially business income and investment income, is much larger. This is simply a reflection of the relatively large number of cases where the value of investment or business income is known to be non-zero but where a value was not provided. More importantly, the impact of imputation is to move our estimates of business income much closer to the ABS estimate. In contrast, the estimate of investment income moves further away.

The third point of interest is that the comparisons with the ABS data suggest that there may be a tendency for the HILDA Survey to overstate income, especially wage and salary income and investment income. Our best assessment is that the HILDA Survey estimate of annual wages and salary income is around 7 to 8 per cent higher than that derived from the ABS Survey of Income and Housing Costs. The estimates of investment income, on the other hand, are more than 50 per cent greater than the comparable ABS estimate following imputation. This suggests either that the HILDA Survey sample differs from the ABS sample in some way not reflected in our weighting structure, or that there are differences in the way respondents report income in the two surveys. Evidence for the first explanation can be seen in data which compares the occupational distribution of the employed sub-sample from the HILDA Survey with the occupational distribution in the ABS Labour Force Survey. As reported in Table 8, it appears that there is an over-representation of persons working in the managerial and professional occupations in the HILDA Survey, and of course persons in these occupational groupings earn, on average, higher wages and typically have greater levels of financial investments. This is important given that neither occupation nor education were used in the weighting structure employed in the HILDA Survey.[16]

The possibility of reporting bias can also not be discounted, though it does appear to provide a less plausible explanation for the differences with ABS estimates. Reporting biases, for example, might stem from measurement errors due to recall, bearing in mind that estimates of financial year income relate to the financial year preceding interview. However, if this were

[15] While the survey is described as the 2000-01 Survey of Income and Housing Costs, the date here refers to the period over which the survey was conducted. The survey thus generates population estimates for current income that cover the period 2000-01, but data on the financial year income of individuals relates to the financial year prior to survey, or 1999-2000.

[16] Both of these variables are subject to measurement errors resulting from the coding process, and thus were not deemed suitable for use in the construction of population weights.

Table 7: Mean Financial Year Income Estimates ($) – SIHC and HILDA Survey Compared

	Survey of Income and Housing Costs		*HILDA Wave 1 (2000/01)*		*Difference from SIHC (HILDA-SIHC)*	
	99/00	*Approx 00/01*[a]	*Without imputation*	*With imputation*	*Without imputation*	*With imputation*
Wages and salaries	18510	19528	20955	21152	1427	1624
Benefits	2312[c]	2451	2202	2219	-249	-232
Business income	1737	1780	1159	1,658	-621	-122
Investment income	1049	1075	1,322	1,537	247	462
Sum of above components	23608	24834	25638	26566	804	1732
Other income[d]	652	668	1,164	1,237	N/A	N/A
Windfall income	N/A	N/A	302	311	N/A	N/A

	Survey of Income and Housing Costs		*HILDA Wave 2 (2001/02)*		*Difference from SIHC (HILDA-SIHC)*	
		Approx 01/02[b]	*Without imputation*	*With imputation*	*Without imputation*	*With imputation*
Wages and salaries		20342	21700	21836	1358	1494
Benefits		2555	2540	2557	-15	2
Business income		1838	1381	1845	-457	7
Investment income		1110	1305	1700	195	590
Sum of above components		25845	26926	27938	1081	2083
Other income[d]		690	1550	1677	N/A	N/A
Windfall income		N/A	1405	1428	N/A	N/A

Notes:
a. SIHC estimates for 00/01 financial year are calculated from 99/00 by applying a 5.5% increase to wages and salaries, a 6.0% increase to benefits and a 2.5% increase to other income components.

b. SIHC estimates for 01/02 financial year are calculated from 99/00 by applying a 9.9% increase to wages and salaries, a 10.5% increase to benefits and a 5.8% increase to other income components.

c. $403 in Family Tax Benefit has been removed from the SIHC estimates (as this is calculated separately in HILDA). Neither SIHC nor HILDA estimates include Child Care Benefit.

d. Income from other sources cannot be directly compared with the ABS as the HILDA survey has not clearly differentiated regular from irregular components. We have only assumed which sources are more likely to be regular and placed them in the 'other' category. Those more likely to be irregular are placed in 'windfall' income.

Source: The ABS data were provided by Roger Wilkins and come from Survey of Income and Housing Costs, 2000/2001, confidentialised unit record file (cat. no. 6541.0.30.001).

Table 8: Occupational Distribution of Employment –
HILDA Survey and ABS Labour Force Survey Compared

Occupation	*ABS LFS*		*HILDA Survey*	
	Aug 2001	*Aug 2002*	*Wave 1*	*Wave 2*
Managers and administrators	7.7	7.4	8.9	8.2
Professionals	18.7	19.0	21.6	21.5
Associate professionals	11.6	11.6	11.8	13.2
Tradespersons and related workers	12.8	12.6	12.0	12.0
Advanced clerical and service workers	4.4	4.2	3.4	3.2
Intermediate clerical, sales and service workers	17.1	17.1	16.1	16.3
Intermediate production and transport workers	8.8	8.5	8.0	8.1
Elementary clerical, sales and service workers	9.8	10.1	9.7	9.4
Labourers and related workers	9.1	9.6	8.5	8.1

Source: ABS data come from *The Labour Force, Australia* (cat. no. 6203.0), August 2001 and August 2002 issues.

so we would expect the HILDA Survey estimates to be more accurate. This is because, on average, the HILDA Survey interviews were held on dates closer to the end of the preceding financial year.[17] Very differently, we would generally expect self-reported surveys to lead to upward biased estimates of financial year wage and salary income given that some respondents will use their current wage and salary income in arriving at an estimate of wage and salary income during the preceding financial year. Nevertheless, it is difficult to see why the HILDA Survey and the ABS Surveys would be different in this respect.

Overall, the income data suggest that in a cross-sectional sense at least, the imputation process seems to have been quite successful (though we have uncovered a potential problem with the population coverage of the sample). However, while the imputation process may have produced unbiased cross-sectional estimates of mean income, this may not be true of the estimated changes in income over time. Indeed, this is almost unavoidable. Imputed estimates invariably have considerable measurement error. This does not matter for cross-section estimates provided the error is random. Estimates of first differences, however, will be biased away from zero even if the cross-section estimates on which they are based are unbiased. Evidence of this problem is documented at much greater length in Watson (2004). In particular, she reports that the cross-wave correlations in reported income are much lower for cases where income had to be imputed. While a lower correlation might be expected, given

[17] Interviews for the ABS Survey of Income and Costs are evenly distributed over a 12-month period and thus for the average respondent, the financial year income date relate to a period that was between 6 and 18 months ago. In contrast, in the HILDA Survey, most interviews are conducted in the period August to December of each year. The average recall period is thus between 3 and 15 months earlier.

the need for imputation is likely to be correlated with both job and spatial mobility, the size of the differential far exceeds what could be thought of as reasonable. For example, for cases where no imputation is required, the year-on-year correlation in total financial year income is 0.70. In contrast, for cases where income had to be imputed in both waves, the correlation was just 0.29. Furthermore, these figures only cover persons who were actually interviewed in both waves. Once we consider non-respondents who were members of partially responding households, the inter-year correlation declines even further. Indeed, for persons who were not interviewed in either year the correlation across years in imputed income was close to zero.

Such results suggest the need for users to exercise extreme caution when using the HILDA data to analyse income mobility. While ignoring cases with missing income data is likely to lead to an understatement of income mobility, reliance on the fully imputed data set will almost certainly to lead to the opposite conclusion.

Missing Wealth Data

Given questions about income were heavily affected by item non-response, it can be expected that the wealth questions included in wave 2 would be similarly affected. As can be seen from Table 9, with respect to almost all of the wealth components, there were significant proportions of respondents who were unable to quantify the value of the type of asset or liability in question. This is reflected in the figures reported in the third column of the table which indicate the percentage of non-zero cases (persons or households which report owning the type of asset or liability in question) not providing a value to the size of that asset or liability. The answers range from a low of 2.1 per cent in the case of credit card debt up to a very high 29 per cent in the case of trusts. Note that the questions on wealth were divided between the Household Questionnaire (HQ) and the Person Questionnaire (PQ). Thus, in total, almost 20 per cent of responding households did not provide a value for at least one of the wealth components included in the HQ while almost 15 per cent of responding individuals failed to provide an answer to one of the components included in the PQ. In total, and given the added complexity of incomplete households, we were unable to directly determine total net household wealth for just over 39 per cent of all wave 2 responding households.

Imputed Wealth Estimates

As for income, it was thus determined that it would be helpful for users if missing values on major wealth components and aggregates could be imputed. The procedure was essentially the same as that employed for income, but undertaken by staff at the Reserve Bank of Australia (RBA). More details are provided in Watson (2004).[18]

As a test of how well the imputation process performed (at least in a cross-sectional sense), we compare, in Table 10, HILDA estimates of household assets, debts and net worth with the national aggregates compiled by the ABS as part of the Australian System of National Accounts. Further, we also report estimates regularly reported by the RBA given these tend to depart from the ABS figures in at least one key respect. All figures are in billions of dollars and the HILDA Survey estimates have been weighted up to reflect the total population of private households.

[18] There is at least one major difference between the methods used to impute income and the methods used to impute wealth. The former made use of responses on income components from both waves and not just the wave for which there was missing values.

Table 9: Sources of Missing Wealth Data, Wave 2

Wealth component	*Missing cases[a] (no.)*	*Valid cases[b] (no.)*	*% of valid cases missing[c]*	*% of all cases missing*
HQ wealth components				
Housing equity	531	5176	10.2	7.3
Equities	455	2978	15.0	6.3
Other cash-type investments	29	241	8.3	0.4
Trusts	123	390	29.0	1.7
Childrens' bank accounts	85	1399	5.8	1.2
Life insurance policies	200	794	24.1	2.8
Vehicles	145	6355	2.2	2.0
Collectibles	150	1050	12.0	2.1
Net business worth	231	1090	20.6	3.2
Total of HQ wealth components	1427	7245	19.7	19.7
PQ wealth components				
Bank accounts	905	12825	7.0	6.9
Superannuation[c]	939	8843	10.5	7.2
Credit card debt	160	7448	2.1	1.2
Personal loans and other debts	174	3679	4.6	1.3
Total of PQ wealth components	1887	13041	14.5	14.5
Total household wealth	2846	7245	39.3	39.3

Notes: a A 'missing case' is any observation where the respondent was unable to either indicate whether they had an asset or liability of the type in question, or were unable to provide a value for that asset or liability.

b A 'valid case' is any observation where the respondent reported owning the asset in question, having a credit card or having personal loans or debts.

c The figures reported in this column do not exactly equal 'missing cases' divided by 'valid cases'. This is because for all components there are a small number of cases where respondents did not answer the key screening question.

d In the case of superannuation assets, respondents were asked first to indicate which of seven broad monetary bands represented the current value of their superannuation. They were then asked to estimate the exact value of these assets within that band. For the purposes of this table we have only treated as missing those cases where individuals could not or would not choose a category. There are a total of 582 cases where a range was provided but not an exact value within that range.

Unfortunately differences of inclusion and exclusion among the different sources make comparisons far from straightforward. The most comparable figures are for financial assets. Here the ABS and the RBA only differ in how they treat unfunded superannuation and pre-paid insurance (the RBA exclude them whereas the ABS do not). Conceptually the HILDA Survey falls between the two, including unfunded superannuation but excluding pre-paid insurance premiums, and as it turns out the HILDA estimate does lie between the ABS and RBA estimates. If we adjust the HILDA data by adding the ABS estimate of pre-paid insurance premiums – just over $28b – we find that the HILDA estimate is about 93 per cent of the ABS estimate. The HILDA Survey thus slightly understates the volume of financial assets. This, however, is to be expected given it has been well established that, without any oversampling of the very wealthy, surveys will understate wealth holdings at the very upper end of the distribution (Juster et al. 1999). Indeed, if anything, the HILDA Survey estimates are a little 'too good'; we would have expected the underestimation of wealth to have been greater.[19]

Turning to non-financial assets, the conceptual differences between the three sources are much more marked. The HILDA Survey, for example, does not provide an estimate of total consumer durables, only vehicles, but does collect data on 'collectibles', whereas such assets are excluded from the other sources. Further, the RBA estimates exclude non-financial assets held through the businesses that individuals own. More importantly, there is much greater divergence in the different estimates. The most significant source for this difference is property values. The RBA, for example, estimated property to be worth $2252b at the end of December 2002 which exceeds the total value of all non-financial assets reported in the National Accounts. As it turns out, the HILDA estimates are much more in line with the RBA, mainly because the aggregate property value derived from the self-reports in the HILDA Survey – $1932b – is also much larger than the valuation derived by the ABS. Overall, once we take account of the marked differences in the composition of the assets

Table 10: Household Wealth: HILDA Estimates and National Aggregate Estimates Compared, 2002 ($ billion)

	ABS[a]	*RBA*[b]	*HILDA*[c]
Financial assets[d]	1236.7	1084	1125.1
Non-financial assets[e]	1955.4[f]	2391	2440.3
Total assets	3192.1	3474	3565.4
Financial liabilities	630.8	640.5	516.5
Net worth	2561.3	2833.5	3048.9

Notes:
a. All figures reported are intended to represent an average of the September 2002 and December 2002 quarters. In addition to households, the assets and liabilities of non-profit organisations are also included.
b. The RBA figures reported here apply to the December quarter of 2002.
c. Figures from the HILDA Survey were collected over the period August 2002 to March 2003, but with October 2002 being the median observation point. All data are weighted and include imputations for missing values. The scope of the survey excludes households living in very remote parts of Australia.
d. The RBA estimate is based on the ABS source but excludes unfunded superannuation claims and pre-paid insurance claims. Conceptually the HILDA Survey includes unfunded superannuation but excludes prepaid insurance premiums.
e. The ABS and RBA figures include an estimate of the value of consumer durables, whereas the HILDA Survey only includes vehicle values. The HILDA Survey, however, includes collectibles whereas the other two sources do not. The RBA estimates do not include business assets.
f. The National Accounts only reports financial year figures for non-financial assets. We have thus interpolated an estimate based on the figures reported for 30 June 2002 and 30 June 2003.

Sources: ABS data derived from Financial Accounts (cat. no. 5232.0), Table 15, and Australian System of National Accounts, 2002-03 (cat. no. 5204.0), Tables 16 and 51.

RBA data taken from *Statement of Monetary Policy*, various issues.

[19] Again, one possible explanation here would be over-sampling of persons from high-earning occupations.

measured, the HILDA Survey estimates are reasonably comparable to RBA estimates; they again are on the low side which again is exactly what would be expected.[20] The ABS estimates are much lower which mainly appears to reflect the marked difference in the way property and housing values are derived.

The final comparison relates to debts. Here the HILDA estimate is much (20 per cent) lower than the official sources. In retrospect the HILDA questionnaire may not have included enough questions on separate types of debt. We asked only about housing debt, business debt, HECS debt, credit and store cards, and 'other' debt. It might have been preferable to ask additionally about overdrafts (excluding housing), vehicle debt, hire purchase, gambling debts and so on (see Juster et al. 1999). Even so, there may be some irreducible tendency for respondents to under-report debt, partly for social desirability reasons. We also believe that relative to official sources, credit card debt may be understated in the HILDA Survey data. Those respondents who said they routinely paid up in the first month and so incurred no interest charges were recorded as having no credit card debt. By contrast, the official sources record card liabilities owed by the nation's households at one moment in time.

Overall, it is our assessment that the HILDA Survey has done a reasonable job in measuring total household wealth. Nevertheless, net worth is almost certainly overstated. This reflects both under-reporting of debts and, though more speculative, over-reporting of assets. On the latter point, it is worth bearing in mind that surveys employing equal probability samples invariably understate the wealth held by the very wealthy. Our wealthiest household, for example, had a reported net worth of $22 million, which is well below the levels recorded for individuals listed in the BRW list of Australia's 200 wealthiest people.

It also needs to be borne in mind that while these very broad-brush comparisons of aggregates suggest that the HILDA data are generating what appear to be highly plausible results, at a more disaggregated level the data may not be so convincing. Indeed, there is clear evidence for this with respect to business assets from the HILDA data itself. Specifically, we can compare answers in the HQ about the presence of household members who own businesses with those provided in the PQ about individual ownership of businesses. This comparison reveals that of the 2049 individual respondents who indicated in the PQ that they owned a business, 32 per cent were members of households where it was indicated that no one owned a farm or business. Such a large discrepancy is both difficult (impossible) to explain away and suggests that the data on both business income and business assets are likely to be subject to serious measurement error.

Finally, we again emphasise the main weakness associated with imputing missing data. Even if imputation does generate unbiased cross-section estimates, estimates of change (assuming the wealth module is repeated at some point in the future) are unlikely to be so well behaved.

[20] The lower RBA estimate reported in Table 10 is simply a function of the exclusion of business assets.

Other Data Issues

Quality of Person Matches

Ensuring the data for the same person is linked across waves is of great concern to a panel study. The items that collectively present a unique key between waves are name, sex and date of birth. Obviously some name changes are expected (typically women change their last name when they marry), but the interviewer is expected to ascertain that they are speaking with the same individual. There are, however, three sources of potential errors in the name, sex and date of birth information recorded during wave 1. First, this information was collected from one individual in the wave 1 household and that person may not have provided the correct details for others in the household (especially likely in share households). The most problematic item here is date of birth. To counteract this problem, the date of birth of each respondent was validated at the end of the wave 2 interview and corrected where necessary. The second source of error is that interviewers may make transcription errors. While little could be done about transcription errors in wave 1, the name, sex and date of birth information was pre-printed on the Household Form (HF) in wave 2. Where corrections were required, the pre-printed fields were crossed out and the revised information recorded. The third source of error is mistakes by data entry operators. In both waves, these data entry errors were minimised through the use of double data entry (where every form was entered by two different operators and discrepancies resolved).

A comparison of the first name, sex and date of birth for people enumerated in waves 1 and 2 is provided in Table 11. While name changes were required for 1.6 percent of respondents to both waves and 1.5 percent of other enumerated people, the vast majority of these were to correct spelling mistakes. Only 19 people had ‘major’ first name changes and all except for one were consistent between the waves for sex and date of birth. Sex was revised for 37 people, which corrected a number of anomalies identified in wave 1 between name and sex. The largest number of changes was made to the date of birth. For the enumerated people not responding in both waves, 2.1 percent had changes to the date of birth recorded. This figure jumps to 3.3 percent for people interviewed in both waves, reflecting the improvement in the date of birth information from the verification at the end of the personal interview. Only eight cases with mismatched name, age or sex required further investigation with ACNielsen.

Cross-form Matching

The interviewer-administered forms for each household were typically kept together, and no problems have been identified in matching the PQs and Household Questionnaires HQs with the appropriate HF. The SCQ forms, however, are quite different in that these forms are often returned by mail. Furthermore, the data entry for the SCQ occurred on a different system from the interviewer-administered forms, and hence the SCQs were separated from the rest of the forms for the household for processing.

The SCQ data thus had to be matched back to the person-level data collected in the PQ. To assist this process, the person identifier was written on the SCQ form and the serial number of the SCQ was written on the relevant PQ. One hundred and four of the 13,162 SCQs that were completed and returned could not be matched to a PQ. In wave 2 the rate of matching was slightly higher with only 83 forms not able to be matched out of 11,718 forms completed and returned.

Table 11: Matches on First Name, Sex and Date of Birth: Waves 1 and 2 Compared

Identifying information	*No changes*	*Minor changes*	*Major changes*
Interviewed in Wave 1 and 2 (N=11993)			
First name	11799 (98.4%)	181 (1.5%)	13 (0.1%)
Sex	11972 (99.8%)		21 (0.2%)
Date of birth	11601 (96.7%)	352 (2.9%)	40 (0.3%)
Not interviewed in both waves (N=5283)			
First name	5202 (98.5%)	75 (1.4%)	6 (0.1%)
Sex	5267 (99.7%)		16 (0.3%)
Date of birth	5174 (97.9%)	99 (1.9%)	10 (0.2%)

Notes: a All changes to sex are considered major.
b Minor changes to date of birth are defined as a change to any one component of day, month or year, a swap between day and month, or a change of plus or minus 1 to more than one of the components. All other changes to date of birth are considered to be major changes.
c People that were not interviewed in both waves, but for which data was collected in both waves include respondents to a single wave, non-respondents in both waves and children.

In wave 2 we also went to more effort to identify the reason for non-matches, and discovered that 27 of the non-matches were the result of interviewers giving a form to children who were not eligible for interview,[21] 45 were actually duplicate forms,[22] and 3 had no PQ to match to (presumably because the SCQ was distributed prior to a PQ interview being attempted). In other words, all but 8 of the non-matches result in no loss of information.

Problems arising from the inability to match the SCQ to the interview-based data thus appear to be of little significance. It is, however, still possible that some of our matches are incorrect matches. This might arise, for example, if respondents within households do not complete the exact form assigned to them (even though their first name is written on the SCQ prior to distribution). We have no way of gauging the extent of this problem in wave 1, but in wave 2 additional questions about age and sex were added to the SCQ instrument as a check on the quality of cross-form matches. The results of comparing the responses on sex with the information recorded in the HF suggest that up to 38 forms may have been incorrectly matched. After further investigation it was determined that 30 of these were cases of within-household form swapping, and hence the forms were re-assigned. For the remaining eight it was assumed that sex was incorrectly answered on the SCQ and the answers amended to be consistent with the HF. Evaluating matches on the basis of age is more difficult given we do not know on which date the SCQ is completed. A simple comparison suggests there may be problems with up to 212 forms. Nevertheless, given we do not know the exact date the SCQ

[21] Mainly 15 year olds who had turned 15 after 30 June.

[22] Duplicates can arise when, as result of a slow return, another form is dispatched to a sample member who then subsequently returns two completed forms.

is completed and given age could easily be incorrectly answered on the SCQ, no changes were made to the data. Users will thus be confronted by what might appear to be an internal inconsistency in the data.

Longitudinal Inconsistencies

The main purpose of longitudinal surveys is to collect data about change. For many data items this can be achieved by asking identical questions at different waves and then comparing responses. This, however, is not as straightforward as it might seem. There are, for example, no guarantees that respondents will respond in the same way in wave 2 as they did in wave 1. Even more problems arise when, in an effort to reduce respondent load, question sequences are limited to respondents who have experienced some change in their circumstances since the previous interview. This now introduces problems associated with recall error given respondents not only have to identify whether the they have experienced the type of change in question, but also whether that change occurred within the reference period.

The data from wave 2 of the HILDA Survey provide clear evidence of the inconsistencies that may arise as result of these sorts of reporting errors. Take, for example, changes in marital status. In wave 1 we collected information on marriages and inferred the respondent's marital status from the current or most recent marriage. A question on current marital status was also asked in wave 2. However, in wave 2 respondents were also asked if their marital status had changed since the date of the wave 1 interview and on what date that change occurred. There were 258 respondents who reported not having changed their marital status since wave 1, even though their reported marital status was different in the two waves. This represents just over half of all respondents that reported a different marital status in the two waves. In addition, there were 39 respondents who reported having changed their marital status since the date of the wave 1 interview, but their reported marital status was the same in each of these two waves.

A similar problem arises with the recording of address changes. Address changes can be identified through two sources in the HILDA data – comparison of actual addresses recorded on the HF in both waves or through a question asked of individual respondents in the PQ about whether they had changed address since the date of the last interview. According to the HF, of all individual respondents interviewed at both waves, 1915 (or 16 per cent) were living at a different address in wave 2. When asked as part of the PQ, however, 119 of these persons indicated that they had not changed address. Further, the PQ identifies another 141 movers even though the HF suggests no movement. Given recall problems with regard to the dating of moves are likely to affect the latter source, we place more weight on the mobility information derived from the HF. Nevertheless, this source is not perfect either. In particular, identical addresses can and are recorded differently each wave, which may lead us to infer movement when no such change has occurred. Also, some difference between these two sources is expected given many PQ interviews are conducted on dates well after the HF is completed.

As a third example, we can also consider changes in employment status. As shown in Table 12, of the wave 2 respondents who were employed in wave 1, 4.6 per cent did not recall being employed at the date of the wave 1 interview. Similarly, among those respondents who were recorded as not being employed in wave 1, almost 7 per cent reported during the wave 2 interview that they thought they had been in employment at the wave 1 interview date. Overall, about 5 per cent of respondents recalled their employment status, as reported in wave

1, incorrectly.[23] However, this error rate can obviously be expected to be relatively low for persons whose employment status has been very stable and thus, conversely, relatively high for those who have experienced changes in their employment state. Indeed, Table 12 indicates that of those persons whose employment state was different at the two interview dates, close to one-third recalled their employment situation differently to what was reported in wave 1.

Table 12: Inconsistencies between Actual Wave 1 Employment Status and Recalled Employment Status

	Employed in wave 2	*Not employed in wave 2*	*Total*
Employed in wave 1			
Recalled being employed in Wave 1	6,578 (97.5%)	427 (71.0%)	7005 (95.4%)
Recalled being not employed in Wave 1	166 (2.5%)	174 (29.0%)	340 (4.6%)
Not employed in wave 1			
Recalled being employed in Wave 1	231 (33.9%)	87 (2.2%)	318 (6.8%)
Recalled being not employed in Wave 1	450 (66.1%)	3880 (97.8%)	4330 (93.2%)

Another related area of the data collection where recall problems can be tested relates to the activity calendar, where information on spells of employment, unemployment and education are collected. This calendar covers the entire preceding financial year (year ended 30 June) plus all additional complete months up to the interview date (but not after 31 December). Given interviews in wave 2 could have been conducted at any time between August 2002 and March 2003, this implies calendars varying in length from 13 to 20 months. More importantly, part of the information collected in the wave 2 calendar will overlap with information collected in wave 1[24], and hence we are able to determine how well these calendars match across the two waves.

In Table 13, therefore, we report a summary of how well job spells recorded in wave 2 could be matched to job spells recorded in wave 1 during the overlapping period of the calendar. Of all persons who had at least one job since 1 July 2001, over 19 per cent provided information about job spells at the beginning of this period that was not consistent with the information provided in the previous wave. Some of these errors were the result of minor recall problems in determining the date when jobs ceased and began. However, in almost 15 per cent of cases the job spells recorded in one calendar simply do not appear in the same period in the other calendar. Furthermore, most of the exact matches are for persons who have been in the same

[23] Not all of these inconsistencies are due to recall errors. Employment status in wave 1 is a derived variable based on International Labour Office (ILO) definitions, and the ILO definition of paid employment will not necessarily accord with what individual respondents think of as employment.

[24] The length of the overlap could vary from anywhere from one month to 5 months.

Table 13: Quality of Calendar Matches

	Persons with at least 1 job since 1 July 2002		*Persons with at least 1 job change since 1 July 2002*	
	No.	*%*	*No.*	*%*
All job spells matched	6607	80.6	1270	44.3
All job spells matched within 1 month	151	1.8	151	5.3
All job spells matched within 3 months	59	0.7	59	2.1
All job spells matched, but error on at least one match exceeded 3 months	173	2.1	173	6.0
At least one job spell cannot be matched	1212	14.8	1212	42.3
TOTAL	8202	100.0	2865	100.0

job for long periods of time. When we restrict our attention to persons who either left or started a job spell since 1 July 2002, the proportion of non-matched job spell data rises to 42 per cent.

Derived Variables

Tax Derivation

A potentially serious flaw in data release 1.0 was the inclusion of a set of derived after-tax income variables that were overly simplistic. Specifically, they only implemented two aspects of the tax code – the marginal rates of income taxation and the Medicare levy rules. As outlined in Headey (2003), this method overestimates taxes by a large margin, and will thus understate disposable income. Release 2.0 provides a completed revised set of after-tax variables for both waves that attempt to provide a more realistic assessment of individual tax burdens. Since the procedure is discussed in much more detail in Headey (2003), including the provision of comparisons with tax office statistics, no more is said here.

Treatment of Owner Managers of Incorporated Businesses

Another flaw in the data release 1.0 was the treatment of wage and salary income from incorporated businesses as part of business income. This is inconsistent with ABS concepts and all wave 1 data have been revised in version 2.0. Wages earned from the respondent's own incorporated business are now treated like all other wage and salary income.

Geographic Identifiers

For wave 1, area information was derived from CD level whereas for wave 2 the most detailed geographic unit provided in the data file supplied by the data collection agency (ACNielsen) is postcode. This has had two effects. First, the geographic data is of a lesser quality in wave 2. This is a problem that cannot be overcome without additional expenditure to enable ACNielsen to devote the resources to geo-coding all address information. Second, derived variables based on area, such as the remoteness scale and the SEIFA scores, are not constructed on an equivalent basis for the two waves. This problem can be rectified by reconstructing the wave 1 versions using the post-code information. We expect that such variables will be included in the next data release (to be issued when the wave 3 data become available).

Other Issues

There were a number of other minor problems in the way some variables were derived in wave 1. Users were notified of these errors as part of the version 2.0 data release. The affected variables were: three labour market history variables (aehtjb, aehtuj, aehto); variables constructed from questions about the amount of contact non-resident parents have with their children; the educational attainment variable for nurses with diplomas; family type (ahhfty); time lived in first de facto relationship (aordflt); and one variable derived from the calendar (acapnlf). In almost cases, the affect of the errors was quite small.

Questionnaire Design Problems

A small number of mostly minor errors were identified with the wave 2 questionnaires that would require amendment prior to the next data release. The most significant of these were:

- HF: Reference to lodgers on page 2 should be to boarders.
- HQ – Childcare grids: Need to allow for a “me or my partner” option in all cases.
- PQ: Date for financial year income was incorrect (but corrected in training).
- PQ – Children living in household: Refers to children living in ‘dwelling’ when reference should be to the ‘household’.
- PQ – date changed address: We asked movers when they moved to their current address, but we did not ask when they left their previous address. Thus for people who move twice in a year we do not know exact length of tenure at former address.
- NPQ – Duration of previous de facto relationship: Inconsistency with wave 1. In wave 1 this question concerned the first de facto relationship, but in the NPQ it now concerns the last (prior to the current relationship).

References

Collins, M., Sykes, W., Wilson, P. and Blackshaw, N. (1988), 'Nonresponse: The UK Experience', in R.M. Groves, P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II and J. Waksberg (eds), *Telephone Survey Methodology*, John Wiley and Sons, New York.

Fitzgerald, J., Gottschalk, P. and Moffitt, R. (1998), 'An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics', *Journal of Human Resources* 33, 251-99.

Frankel, J. and Sharp, L.M. (1981), 'Measurement of Respondent Burden', *Statistical Reporter* 81, 105-111.

Headey, B. (2003), 'How Best to Imopute Taxes and Measure Public Transfers?', HILDA Project Discussion Paper Series No. 2/03, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Juster, F.T., Smith J.P. and Stafford F. (1999), 'The Measurement and Structure of Household Wealth', *Labour Economics* 6, 253-76.

Laurie, H., Smith, R. and Scott, L. (1999), 'Strategies for Reducing Nonresponse in a Longitudinal Panel Survey', *Journal of Official Statistics* 15, 269-282.

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. (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. and Wooden, M. (2004), 'Wave 2 Survey Methodology', HILDA Project Technical Paper Series No. 1/04, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Appendix 1: Item Non-response Rates, Waves 2
(questions where incidence of missing cases exceeds 2% of expected N)

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
Household Questionnaire				
Q4a	Difficulty finding good quality childcare	7.5	982	The high incidence of missing cases reported here is misleading. A "don't know / not applicable" option was provided. As a result, we cannot separate don't know answers from cases where the question does not apply. For example, 78% are missing on Q4h, not because the respondent does not know, but because their child is not a special needs child.
Q4b	Difficulty finding the right person to take care of child	6.1	982	
Q4c	Difficulty getting care for hours needed	4.3	982	
Q4d	Difficulty finding care for a sick child	17.3	982	
Q4e	Difficulty finding care during school holidays	21.6	982	
Q4f	Difficulty with the cost of child care	7.9	982	
Q4g	Difficulty juggling multiple childcare arrangements	29.4	982	
Q4h	Difficulty finding care for a special needs child	78.1	982	
Q4j	Difficulty finding a place at the childcare centre of choice	26.2	982	
Q4k	Difficulty finding a childcare centre in the right location	23.8	982	
Q4m	Difficulty finding care my child is happy with	9.1	982	
Q7	Childcare total cost for all school-age children during term time	6.0	369	
Q15	How is child care payment made	3.2	526	
R9b	Per cent of property non HH member owns	5.8	189	
R12	Price of home when purchased	5.2	4944	
R13	Approximate value of home today	7.6	4944	
R15	How much was home loan originally	7.0	3481	
R17	Approximate outstanding on home loan	4.9	2101	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
R20	When expect to have home loan paid off	15.7	2101	
R22	Amount of other loan still owed	3.5	343	
R24	How much is left to pay on second loan	2.8	424	
R25a	How much is usual repayment on second loan	8.2	417	
R28	How much weekly rent would have to pay if renting	20.0	220	
R31	Approximate value of other property today	4.2	1222	
R35	Total amount of debt on other property	6.1	575	
S3	Year investments first acquired	5.0	3859	
S5	Total current value of all investments	15.0	2978	
S7	Total current value of trusts	27.9	390	
S8b	Per cent of trust fund for benefit of household	5.6	390	
S10b	Amount in children's bank accounts	5.8	1399	
S12	Current value of other investments for all household members	8.3	241	
S14	Value of businesses	18.3	1090	
S16	Amount of debt as result of business	9.0	1090	
S22	Current worth of other vehicles	5.8	310	
S24	Current value of insurance policies	24.1	794	
S26	Current value of other assets	13.5	1068	

Person Questionnaire				
C7b	Hours worked per week on average in main job (if has more than one job and hours vary from week to week in main job)	5.9	34	Very small sample. The item non-response rate for the same question for persons with one job is very low.
C18	Hours worked per week worked at home (if hours vary)	7.5	134	Small sample size.

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
C18	Hours worked per week worked at home (if hours vary)	7.5	134	
C33	How many people work at locations throughout Australia	6.5	4479	But most respondents were able to tell us whether firm size was more or less than 100.
D15	If offered suitable job, could start work in next four weeks	4.3	1141	
D16	Lowest acceptable wage per hour	9.4	1773	Difficult concept.
D18	Percent chance of finding suitable job in next 12 months	2.0	1773	Difficult concept.
F3	Total gross amount of most recent pay before deductions	7.4	7323	72.0 per cent of these knew their net pay.
F7	Total amount of your most recent gross pay for all other jobs	11.9	681	32.0 per cent of these knew their net pay.
F10	Gross or net pay from workers compensation, accident insurance or sickness	9.9	91	
F17b	Amount of latest payment - War Widows Pension	2.5	79	
F17b	Amount of latest payment - Other Government pensions/benefits	3.9	127	
F19	Gross wage income last financial year (LFY)	7.2	7764	13.4 per cent of these knew their net income.
F24	Gross wage income from incorporated businesses LFY	17.7	560	
F25b	Dividend income from incorporated businesses LFY	19.4	108	
F26b	Total share of profit or loss from non-incorporated businesses LFY	26.1	1402	
F28b	Dividend income LFY	18.5	3196	
F29b	Royalty income LFY	11.9	67	
F29d	Dividend income LFY	14.4	3498	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
F30b	Profit or loss from rental income	13.9	1357	
F32	Mature Age Allowance	4.8	62	Small samples.
F32	Wife Pension	2.1	47	
F32	Sickness allowance	2.9	35	
F32	War widows pension	3.7	82	
F32	Special benefit	3.4	29	
F32	Austudy	2.1	140	
F32	Parenting payment	2.9	706	
F32	Foreign government pensions	3.6	223	
F32	Other pensions (specify)	6.7	90	
F33b	Annuity income FY	3.7	751	
F33b	Child support income FY	4.9	349	
F33b	Workers compensation etc. income FY	9.8	143	
F33b	Inheritance / bequests FY	3.7	189	
F33b	Income from parents FY	3.3	602	
G5a	Regular child support payments	3.8	450	
G7a	Other child related expenses	10.6	396	
G15d	Distance of non-resident parent (first in grid)	6.4	832	
G18a	Regular child support received	5.6	338	
G20a	Receipt of income for other child-related expenses	17.9	117	
CPQ :H11b	Month when stopped living with former partner	2.0	245	
CPQ: H13a	Year started living together in most recent relationship	5.8	52	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
CPQ: H13b	Month stopped living together in most recent relationship	5.9	51	
CPQ: H13b	Year stopped living together in most recent relationship	7.7	52	
J2	Total amount held in bank accounts	4.3	9174	
J7a	Per cent share of joint accounts	3.8	104	
J7b	Per cent share of joint accounts held by other in household	4.8	104	
J11	Maximum borrowing limit on all credit cards in own name	2.1	5245	
J13	Amount owing on own credit cards	3.4	2129	
J16	Maximum borrowing limit on all joint credit cards	3.8	2976	
J18	Amount owing on joint credit cards	9.5	907	
J21	Amount owed on HECS or student loans	10.2	1034	
J24	Value of all other personal debts	2.1	2979	
J26	Last month's repayments on other personal debts	2.7	2728	
J29	Total value of capital in all superannuation funds (if retired)	20.0	671	
J32	Member of the Commonwealth or Public Superannuation Schemes	9.2	336	
J34	Employer contribution to superannuation funds as per cent of wages/salary	22.3	6049	36.5 per cent of these knew the absolute dollar amount.
J37	Personal contribution to superannuation funds as percentage of wage/salary	28.3	1314	87.6 per cent of these knew the absolute dollar amount.
J38b1	Personal contributions to superannuation fund ($)	3.1	416	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
J41a	Personal contributions to own superannuation fund ($)	22.5	846	
J42	Per cent of income contributed to own superannuation fund	55.4	846	
J45a	Partners contributions to superannuation fund	31.0	116	
J46	Estimated value of all super funds	8.0	9987	
J48	Type of largest super fund	18.8	8162	
K9	Month moved to this address	3.8	2903	
T5b	When expect to live at new address	43.2	2158	This question is about the future, so a don't know answer is an entirely valid response.
NPQ BB4	How old were you at the time your parents first separated	4.1	296	
NPQ BB12	Father's occupation	2.1	985	
NPQ BB13	Father was unemployed for 6 months or more while you were growing up	6.0	1005	
NPQ BB14	Mother in paid employment when you were 14	2.2	1048	
NPQ BB15	Mothers occupation	5.4	875	
NPQ D23	How long since last worked for pay	2.7	74	
NPQ D27a	How much paid in last job before tax	11.9	134	
NPQ H5a	Month - Present or most recent marriage	5.6	323	
NPQ H5c	Lived together - First marriage if married more than once	4.3	138	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
NPQ H5f	Year separated (separated/divorced) - Present or most recent marriage	7.5	93	
NPQ H5f	Year separated (separated/divorced) - First marriage if married more than once	11.3	53	
NPQ H7b	Month began living with current partner	6.3	208	
NPQ H13	Year began living together	2.6	192	
NPQ H14	Length lived together as a couple	9.8	192	
Self Completion Questionnaire (matched sample only)				
A1	General health	2.7	11636	
A2	Health compared to one year ago	2.5	11636	
A3a	Physical functioning: Whether health limits vigorous activities	3.8	11636	
A3b	Physical functioning: Whether health limits moderate activities	3.3	11636	
A3c	Physical functioning: Whether health limits lifting or carrying groceries	3.4	11636	
A3d	Physical functioning: Whether health limits climbing several flights of stairs	3.7	11636	
A3e	Physical functioning: Whether health limits climbing one flight of stairs	4.2	11636	
A3f	Physical functioning: Whether health limits bending kneeling or stooping	3.6	11636	
A3g	Physical functioning: Whether health limits walking more than one kilometre	3.5	11636	
A3h	Physical functioning: Whether health limits walking half a kilometre	3.9	11636	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
A3i	Physical functioning: Whether health limits walking 100 metres	4.0	11636	
A3j	Physical functioning: Whether health limits bathing or dressing yourself	3.3	11636	
A4a	Role-physical: Cut down the amount of time spent on work or other activities	3.5	11636	
A4b	Role-physical: Accomplished less than would like	3.5	11636	
A4c	Role-physical: Were limited in the kind of work	3.9	11636	
A4d	Role-physical: Had difficulty performing work or other activities	3.6	11636	
A5a	Role-emotional: Cut down the amount of time spent on work/other activities	3.5	11636	
A5b	Role-emotional: Accomplished less than would like	3.4	11636	
A5c	Role-emotional: Didn't do work or other activities as carefully as usual	3.9	11636	
A6	Role-emotional: Health has interfered with social activity	2.3	11636	
A7	Physical functioning: Had bodily pain in last 4 weeks	2.2	11636	
A8	Physical functioning: Pain interfered with normal work	2.3	11636	
A11a	Physical functioning: Get sick a little easier than others	2.3	11636	
A11c	Physical functioning: Health expectations	2.2	11636	
A11d	Physical functioning: Health is excellent	2.1	11636	
B3	Cigarettes smoked	5.3	2858	
B9a	Neighbourhood: Neighbour helping each other out	9.0	11636	Specific DK option provided.

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
B9b	Neighbourhood: Neighbours doing things together	11.1	11636	Specific DK option provided.
B9c	Neighbourhood: Traffic noise	2.3	11636	
B9d	Neighbourhood: Noise from airplanes, trains or industry	2.1	11636	
B9e	Neighbourhood: Homes and gardens in bad condition	3.7	11636	
B9f	Neighbourhood: Rubbish and litter	2.1	11636	
B9g	Neighbourhood: Teenagers hanging around on the streets	2.8	11636	
B9h	Neighbourhood: People being hostile and aggressive	4.0	11636	
B9i	Neighbourhood: Vandalism and deliberate damage to property	3.7	11635	
B9j	Neighbourhood: Burglary and theft	8.5	11636	
B11a	Satisfaction with: Partner	2.1	8379	Some of the non-response is due to people for whom the question did not apply not checking the NA box.
B11b	Satisfaction with: Children	2.3	7928	
B11c	Satisfaction with: Partners relationship with children	3.4	6678	
B11d	Satisfaction with: Relationship with step children	24.2	1653	
B11e	Satisfaction with: Children in HH get along with each other	5.8	5075	
B11f	Satisfaction with: Relationship with parents	3.5	8208	
B11g	Satisfaction with: Relationship with step parents	21.9	1958	
B11h	Satisfaction with: Relationship with most recent former spouse/partner	9.4	3638	
B17a	Hours per week spent on: Paid employment	3.3	11636	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
B17b	Hours per week spent on: Travelling to / from work	4.0	11636	
B17c	Hours per week spent on: Household errands	3.5	11636	
B17d	Hours per week spent on: Housework	2.8	11636	
B17e	Hours per week spent on: Outdoor tasks	3.1	11636	
B17f	Hours per week spent on: Playing with your children	5.8	11636	
B17g	Hours per week spent on: Playing with other children	6.5	11636	
B17h	Hours per week spent on: Volunteer / charity work	6.6	11636	
B17i	Hours per week spent on: Caring for disabled family	7.5	11636	
C3a	Difficulty in raising $2000	2.2	11636	
C4	Savings habits	2.3	11636	
C5	Current reason for saving	2.5	8332	
C6	Savings time horizon	3.1	11636	
C7	Willingness to take financial risk	2.1	11636	
C8a	Attitudes to borrowing: For a holiday	7.6	11636	
C8b	Attitudes to borrowing: For living expenses	6.0	11636	
C8c	Attitudes to borrowing: For clothes or jewellery	9.2	11636	
C8d	Attitudes to borrowing: For car	5.6	11636	
C8e	Attitudes to borrowing: For education	7.5	11636	
C9b	Household decision maker for decisions on large household purchases	3.8	11636	
C9c	Household decision maker for decisions on savings, investments and borrowing	4.3	11636	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
D1	Is in paid work	3.0	11636	
D3a	Work entitlements: Paid maternity leave	26.5	6933	Specific don’t know options provided. Further, it is entirely reasonable to expect that many workers will not know whether all of entitlements (especially if they have never used them themselves) are available in their workplace.
D3b	Work entitlements: Unpaid maternity leave	30.3	6933	
D3c	Work entitlements: Parental leave	32.5	6933	
D3d	Work entitlements: Special leave for caring for family members	26.6	6933	
D3e	Work entitlements: Permanent part-time work	17.5	6933	
D3f	Work entitlements: Home-based work	18.1	6933	
D3g	Work entitlements: Flexible start finish times	10.6	6933	
E1	Has parenting responsibilities for child < 18	3.1	11636	
E4a	Working and family: Makes me a more rounded person	2.4	2882	
E4b	Working and family: Gives my life more variety	2.3	2882	
E4c	Working and family: Makes me feel competent	2.4	2882	
E4d	Working and family: Have to turn down work/opportunities	2.4	2882	
E4e	Working and family: Time working less enjoyable/more pressured	2.4	2882	
E4f	Working and family: Miss out on home/family activities	2.5	2882	
E4g	Working and family: Family time less enjoyable/more pressured	2.4	2882	
E4h	Working and family: My work has a positive effect on my children	2.4	2882	
E4i	Working and family: Better appreciate time spent with children	2.6	2882	

Form / Qstn #	*Variable*	*Missing cases (%)*	*Expected N*	*Notes*
E4j	Working and family: Working makes me a better parent	2.4	2882	
E4k	Working and family: Worry about children while at work	2.5	2882	
E4l	Working and family: Too little time or energy to be good parent	2.4	2882	
E4m	Working and family: Miss out on the rewarding aspects of being parent	2.4	2882	