FACULTY OF
BUSINESS &
ECONOMICS

HILDA PROJECT TECHNICAL PAPER SERIES
No. 2/12, December 2012

Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey

Nicole Watson

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

Acknowledgements

This paper uses near final Release 11 data of the Household, Income and Labour Dynamics in Australia (HILDA) Survey, a project initiated and funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs (FaHCSIA) and managed by the Melbourne Institute of Applied Economic and Social Research. The work has been supported by funding for the HILDA Survey from FaHCSIA, and in part, by funding from an Australian Research Council (ARC) Discovery Grant (#DP1095497).

The findings and views reported in this paper, however, are those of the author and should not be attributed to either FaHCSIA, the ARC, or the Melbourne Institute.

Contents

INTRODUCTION **1**

SURVEY METHODOLOGY **1**

SUMMARY OF THE SAMPLE DESIGN 1
FOLLOWING RULES 2
WHO ARE INTERVIEWED? 2
EVOLUTION OF THE POPULATION 2
EVOLUTION OF THE SAMPLE 2
WHO ARE STRUCTURALLY MISSING? 5
RESPONSE RATES 5

GENERAL STEPS IN WEIGHTING **8**

CROSS-SECTIONAL WEIGHTS FOR WAVE 1 ..... **9**

HOUSEHOLD AND ENUMERATED PERSON WEIGHTS 9
RESPONDING PERSON WEIGHTS 12

LONGITUDINAL WEIGHTS **13**

LONGITUDINAL RESPONDING PERSON WEIGHTS 13
LONGITUDINAL ENUMERATED PERSON WEIGHTS 15

CROSS-SECTIONAL WEIGHTS FOR WAVES 2 TO 10 ..... **15**

HOUSEHOLD AND ENUMERATED PERSON WEIGHTS 17
RESPONDING PERSON WEIGHTS 19

CROSS-SECTIONAL WEIGHTS FOR WAVE 11 (INTEGRATION OF ORIGINAL AND TOP-UP SAMPLES) **19**

COMMON STEPS IN WEIGHTING PROCESS COMPARED TO EARLIER WAVES 19
CLASSIFY POPULATION OVERLAP 21
COMBINE SAMPLES 24
CAUTION IN USING CROSS-SECTION WEIGHTS TO PRODUCE A SERIES OF ESTIMATES 25

COMPARISON OF HILDA AND ABS CROSS-SECTIONAL ESTIMATES FOR 2011 ..... **26**

WEIGHTS IN THE HILDA DATA RELEASE **30**

WEIGHTS PROVIDED 30
PLAN FOR INCLUSION OF THE TOP-UP SAMPLE INTO THE WEIGHTS IN FUTURE RELEASES 31
WHICH WEIGHT TO USE 32
CALCULATING STANDARD ERRORS 33

REFERENCES **35**

APPENDIX 1: COMPARISON OF DESIGN-ADJUSTED HILDA ESTIMATES TO ABS ESTIMATES **36**

APPENDIX 2: RESPONSE MODELS **40**

APPENDIX 3: COMPARISON OF INTEGRATION OPTIONS FOR WAVE 11 CROSS-SECTIONAL WEIGHTS **46**

INTEGRATION METHODS 46
PANEL ALLOCATION AND ADJUSTMENT FACTORS 47
EVALUATION OF INTEGRATION OPTIONS 49

Introduction

Weights are used to make inferences about the population from a sample. They adjust for unequal probabilities of selection and for non-response. Data users will typically use them in tabulations or summary statistics and they may sometimes use them in regressions.

This paper describes the weighting methodology for the HILDA Survey sample. The HILDA Survey is a longitudinal household-based panel study that follows individuals over time. It began in 2001 with 7682 responding households and, in 2011, the sample was extended through the recruitment of an additional 2153 responding households. Annual interviews are conducted with all people aged 15 and over and one person also answers questions about the household as a whole. A series of longitudinal and cross-sectional weights are provided on the datasets.

We begin with a brief overview of the sample design, then examine how the sample changes over time and reflect on the response rates achieved over the first 11 waves of the HILDA Survey. The general steps in the weighting process are then described along with how these apply to the cross-sectional weights in wave 1, the longitudinal weights, and the cross-sectional weights in waves 2 to 10. The cross-sectional weights for wave 11 integrates the original ('main') sample with the top-up sample. We compare how the wave 11 cross-section matches estimates from several surveys conducted by the Australian Bureau of Statistics (ABS). The paper concludes with a description of the weights provided on the datasets and provides some advice on using these weights.

Survey methodology

Summary of the sample design

The original HILDA Survey sample was selected in 2001 via a stratified three-stage clustered design (see Watson and Wooden (2002) for details). The sample was restricted to households living in private dwellings, excluding very remote parts of Australia. It was stratified by state and within the five most populous states by metropolitan and non-metropolitan areas. In the first stage of selection, 488 Census Collection Districts (CCDs) were selected with probability proportional to the number of (occupied and unoccupied) dwellings.[1] The CCDs were sorted in a serpentine order and selected systematically to ensure the sample had a wide spread across Australia. A list of the dwellings in each of these CCDs was constructed and a sample of approximately 25 dwellings was systematically selected (with a random start). For five of the very large and remote CDs a number of blocks were systematically selected prior to the dwelling listing process. When the interviewer approached the dwelling and found more than three households living there, a random sample of three households was chosen.

The top-up sample was selected in 2011 using a similar design as the 2001 sample with 125 CCDs selected (see Watson, 2011). There are three small differences in the two designs. First, the boundaries used for the CCDs in the top-up sample were the 2006 Census boundaries.[2] Second, the size measure for the CCDs for the top-up sample was the total number of occupied dwellings (rather than occupied and unoccupied as was done for the original sample) which will slightly reduce the variability in the design weights. Third, the top-up sample was not stratified due to the smaller sample size involved but the systematic selection was ordered according to state and, within the five most populous states, by major statistical region. This will have a similar effect.

[1] The CCD boundaries used for the selection of the 2001 HILDA sample are those used for the 1996 Census.

[2] While the 2011 CCD boundaries were available as of December 2011, the information about the number of dwellings in each CCD did not become available until around mid 2012, so it was not possible to use the 2011 CCD boundaries.

Following rules

The original sample has evolved over time with household structure changes: some individuals move out to form their own households, others move overseas or die, and other individuals move in, or are born. The following rules adopted in the HILDA Survey are intended to ensure the sample mimics the changes in the population as much as possible and allows for the study of family dissolution. All members of the responding households in 2001 are considered Permanent Sample Members (PSM) and these people are followed over time, even if they move into non-private dwellings or very remote parts of Australia. In addition, others are converted to PSM status if they are:

- born to or adopted by a PSM;
- the other parent of a PSM birth or adoption if they are not already a PSM;
- recent arrivals to Australia since the survey began in 2001.[3]

All other sample members are Temporary Sample Members (TSMs), and are considered part of the sample for as long as they share a household with a PSM.

Who are interviewed?

Each wave, we aim to interview all adults (aged 15 and over at the 30th June preceding the interview) living in a household with a PSM. There are three specific cases that are worth clarifying at this point. First, if a child PSM moves out of the household without an adult PSM, we will seek to interview the adult TSMs that live with the child PSM. Second, if a PSM moves into an institution (such as a nursing home, or staff quarters) or very remote parts of Australia we will seek to continue to interview them. And third, we only interview people who are living in Australia: if a PSM moves overseas we keep in touch with them so that if they return then we can resume interviewing them.

Evolution of the population

Over the last 10 years, the Australian population has changed in a number of ways. Some people have died, emigrated from Australia, moved into institutions, or moved into very remote parts of Australia. Others have been born, immigrated to Australia, moved out of institutions, or moved out of very remote parts of Australia. There have also been changes in how these individuals collect themselves into households, with some households merging, others splitting, and some doing both.

Evolution of the sample

The HILDA sample evolves over time due to the following rules, population changes, household changes, and sample attrition. It is relevant for users to understand how the sample has evolved when using the data.

Table 1 shows the evolution of households in the main sample between waves 1 and 11 over time. Some key points about the sample of households include:

- The number of split and empty households has been fairly stable in recent waves as the number of households issued to field has stabilised.
- We have lost contact with 525 households and all tracking attempts have been exhausted with at least 455 of them.
- The number of responding households has been increasing in recent waves as the number of dead, empty or newly non-responding households has not exceeded the number of household splits.

[3] The inclusion of recent arrivals (i.e., immigrants who arrived to Australia after 2001) into our following rules occurred in wave 9 and was applied retrospectively. There were some recent arrivals who entered the sample in earlier waves but had moved out by wave 9 so could not be followed.

In a similar fashion, Table 2 shows the evolution of the sample of individuals over time. In wave 1, we began with 19,914 people who were part of responding households and these people form the basis of the sample followed over time. We note the following with respect to wave 11:

- We have added 3482 Permanent Sample Members to the HILDA sample. The great majority of these conversions are births to original Permanent Sample Members.
- Over half of the Temporary Sample Members who have joined the sample for one or more waves have since left.
- 937 of our sample members have died and 557 have moved overseas.
- Relatively few of our sample members have moved into non-private dwellings.
- Similarly relatively few have moved to very remote parts of Australia (we send a face-to-face interviewer to areas that were included in wave 1 if the sample is large enough to warrant this, otherwise the contact is made via the phone).
- We have observed 2457 births into the sample (some belong to Temporary Sample Members but the vast majority are be to Permanent Sample Members).
- We have identified 239 adults who are recent arrivals to Australia (i.e. born overseas and arrived in Australia for the first time after 2001) and they have 31 children.
- The main reason sample members are not issued to field is because of adamant refusals, though 768 sample members have been lost to tracking efforts.
- 69 per cent of the Permanent Sample Members and active Temporary Sample Members (i.e. known not to have left the household of a PSM) were part of a responding household in wave 11.
- There has been a decrease in the percentage of children in responding households over time, with 24 per cent of people in responding households being children in wave 1 compared to 20 per cent in wave 11. This is due to greater propensity for people in households without children to split to new households (13 per cent of people in households without children split into new households in wave 2, compared to 9 per cent of those with children).

Table 1: Composition of main household sample

	Wave										
	1	2	3	4	5	6	7	8	9	10	11
Eligible households											
Households from previous wave	-	7682	8368	8764	9037	9300	9584	9789	9995	10281	10526
Plus split households	-	712	466	371	388	394	321	350	405	380	368
Less dead or empty	-	26	70	98	125	110	116	144	119	135	121
Less households overseas	-	42	85	150	169	241	288	304	316	321	333
Total	11693	8326	8679	8887	9131	9343	9501	9691	9965	10205	10440
Outcomes											
Not issued to field	-	-	400	808	1079	1444	1785	1970	2062	2216	2344
Not issued as lost	-	-	221	279	359	399	425	438	435	441	455
Lost to tracking	-	250	146	119	79	73	49	60	103	76	70
Responding	7682	7245	7096	6987	7125	7139	7063	7066	7234	7317	7390

Note: When a household is no longer issued to field, we keep the same structure as the last issued wave so splits or empty households only occur in the issued sample.

Table 2: Composition of main individual sample

	Wave										
	1	2	3	4	5	6	7	8	9	10	11
All sample members	19914	21045	22062	22958	23903	24852	25702	26523	27518	28530	29489
Original Permanent Sample Members	19914	19914	19914	19914	19914	19914	19914	19914	19914	19914	19914
Converted Permanent Sample Members	-	232	496	768	1075	1393	1764	2182	2593	3027	3482
Active Temporary Sample Members[1]	-	899	1323	1529	1705	1899	2002	1972	2244	2417	2529
Inactive Temporary Sample Members	-	-	329	747	1209	1646	2022	2455	2767	3172	3564
Sample changes[2]											
Deceased	-	68	174	293	397	491	579	683	767	843	937
Moved overseas	-	74	233	374	387	430	483	501	491	513	557
Moved into non-private dwelling[3]	-	26	38	39	41	44	48	52	58	75	113
Moved into very remote Australia[3]	75	94	98	107	115	116	114	125	115	110	114
Births	-	219	450	674	920	1157	1413	1661	1926	2184	2457
Recent arrivals aged 15+[4]	-	13	26	43	53	85	85	109	151	187	239
Recent arrivals aged 0-14[4]	-	2	5	6	4	9	8	17	22	29	31
In responding HH											
Responding adult	13969	13041	12728	12408	12759	12905	12789	12785	13301	13526	13603
Non-resp. adult	1158	978	873	913	812	792	800	785	706	729	749
Child	4787	4276	4089	3888	3897	3756	3691	3574	3623	3600	3601
Not issued to field											
Lost	-	-	290	382	488	544	590	611	614	759	768
Permanent refusal	-	-	322	1080	1263	2093	2680	3124	3338	3608	3828
Permanent illhealth	-	-	19	27	61	91	113	26	44	100	131
Permanently overseas	-	-	60	74	129	198	283	318	280	145	251
Child (of one of the above)	-	-	158	307	363	447	503	521	518	502	460
Child permanently overseas	-	-	10	16	25	38	52	53	47	48	50

Note: 1. Active TSMs includes all TSMs not known to have left PSM households. TSMs in non-responding or not issued households may no longer belong to the household, but we are also not picking up any new entrants to these households. Inactive TSMs are TSMs known to have left the household of a PSM.
2. Excludes inactive TSMs.
3. Information has been carried over for non-responding and not issued households.
4. As recent arrivals were not followed if they left the household of PSM in waves 1 to 8, the number could go down (this occurs for children in waves 5 and 7).

Who are structurally missing?

The population has evolved in a number of ways that the following rules cannot emulate. In particular, the population now includes i) immigrants permanently settling in Australia since 2001; ii) long-term visitors arriving since 2001; iii) Australians not in Australia in 2001 who have since returned from overseas; iv) people who have moved out of non-private dwellings; v) people who have moved out of very remote Australia; and vi) Australian-born children of these groups. It is estimated that these groups form about 7 per cent of the Australian population in 2011, with permanent immigrants being by far the largest missing group.

The lack of recent immigrants was a motivating factor for the inclusion of the top-up sample in 2011. A number of options were canvassed for this top-up sample (see Watson, 2006) and ultimately it was decided that a general top-up sample would be added. A general top-up sample not only allows for the new portion of the population to be represented, but it will also increase the sample size for some analyses going forward and permits the study of the impact of non-response and attrition on our main sample.

Response rates

Recruitment of new sample in wave 1 and wave 11

Table 3 and 4 show the fieldwork outcomes for the recruitment of the original sample in wave 1 and the top-up sample in wave 11. A household response rate of 66 per cent was obtained in wave 1 (see Table 3). This rate was exceeded in wave 11 where we obtained a household response rate of 69 per cent. We attribute this increase to the experience the fieldwork team has gained over the past 10 years and the longer fieldwork period for wave 11 (28 weeks in wave 11 compared to 21 weeks in wave 1). Within the responding households, individual interviews were obtained with the vast majority of adults. In wave 1, 92 per cent of the adults provided an interview and in the wave 11 top-up sample this rate increased to 94 per cent (see Table 4).

Table 3: Household outcomes for new sample, wave 1 and wave 11 top-up compared

	Wave 1		*Wave 11 Top-Up*	
Sample outcome	*Number*	*%*	*Number*	*%*
Addresses issued	12,252		3,250	
Less out-of-scope (vacant, non-residential, foreign)	804		212	
Plus multi-households additional to sample	245		79	
Total households	*11,693*	*100.0*	*3,117*	*100.0*
Refusals to interviewer	2,670	22.8	885	*28.4*
Refusals to fieldwork company (via 1800 number or email)	431	3.7	16	*0.5*
Non-response with contact	469	4.0	16	*0.5*
Non-contact	441	3.8	47	*1.5*
Fully responding households	6,872	58.8	1,963	*63.0*
Partially responding households	810	6.9	190	*6.1*
Total responding households	*7,682*	*65.7*	*2,153*	*69.1*

Table 4: Person outcomes for new sample, wave 1 and wave 11 top-up compared

	Wave 1		*Wave 11 Top-Up*	
Sample Outcome	*Number*	*%*	*Number*	*%*
Enumerated persons	19,914		5,451	
Ineligible children (under 15)	4,787		1,171	
Eligible adults	*15,127*	*100.0*	4,280	*100.0*
Refusals to interviewer	597	3.9	228	5.3
Refusals to fieldwork company (via 1800 number or email)	31	0.2	0	0.0
Non-response with contact	218	1.4	23	0.5
Non-contact	312	2.1	20	0.5
Responding individuals	13,969	92.3	4,009	93.7

Main sample in waves 2 to 11

A common measure of the re-interviewing success is the re-interview rate, calculated as the percentage of respondents in the previous wave that provide an interview in the current wave, excluding those that are out of scope (that is, those that have died or moved overseas). As shown in Table 5, this re-interview rate has increased from 86.8 per cent in wave 2 to 96.5 per cent in wave 11.

In terms of how this re-interview rate compares to other studies, Figure 1 shows the HILDA experience (black line) against that of the British Household Panel Study (BHPS) (dark grey solid line and dark grey dashed line) and the German Socio-Economic Panel (SOEP) (light grey solid line and light grey dashed line). The BHPS started in 1991 and two response rates have been provided to demonstrate the effect of the inclusion of proxy interviews and short telephone interviews in the BHPS. Conceptually, the closest BHPS measure to the HILDA Survey excludes both proxy interviews and short telephone interviews (dark grey solid line) as we do not allow either of these options in HILDA. The two SOEP response rates are for their original AB sample (started in 1984) together with their large general refreshment sample F (started in 2000). The HILDA re-interview rates are reasonably similar to the BHPS and SOEP AB samples in the early waves and have surpassed the other studies in the last few waves. We believe the early HILDA rates compare favourably to the other studies given the comparative waves were conducted 10 to 17 years earlier and it has been generally accepted that response rates to surveys have been falling (eg, De Leeuw and De Heer, 2002).

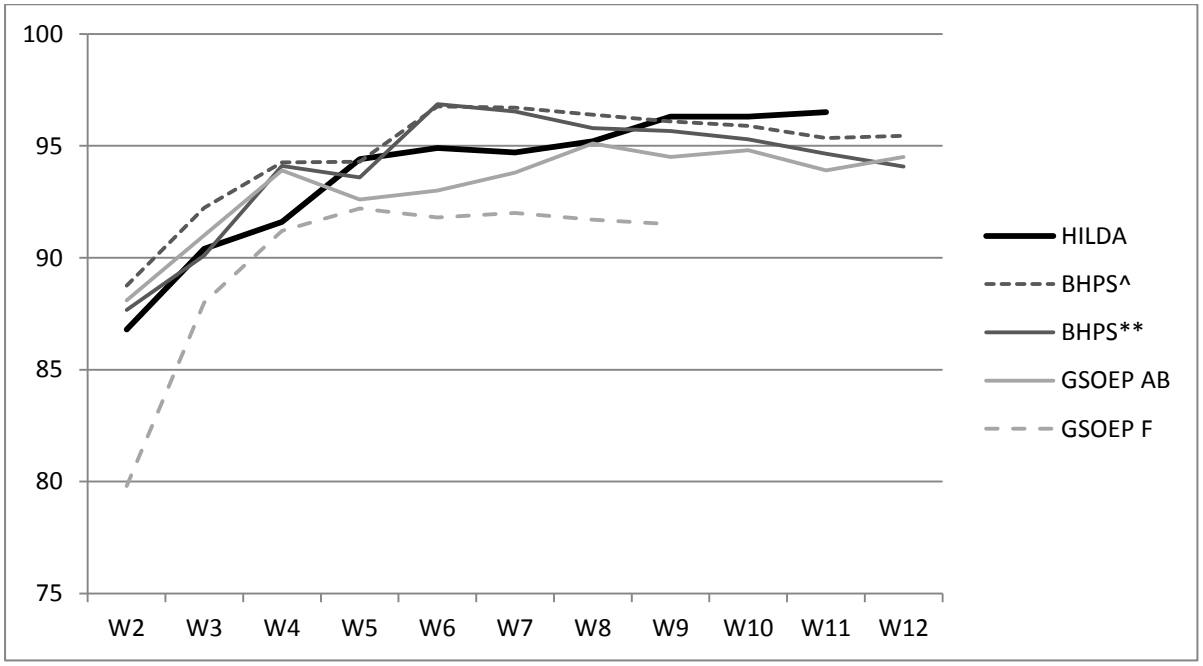
Another measure of the re-interview success is the proportion of wave 1 respondents re-interviewed (excluding those that have died or moved overseas). These rates for HILDA are compared to the BHPS and SOEP experience in Figure 2. By wave 11, we are still interviewing 68 per cent of the in-scope wave 1 respondents. This closely matches the BHPS rate, but is markedly different from the SOEP AB and F samples.

Returning now to the other response rates provided in Table 5, the rates that are most comparable over time are those in the bottom half of the table for people who are attached to a household that responded in the previous wave. Around 15 to 20 per cent of those individuals who did not respond in the previous wave (but belonged to a household where someone else responded) are re-engaged with the study each wave. The response rate for children turning 15 ranges from 80 to 93 per cent and for adults joining the household the response rate is around 75 to 85 per cent.

Table 5: Individual response rates for the HILDA Survey, waves 2 to 11 compared

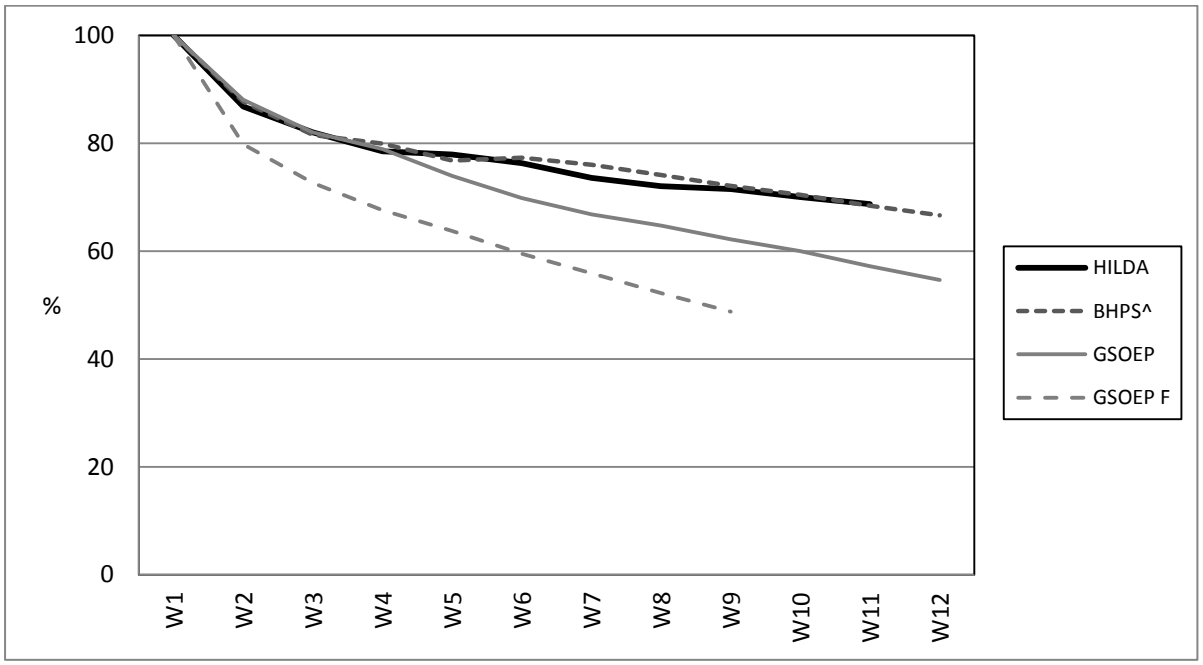
	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11
All people										
Previous wave respondent	86.8	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5
Previous wave non-respondent	19.7	17.6	12.7	14.7	8.4	5.6	5.7	8.5	4.5	3.8
Previous wave child	80.4	71.3	70.7	74.6	75.4	70.8	73.7	73.4	72.0	70.0
New entrant this wave	73.3	76.1	70.4	81.7	81.1	79.7	79.5	81.3	82.9	80.7
People attached to responding household in previous wave										
Previous wave respondent	86.8	90.4	91.6	94.4	94.9	94.7	95.2	96.3	96.3	96.5
Previous wave non-respondent	19.7	19.8	18.1	25.3	18.3	13.2	15.0	25.9	15.9	15.4
Previous wave child	80.4	81.8	81.2	87.3	89.5	90.5	90.9	93.0	92.3	93.0
New entrant this wave	73.3	78.5	71.8	85.4	81.0	80.2	81.2	81.4	83.5	82.0

Figure 1: Wave-on-wave response rates, HILDA, BHPS and SOEP compared



Notes: ^ Includes proxies and short telephone interviews.
** Excludes proxies and short telephone interviews.

Figure 2: Percentage of wave 1 respondents re-interviewed, HILDA, BHPS and GSOEP compared



Note: Deaths and moves out of country are excluded from the denominator.

Attrition is generally only a serious concern when it is non-random (that is, when the persons that attrit from the panel have characteristics that are systematically different from those who remain).

The *HILDA User Manual* regularly provides some information on differential response rates for various respondent characteristics (see Summerfield et al. 2012, Table 8.24). The reinterview rate is typically lowest among people who were:

- relatively young (aged between 15 and 24 years);
- born in a non-English speaking country;
- of Aboriginal or Torres Strait Islander descent;
- single;
- unemployed; or
- working in low-skilled occupations.

More details on factors affecting attrition are provided in Watson and Wooden (2004; 2009; 2011).

As attrition is not random, we need to make adjustments for attrition in the analysis we do, though of course these adjustments are only as good as our ability to measure differential attrition. One such way to make adjustments for attrition is through the use of sample weights.

General steps in weighting

There are four main steps in developing weights:

1. Determine which sample units are in-scope of the population.
2. Calculate the initial weights as the inverse of the probability of selection.
3. Adjust for non-response by developing response homogenous groups or modeling response propensities.
4. Calibrate to known benchmarks to ensure the certain weighted estimates match (typically external) high quality totals.

Each of these steps are undertaken in preparing the various cross-sectional and longitudinal weights for the HILDA datasets.

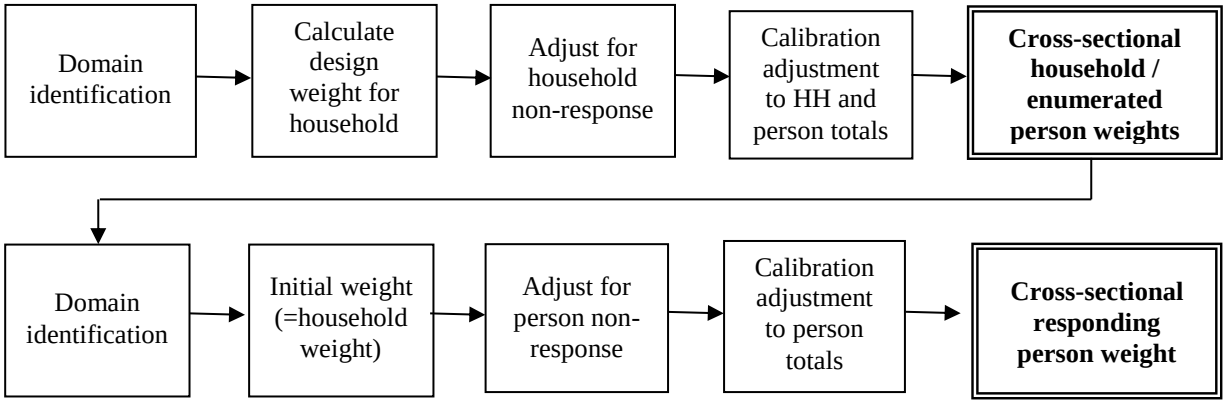
For longitudinal purposes, weights are provided at the individual level for those who provide an interview ('responding persons') and for those who are part of a responding household ('enumerated persons') where at least one individual provides an interview. Due to the changing nature of households and many research-specific definitions that could be used for a longitudinal household, we do not provide longitudinal household weights.

Cross-sectional weights are provided for households, responding persons and enumerated persons. New entrants who join the households (i.e., TSMs) are included in the cross-sectional estimates. It has been shown using Canadian data that including these cohabitants in the cross-sectional estimates helps to make them more representative (LaRoche, 2003).

Cross-sectional weights for wave 1

Figure 3 outlines the process for constructing the wave 1 cross-sectional weights. The household and enumerated person weights are determined together, followed by the responding person weights. Each of these steps are now discussed in detail.

Figure 3: Process for calculating cross-sectional weights for wave 1



Household and enumerated person weights

Domain identification

The first step in the wave 1 weighting process is to determine which households are considered in-scope and which are not. As shown in Table 3, there were 12,252 addresses issued which resulted in 804 addresses being identified as out of scope (as they were vacant, non-residential, or all members of the household were not living in Australia for 6 months or more). In addition, there were 245 households added to the sample due to multiple households living at one address. This resulted in 11,693 in-scope households of which 7682 responded.

Within each of these responding households, we need to determine which people are considered in-scope. To be enumerated, an individual needs to belong to the household. A household is defined as "a group of people who usually reside and eat together".[4] The ABS clarifies how this definition is operationalised. Specifically, a household is either:

- a one-person household, that is, a person who makes provision for his or her own food or other essentials for living without combining with any other person to form part of a multi-person household; or

[4] See *Statistical Concepts Library*, ABS Cat. No. 1361.30.001, ABS, Canberra.

- a multi-person household, that is, a group of two or more persons, living within the same dwelling, who make common provision for food or other essentials for living. The persons in the group may pool their incomes and have a common budget to a greater or lesser extent; they may be related or unrelated persons, or a combination of both.

We differed from the ABS definition in one respect. We include children attending boarding schools or halls of residences while studying as members of the sampled households provided they spent at least part of the year there. Using these definitions, there were 19,914 people enumerated in the responding households in wave 1 (as shown in Table 4).

Calculate initial (design) weights

The initial (or design) weights are inverse of the probability of selecting the households into the sample ($w_{1h,design} = 1/p(select\ HH)$). Given the three stage design, the probability of selecting household *j* is:

$$\begin{aligned} p(select\ HH) &= P(select\ CCD)\ P(select\ block|select\ CCD) \\ &\quad P(select\ dwelling|select\ block)\ P(select\ HH|select\ dwelling) \\ p_h &= \left(488\frac{\widehat{D}_c}{\sum_i \widehat{D}_i}\right)\left(\frac{b_c}{B_c}\right)\left(\frac{d_b}{D_b}\right)\left(\frac{h_d}{H_d}\right) \end{aligned} \quad (1)$$

where $\widehat{D}_c$ is the number of (occupied and unoccupied) dwellings in CCD *c* based on the 1996 Census, $\sum_i \widehat{D}_i$ is the number of dwellings in all CCDs (excluding sparsely populated and remote areas). b_c is the number of blocks selected in CCD *c* which equals the total number of blocks in the CCD (B_c) for all but five CCDs. D_b is the number of dwellings listed in all selected blocks *b* and d_b is the number of dwellings selected from the listed dwellings. H_d is the number of households in dwelling *d* with h_d selected.

Table A1.1 in Appendix 1 compares the distribution for a range of variables for responding and non-responding households. We find that responding households are more likely to be in rural areas, live in separate houses, and do not have a locked gate, security door or no junk mail sign. As a check on the sample selected, the 2001 Census figures for geographical area and dwelling type are also provided and the selected sample distributions align closely with those from the Census.

Adjust for non-response

The design weights are adjusted for differential household non-response using information collected or known about all selected households (both responding and non-responding). A logistic regression model for predicting household response was developed. It includes the following covariates that the interviewers observed for all selected households: dwelling type, external condition of the dwelling, security features of the dwelling (for example, locked gate, security guard, security door, dangerous dog, no junk mail sign, bars on windows, etc.), and the proportion of high-rise buildings in the area. The model also included covariates about the CCD including the geographical location, population density, proportion of people speaking a language other than English, proportion of people not in the labour force, proportion of people unemployed and SEIFA indicators of advantage. Table A2.1 in Appendix 2 provides details of the estimated model. This model is used to predict the probability of response for each household ($\hat{p}_{hr}$). The inverse of the probability of response is multiplied by the household design weight ($w_{1h,design}$) to give the response adjusted household weight:

$$w_{1h,adj} = w_{1h,design}\ \frac{1}{\hat{p}_{hr}}$$

Calibration to known benchmarks

The final step in the weighting process is to fix the response-adjusted weights to several known external population totals. The benchmarks used in the weighting process are listed in Table 6.[5] A SAS macro developed by the Methodology Division at the ABS (GREGWT) is used to calibrate the weights to multiple benchmarks.[6] The household and enumerated person weights are calibrated at the same time resulting in the same weight for the household as for every enumerated person in that household.[7]

The person benchmarks for State, part of State, sex and age are from the Estimated Residential Population figures produced by the ABS based on the 2001 Census and the 2006 Census, updated for births, deaths, immigration, emigration and interstate migration.[8] The household benchmarks are derived from these person benchmarks by the ABS. The person benchmarks for household composition are derived from the household benchmarks. The person benchmarks for labour force status and marital status come from the ABS Labour Force Survey.

These benchmarks have two population exclusions that give rise to zero weights for some cases. First, the very remote parts of New South Wales, Queensland, South Australia, Western Australia and the Northern Territory have been excluded from the benchmarks, which is in line with the practice adopted in similar large-scale surveys run by the ABS. Second, these benchmarks exclude people living in non-private dwellings. For wave 1, only the first exclusion has an impact where some households selected into the sample are given zero weight due to a small change in the definition of areas considered very remote.[9] In subsequent waves, both of these exclusions will cause people living in non-private dwellings and those living in very remote areas to be given zero cross-sectional weights.

The benchmarks may change a little from release to release resulting in changes to the weights. This is because of changes to the methodology used to create the benchmarks or updates to the underlying sources of information that feeds into the estimates. Apart from methodological changes, the benchmarks used for the weights in the first five waves of HILDA have been stable since Release 8 following final revisions given the 2006 Census data.

[5] The Demography Section and the Labour Force Estimates team from the ABS provide the benchmarks used in the weighting process.

[6] The GREGWT macro performs generalized regression weighting as described by Stukel, Hidiroglou and Sarndal (1996).

[7] This is known as integrated weighting and allows for identical estimates where the same concept (such as the number of people living in two person households) can be determined from different level files (household and enumerated files). Due to the demands placed on the weights through the integrated weighting process some of the benchmarks initially specified by Watson and Fry (2002) have been simplified. Further, additional benchmarks on marital status and household composition have been included due to concerns about the representativeness of the sample.

[8] See *Population Estimates: Concepts, Sources and Methods*, ABS Cat.No. 3228.0.55.001, ABS, Canberra.

[9] This stemmed from a change in the benchmarks available from the ABS to align with the 'very remote' category of the Remoteness Area classification (based on the Accessibility / Remoteness Index for Australia) rather than a 'remote and sparsely settled' definition that was originally used.

Table 6: Benchmarks used in weighting

	Household weights	*Enumerated person weights*	*Responding person weights*
Cross-sectional weights	• Number of adults by number of children* • State by part of State* *Determined jointly with enumerated person weights*	• Sex by broad age • State by part of State • Labour force status • Marital status *Determined jointly with household weights*	• Sex by broad age • State by part of State • State by labour force status • Marital status • Household composition (number of adults and children)
Longitudinal weights	Not applicable	• Sex by broad age • State by part of State • Labour force status • Marital status • Household composition (number of adults and children)	• Sex by broad age • State by part of State • State by labour force status • Marital status • Household composition (number of adults and children)

* Due to updates to the household propensities used by the ABS to create the household benchmarks, the total number of households based on the 2006 Census is somewhat different from that based on the 2001 Census. For example, the number of households in Australia in September 2001 based on the 2001 Census was 7.43 million, whereas the corresponding number based on the 2006 Census was 7.32 million. In order to minimise the impact on our estimates caused by changes to the benchmarks, an incremental combination of the two sets of household benchmarks has been taken.

Responding person weights

Domain identification

For the responding person weights, we determine which people are in-scope to be interviewed and which are not. In wave 1 there were 15,127 adults who were eligible to be interviewed (that is, aged 15 and over at the 30th June 2001). Of these 13,969 were interviewed.

Calculate initial weights

The initial weight for the responding person weights is the final household weight determined above.

Adjust for non-response

Individual level characteristics for enumerated and responding persons are compared in Table A1.2 in Appendix 1. We find differences in response with respondents more likely to be living in rural areas, male, older, married, and in households where children are present. The ABS Labour Force estimates are also provided for comparison purposes and we find that respondents are less likely to be born in countries where the main language is not English, employed full time, or own account workers.[10]

As a result, the initial responding person weights require a response adjustment for person-level non-response in responding households. This is only undertaken in households with two or more adults (as adults in one adult households by definition respond). The covariates used in this logistic regression model are derived primarily from the Household Form and include geographical location, labour force status, sex, age, number of adults, number of children, marital status, English language ability, and dwelling type. Table A2.2 in Appendix 2 provides details of this estimated model. To get the response

[10] Some of these differences may be explained, in part, by differences in the scopes of the two surveys, with the Labour Force Survey including people in institutions and very remote Australia. Note that there are three Labour Force Survey estimates – relationship in household, country of birth and indigenous status – that exclude the institutionalised population, so are closer in scope to the HILDA Survey. Further, the definition of part-time employment versus full-time employment is based on the number of actual and usual hours, whereas in the in HILDA Survey it is based just on usual hours.

adjusted responding person weight, the final household weight ($w_{1h,bm}$) is adjusted by the probability of the person providing a response ($\hat{p}_{pr}$) in the following way:

$$w_{1r,adj} = \begin{cases} w_{1h,bm}\dfrac{1}{\hat{p}_{pr}} & in\ 2+\ adult\ households \\ w_{1h,bm} & in\ 1\ adult\ households \end{cases}$$

Calibration to known benchmarks

The response-adjusted responding person weights are calibrated to the population benchmarks indicated in the third column in Table 6. As noted in the calibration section for the household and enumerated person weights, some weights can be zero and the weights may vary from release to release.

Longitudinal weights

Longitudinal responding person weights

Domain identification

A number of longitudinal weights are constructed and these are defined by this domain identification stage where we decide who is considered an acceptable 'response' and who is not.

Most users will be interested in the longitudinal responding person weight for the continuous panel from wave 1 to wave *t*. The continuous panel from wave 1 is defined as the group of people who were interviewed in wave 1 and then were interviewed, overseas or dead at every wave to wave *t*. These people are counted as 'responses' and every other person interviewed in wave 1 are 'non-responses'.

The longitudinal responding person panels for which weights are provided (including the one just mentioned) are presented in the top half of Table 7 along with the definition of 'responses' and 'non-responses' for each panel. The panels include:

- Continuous balanced panel of respondents from wave 1 to *t*;
- Continuous balanced panel of respondents from wave t_1 to t_n;
- Paired balanced panel of respondents for wave t_1 and t_n;
- Balanced panel of respondents for the retirement module (waves 3, 7 and 11); and
- Balanced panel of respondents for the fertility module (waves 5, 8 and 11).

Calculate initial weights

The initial weights for the longitudinal responding person weights are the final cross-sectional responding person weights for the starting wave of the panel (for example, for the balanced panel from wave 1 to 5, the starting wave is wave 1). The calculation of the final cross-sectional weight has been described elsewhere but it is essentially the design weight adjusted for non-response and benchmarked to known population totals.[11]

Adjust for non-response

The longitudinal responding person weights are adjusted for attrition from the initial wave. This is done by constructing a logistic model to predict the probability each individual had of responding. The model includes covariates from the initial wave of the panel and some information about changes after the initial wave. These covariates about the individual include: age, sex, marital status, ability of speak English, employment status, hours worked, number of children, country of birth, highest level of education, relationship in household, health status, likelihood of moving, number of times moved

[11] If the panel starts in wave 2 or later, then the weight has also been adjusted for TSMs as described in the section on cross-sectional weights for wave 2 and later. If the TSM subsequently leaves (that is, becomes inactive) during the particular balanced panel, this adjustment is reversed.

Table 7: Domain identification for longitudinal panels

	Responses	Non-responses
Responding persons		
Continuous balanced panel from wave 1 to wave *t*	Interviewed in wave 1 Interviewed, overseas or dead in waves 2 to wave *t*	All other individuals interviewed in wave 1
Continuous balanced panel from wave t_1 to wave t_n	Interviewed in wave t_1 Interviewed, overseas or dead in waves t_{1+1} to t_n	All other individuals interviewed in wave t_1 excluding TSMs who become inactive between t_{1+1} and t_n
Paired balanced panel for wave t_1 and t_n	Interviewed in wave t_1 Interviewed, overseas or dead in wave t_n	All other individuals interviewed in wave t_1 excluding TSMs who become inactive by t_n
Balanced panel for retirement module waves (3, 7, 11)	Interviewed in wave 3 Interviewed, overseas or dead in wave 7 and 11	All other individuals interviewed in wave 3 excluding TSMs who become inactive by wave 7 or 11
Balanced panel for fertility module waves (waves 5, 8, 11)	Interviewed in wave 5 Interviewed, overseas or dead in wave 8 and 11	All other individuals interviewed in wave 5 excluding TSMs who become inactive by wave 8 or 11
Enumerated persons		
Continuous balanced panel from wave 1 to wave *t*	Enumerated (i.e., part of responding household) in wave 1 Enumerated, overseas or dead in waves 2 to wave *t*	All other individuals enumerated in wave 1
Continuous balanced panel from wave t_1 to wave t_n	Enumerated in wave t_1 Enumerated, overseas or dead in waves t_{1+1} to t_n	All other individuals enumerated in wave t_1 excluding TSMs who become inactive between t_{1+1} and t_n
Paired balanced panel for wave t_1 and t_n	Enumerated in wave t_1 Enumerated, overseas or dead in wave t_n	All other individuals enumerated in wave t_1 excluding TSMs who become inactive by t_n
Balanced panel for wealth module waves (2, 6, 10)	Enumerated in wave 2 Enumerated, overseas or dead in wave 6 and 10	All other individuals enumerated in wave 2 excluding TSMs who become inactive by wave 6 or 10

in last 10 years, whether flagged as reference person for household. Details of the interview situation, as recorded by the interviewer, are also included, these being: the level of cooperation, whether the interview was assisted, whether there were difficulties during the interview (e.g., with eyesight, hearing, reading), whether the respondent was suspicious of the study, how well they understood the questions, whether their answers were influenced by others, the length of the interview, and whether the Self-Completion Questionnaire was returned. Household characteristics are also included in the model, such as geographical location, remoteness area, SEIFA index of disadvantage, dwelling type, dwelling condition, number of bedrooms, number of calls made, whether the household was partially responding, number of adults, number of children, household type, housing tenure, whether any household members are benefit recipients, household income, household splits and household moves. Details of the models between wave 1 and 2 are provided in Table A2.3 in Appendix 2. The response-adjusted weight for, say, the longitudinal panel of respondents from wave t_1 to t_n is given by:

$$w_{r,t_1 t_n} = w_{r,t_1} \frac{1}{p_{r,t_1+1 t_n}}$$

where w_{r,t_1} is the cross-sectional responding person weight for wave t_1 and p_{r,t_1+1t_n} is the probability of observing the person in waves t_1+1 to t_n.

Calibration to known benchmarks

These response-adjusted responding person weights are then benchmarked back to the key characteristics of the initial wave according to the benchmarks set out in Table 6.

Longitudinal enumerated person weights

Domain identification

The longitudinal enumerated person panels for which weights are provided are presented in the bottom half of Table 7 along with the definition of what is considered 'responses' and 'non-responses' for each panel. The panels include:

- Continuous balanced panel of enumerated persons from wave 1 to *t*;
- Continuous balanced panel of enumerated persons from wave t_1 to t_n;
- Paired balanced panel of enumerated persons for wave t_1 and t_n; and
- Balanced panel of enumerated persons for the wealth module (waves 2, 6, and 10).

Calculate initial weights

The initial weights for the longitudinal enumerated person weights are the final cross-sectional enumerated person weights for the starting wave of the panel as described elsewhere.[12]

Adjust for non-response

The longitudinal enumerated person weights are adjusted for attrition from the initial wave. Models predicting response are constructed using covariates from the initial wave and mobility indicators from subsequent waves. The models are split by whether the person was a respondent in wave t_1 or not to allow for much greater use of respondent covariates where they are available. Details of the models between wave 1 and 2 are provided in Table A2.3.

Calibration to known benchmarks

These response-adjusted weights are then benchmarked back to the key characteristics of the initial wave according to the benchmarks set out in Table 6.

Cross-sectional weights for waves 2 to 10

While we provide cross-sectional weights on the data files, using a longitudinal survey for cross-sectional purposes is not ideal. Over time, there are issues of increasing magnitude with the coverage of the population unless top-up samples which include recent immigrants are added. As it is quite costly to add a top-up sample, the first one for the HILDA Survey was only added in 2011.[13] Since the study began in 2001, the cross-sectional HILDA estimate for the proportion of people aged 15 and older that are born overseas and arrive in Australia in 2001 or later is markedly different from the ABS Labour Force Estimate (see Figure 4). By 2010, there is a 7 percentage point gap between these two estimates. There is only a small increase in the HILDA estimate over time as some recent arrivals join the households we have sampled. This helps reduce the size of the gap, but only by about 1.5 percentage points.

The amount of bias that this lack of coverage can inject into the cross-sectional estimates will vary across different variables depending on how strongly associated they are to immigration. An example

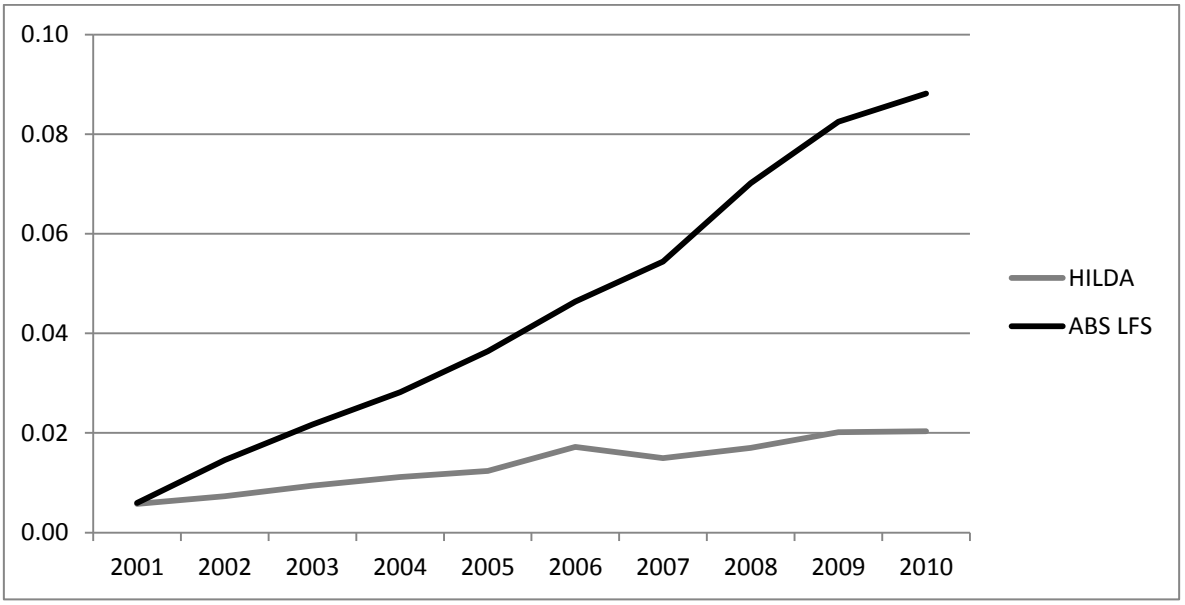
[12] Essentially these weights are the design weights adjusted for non-response and benchmarked to known population totals. If the panel starts in wave 2 or later, any adjustment for TSMs who later become inactive is reversed.

[13] A range of options were canvased (Watson, 2006) and we ultimately chose a general top-up rather than focusing explicitly on immigrants. The most obvious source of immigrants (the Department of Immigration and Multicultural Affairs Settlement Database) excludes New Zealanders. As New Zealanders make up about a quarter of all immigrants, this exclusion was significant so we chose not to pursue this option.

of a variable that is highly affected is country of birth. Figure 5 shows the proportion of the Australian population aged 15 and over who were born in Australia. Over the 9 year period between 2001 and 2010, the HILDA estimates would suggest that the proportion of the adult population born in Australia is increasing over time, whereas for the same period, the ABS Labour Force Estimate is declining. By 2010, these two estimates diverge by 6 percentage points.

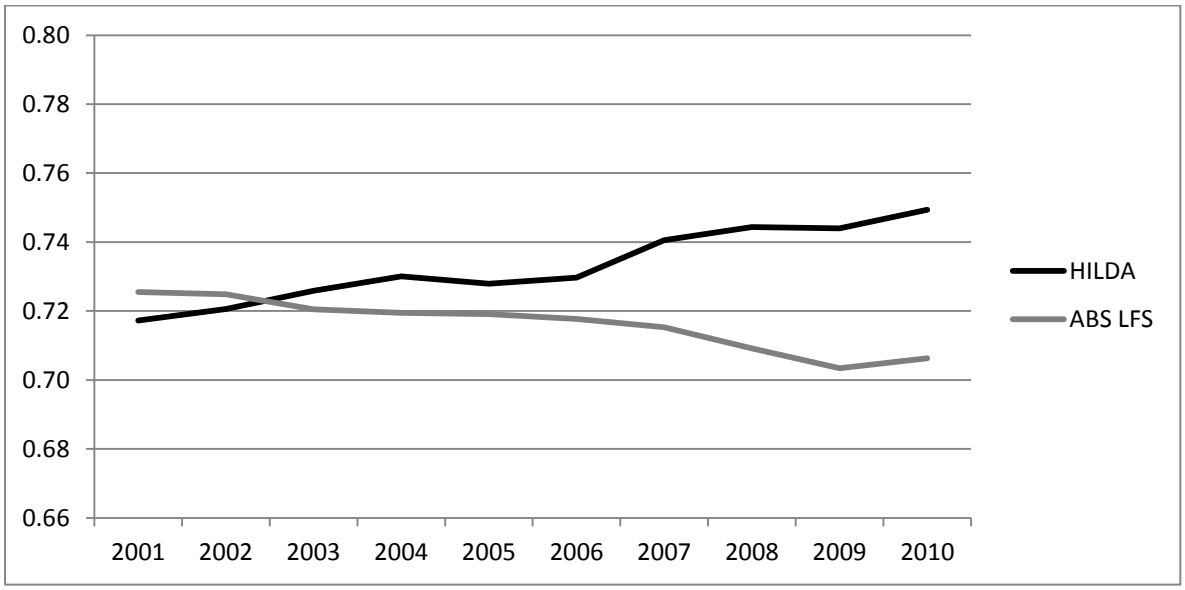
These two variables are most likely the worst affected by these coverage issues. There will be some variables that are only slightly affected and many that are not affected at all.

Figure 4: Proportion born overseas and arrived in 2001 or later (aged 15+), years 2001 to 2010



Notes: 1. ABS Labour Force estimates exclude institutionalised but does include very remote parts of Australia. (ABS Cat.No. 6291.0.55.001, Data cube LM4, September.)
2. HILDA estimates exclude both the institutionalised and very remote parts of Australia.

Figure 5: Proportion born in Australia (aged 15+), years 2001 to 2010



Notes: 1. ABS Labour Force estimates exclude institutionalised but does include very remote parts of Australia. (ABS Cat.No. 6291.0.55.001, Data cube LM7, September.)
2. HILDA estimates exclude both the institutionalised and very remote parts of Australia.

Household and enumerated person weights

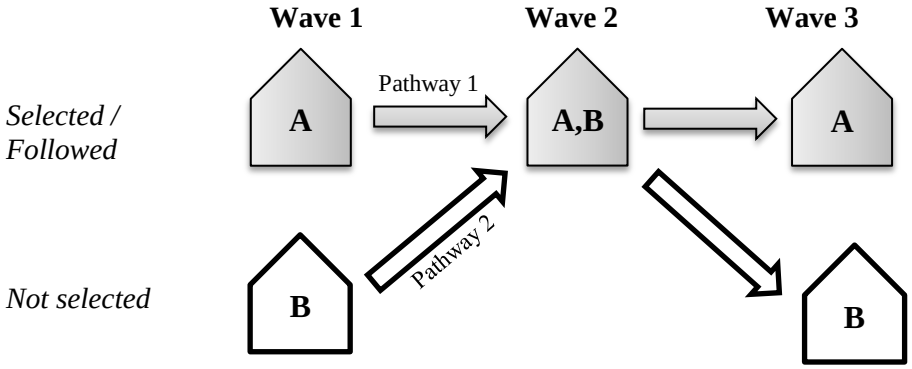
Domain identification

Keeping in mind the potential for biased cross-sectional estimates derived from longitudinal surveys, users may still find these estimates useful and we provide cross-sectional weights for each wave on the datasets. These weights opportunistically include temporary members into the sample (i.e., those people who are part of the sample only because they currently live with a PSM). Responding households are counted as those where at least one person in the household provided an individual interview in wave *t*. Enumerated person weights are assigned to those individuals who belong to these responding households. All other people are treated as non-respondents with the exception of who are dead, overseas or are TSM leavers from the household (who are treated as out of scope).

Calculate initial weights

The underlying probability of selection for households in wave 2 onwards needs to be corrected to account for the various pathways from wave 1 into the specific wave household. Consider the following situation (displayed in Figure 6): we select a household that contains person *A* in wave 1; in wave 2 person *B* moves in; and in wave 3 person *B* moves out. In wave 2, the household weight needs to be adjusted downwards to allow for the probability of observing the household via person *A* (which we did observe) or via person *B* (which we did not observe). That is, we need to estimate the probability of selection that person *B* had in wave 1. In wave 3, this adjustment is not needed as person *B* has left the sample. If, however, person *B* who is a TSM is converted to a PSM (i.e. by having child *C* with person *A*), then when person *B* leaves the household, they will be followed and their adjusted weight is retained.

Figure 6: Examples of pathways into and out of households over time



The correction to the initial household weight involves the following steps:

1. Step 1: Identify family groups within the new entrants joining the household. Related people are assumed to join the wave *t* household together. Unrelated people are assumed to join the household separately. Newly born babies, adoptions and recent immigrants (since 2001) are considered part of the 'intact' household group (they are organic additions to the sample).
2. Step 2: Identify a reference person within each of these new entrant family groups. The reference person is the first within the family group to satisfy the following ordered requirements: couple, lone parent, non-dependent child, dependent child, other related, not related. A preference for a respondent as the household reference person was taken over a non-respondent (so that as much personal information could be used as possible).
3. Step 3: Construct a regression model to predict a 'quasi-selection' probability for the new entrant family groups. This consists of the following steps:

- Step 3a: Identify a reference person within the intact group from the selected wave 1 household, using similar criteria as above.
- Step 3b: Convert the final wave 1 household weight to a 'quasi-selection' probability by taking the inverse of the weight (that is, $p_{1h} = \frac{1}{w_{1h,bm}}$).[14] As the 'quasi-selection' probability is bounded by 0 and 1, transform it into a new variable *y* which has a continuous scale, via the following:

$$y = ln\left[\frac{p_{1h}}{(1-p_{1h})}\right]$$

- Step 3c: Construct a regression model of the transformed variable *y* using the wave *t* person information for the reference person of the intact group and the wave *t* household information.
- Step 3d: Use this model to predict a wave 1 'quasi-selection' probability ($\hat{p}_{1fi}$) for the new entrant family groups (i.e., for cases like *B* in the above illustration). From the model of *y*, obtain an estimate $\hat{y}$ given the characteristics of the household and the reference person of the new entrant family group. Transform this into the probability for the new entrant family group using:

$$p_{1fi} = \frac{e^{\hat{y}}}{(1+e^{\hat{y}})}$$

4. Step 4: Construct the revised wave *t* household weight which adjusts for the multiple pathways into the wave *t* household. This adjustment is done via the following formula which accounts for the joint selection probabilities of these family groups:

$$w_{th,init} = \frac{1}{\left[1-(1-p_{1h})\left(1-\hat{p}_{1f1}\right)\ldots\left(1-\hat{p}_{1fn}\right)\right]}$$

where p_{1h} is the 'quasi-selection' probability for the intact family group, and $\hat{p}_{1fi}$ is the estimated 'quasi-selection' probability for the new entrant family *i*. For new entrant family groups where nobody responded in wave t, the wave 1 'quasi-selection' probability is taken to be zero as it is likely they would not have responded in wave 1 (so would not have been followed along that pathway into wave *t*).

For wave 1 households that have merged with other wave 1 households by wave *t*, we make similar adjustments to the wave *t* household weight as described above. In this instance, we do not need to model the wave 1 'quasi-selection' probability as they are known.

Adjust for non-response

Following this correction to the initial household weights due to the effect of new entrants, the weights are adjusted for the probability that the household response in wave *t*. This is done via a logistic regression model of household reference persons using the same wave 1 household characteristics, wave 1 reference person characteristics and household splits and moves since wave 1 as used in the longitudinal response probability models described earlier. This household staying probability is used to construct an interim weight:

$$w_{th,interim} = \frac{w_{th,init}}{\hat{p}_{th,stay}}$$

[14] As we have incorporated both selection and response probabilities into this wave 1 weight, we refer to the inverse as a 'quasi-selection' probability.

Calibration to known benchmarks

The interim household weight is then calibrated to household and enumerated person benchmarks as set out in Table 6.

As the HILDA sample is not representative of the population in institutions or very remote areas and these parts of the population can be excluded from the benchmarks, any sample members who move into these places receive a zero cross-sectional weight.

Responding person weights

Domain identification

Cross-sectional responding person weights for wave *t* are assigned to people who provided an interview in wave *t*. All other people in wave *t* are treated as non-respondents with the exception of who are dead, overseas or are TSM leavers from the household (who are treated as out of scope).

Calculate initial weights

As in wave 1, the initial cross-sectional responding person weight is taken as the final cross-sectional household weight for wave t.

Adjust for non-response

The initial responding person weights are adjusted in responding households with two or more adults to allow for differential response propensities using the same variables as described for this step in the wave 1 cross-sectional weights, but this time for wave *t*.

Calibration to known benchmarks

The response adjusted responding person weights are calibrated to the responding person benchmarks listed in the last column of Table 6.

As the HILDA sample is not representative of the population in institutions or very remote areas and these parts of the population can be excluded from the benchmarks, any sample members who move into these places receive a zero cross-sectional weight.

Cross-sectional weights for wave 11 (integration of original and top-up samples)

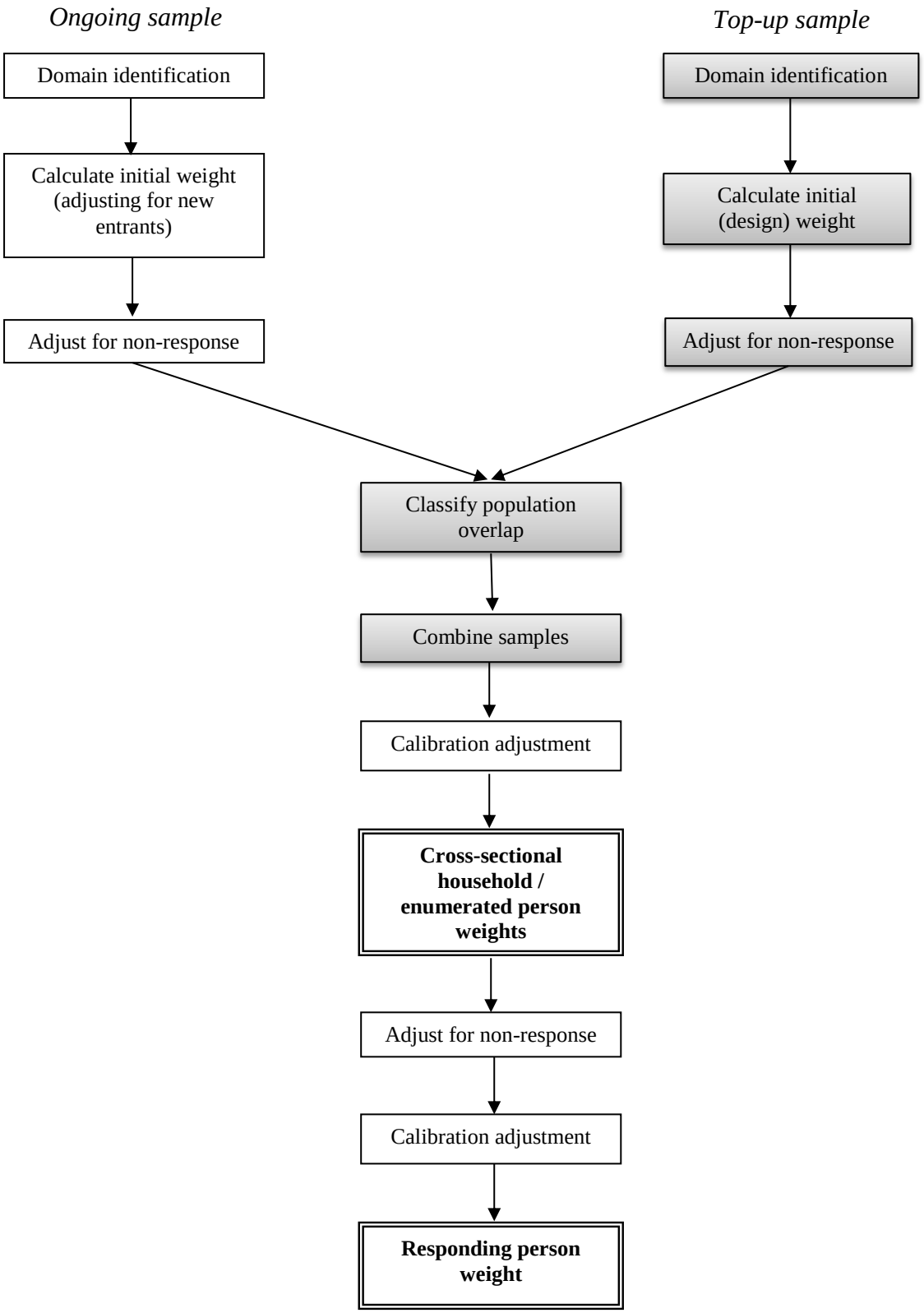
With the introduction of the top-up sample, the process for calculating the cross-sectional weights for wave 11 from the integrated sample becomes more complicated. As shown in Figure 7, the two samples have separate treatment through to the non-response adjustment stage. The samples are then combined together prior to the calibration stage which produces the integrated household and enumerated person weights. The new steps in this process compared to those just described for the wave 2 to 10 cross-sectional weights are shown by the grey shaded boxes. The first three grey boxes on the top right hand side of Figure 7 correspond to the weighting steps undertaken for the original sample in wave 1 but this time for the wave 11 top-up sample. The next two shaded boxes are particular to wave 11.

Common steps in weighting process compared to earlier waves

There are a number of steps in the weighting process for the wave 11 cross-sectional weights that are the same as (or very similar to) those undertaken for waves 1 to 10. These are listed below.

1. The domain identification, initial household weights, and adjustment for household non-response in the main sample in wave 11 is the same as described for waves 2 to 10.
2. The domain identification, initial household weights, and adjustment for household non-response in the wave 11 top-up sample is very similar to that described for wave 1. There are a few differences, which are as follows:

Figure 7: Process for calculating cross-sectional weights for wave 11



a) The probability of selecting the CCD for the wave 11 top-up sample was proportional to the number of occupied private dwellings, resulting in a small change to the first term in equation (1) to refer to occupied private dwellings only.

b) The interviewer observations in wave 11 are a little different from wave 1 and there have been some changes in which variables are associated with response. For the additional interviewer observations recorded in the wave 11 top-up sample, we find that responding households are less likely to have a garden, be on a main road and more likely to contain children. We also find that responding and non-responding households in the wave 11 top-up sample are not significantly different in terms of dwelling type, condition of dwelling, and having a no junk mail sign (as shown in Table A1.1 in Appendix 1) whereas they were in the wave 1 sample.

c) The adjustment for household non-response for the wave 11 top-up sample can include some additional household observations that were collected by the interviewer. The model results are shown under "Model A" in Table A2.1 in Appendix 2. These additional observations helped to improve the pseudo-R^2 from 0.036 to 0.046.

d) Person level differences in response in the wave 11 top-up are similar to the wave 1 experience with one exception. The exception is that people born in Australia were less likely to participate in the survey and those born in countries where English was not the main language were also somewhat less likely to participate (see Table A1.2 in Appendix 1).[15]

3. The calibration of household and enumerated person weights to known population benchmarks is the same as described for waves 1 to 10.

4. The adjustment for individual-level non-response in households with two or more adults and the calibration of the responding person weights to known population benchmarks is the same as described for waves 1 to 10. The model used to adjust the weights for the probability of an individual interview is provided in Table A2.2 in Appendix 2. Fewer significant differences for the wave 11 top-up sample are likely due to the smaller sample size involved.

As a result, we will not discuss these issues in further detail but will focus our attention on the areas that are different, namely in identifying the population overlap and combining the samples.

Classify population overlap

Let us begin with a catalogue of the differences between the underlying survey population in 2001 and 2011 and then consider how the main sample and the top-up sample differ in 2011.

Figure 8 shows the differences in the survey population in 2001 and 2011. The population in common is all persons living in private dwellings in both 2001 and 2011, excluding the very remote parts of Australia. That is, in 2011, this population is aged 10 or older. The parts of the 2001 population that will not be present in the 2011 population are: i) deaths that have occurred between 2001 and 2011; ii) individuals that have moved overseas after 2001 and have not returned; and iii) individuals that have moved into non-private dwellings or very remote parts of Australia since 2001.

Equivalently, the parts of the 2011 population that were not present in the 2001 population are: i) births that have occurred after 2001; ii) individuals that have moved to Australia from overseas after 2001; and iii) individuals that have moved out of non-private dwellings or very remote parts of Australia since 2001.

[15] Most of this difference appears to be from households containing recent arrivals. It seems that people who arrived recently were more motivated to respond to the top-up interview perhaps because they saw the study as more relevant to them. An adjustment factor of 0.8 was applied to the design weights of households that contain a majority of recent arrivals.

Given the following rules that have been applied to the original sample, the population represented by the main sample in 2011 overlaps with the survey population for the top-up sample, as shown by the dashed line in Figure 8, and includes:

- individuals living in private dwellings in both 2001 and 2011, excluding very remote parts of Australia;
- a portion of the 'recent arrivals' (people born overseas and arrived in Australia after 2001);[16] and
- births since 2001, excluding a portion to 'recent arrivals' not living with someone who was in Australia in 2001.

The part of the survey population represented by the 2011 main sample that is not included in the top-up sample are individuals who move into non-private dwellings or very remote parts of Australia since 2001. Conversely, the parts of the survey population represented by the 2011 top-up sample that are not included in the main sample are:

- the recent arrivals that were born overseas and arrived in Australia after 2001 who did not form a household with a person living in Australia since 2001;
- births to recent arrivals who did not form a household with a person living in Australia in 2001; and
- individuals that have moved out of non-private dwellings or very remote parts of Australia since 2001.

In integrating these samples, we determine if and how each of these groups will be treated.

We begin by identifying which parts of the population the various sample members are from in the two samples. In the main sample, we can identify those who have moved into very remote parts of Australia, moved into non-private dwellings, died or moved overseas (though these two groups will not be part of responding households), been born since 2001, arrived from overseas since 2001, and been born to recent arrivals. The group that we cannot identify in the main sample which will overlap with the top-up sample is Australians who were overseas in 2001 and have since returned and started living with a PSM. It is estimated that this group is relatively small and is of negligible consequence.[17]

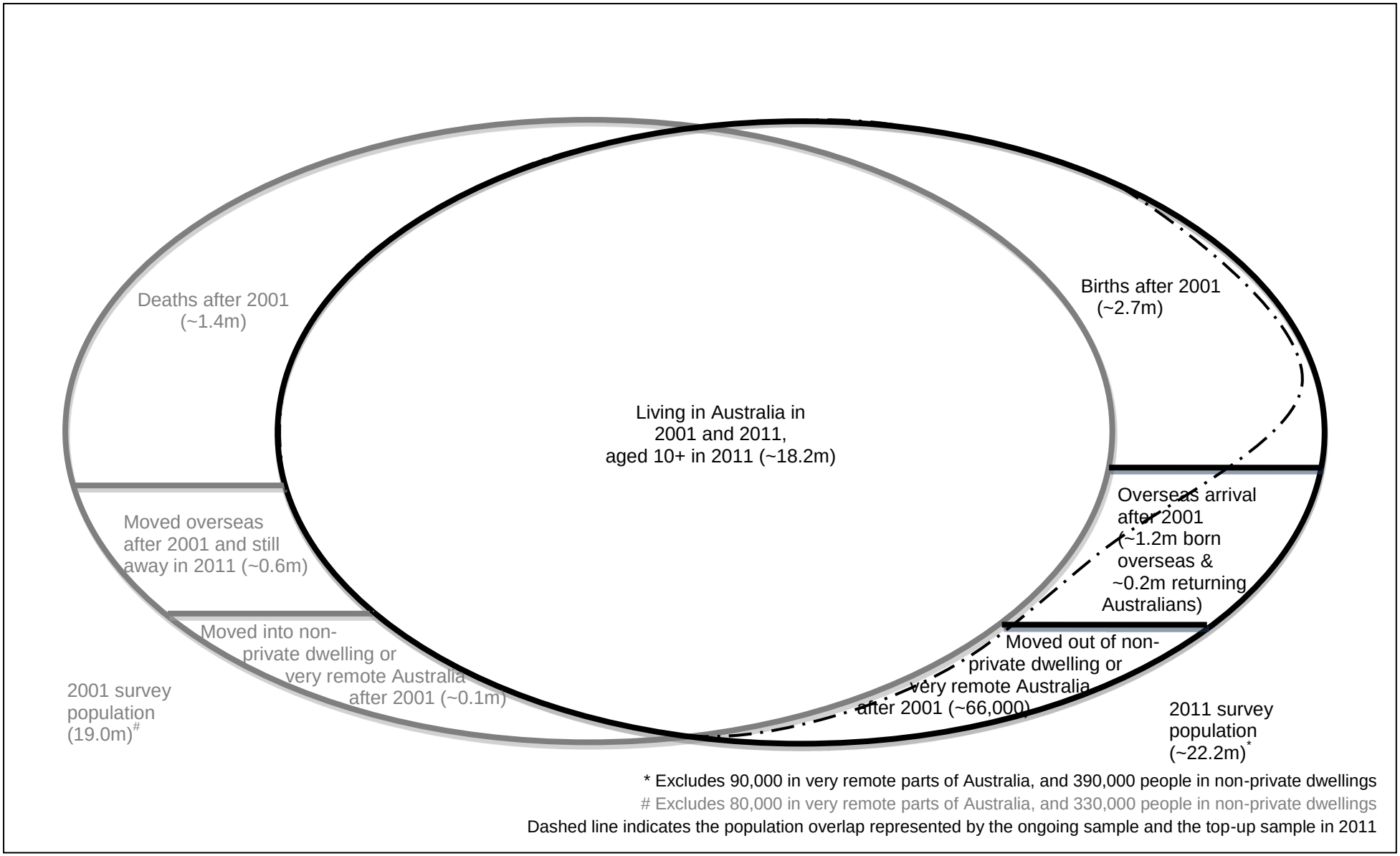
In the top-up sample, we can identify individuals who have been born since 2001, and those who arrived from overseas since 2001. We cannot identify the following groups: i) Australians who were overseas in 2001 and have since returned; ii) individuals who were living in very remote parts of Australia in 2001 and have moved to other parts of Australia by 2011; and iii) individuals who were living in non-private dwellings in 2001 and have moved to private dwellings by 2011. It is difficult to determine how many of these people there might be in the population, but based on the experience of the first 11 waves of the main sample, we expect to have 28 individuals in the top-up sample in group 1, and 12 individuals in group 2 and 3 combined. Again, these numbers are fairly inconsequential so we do not consider them further.

As all individuals in the household are selected, the household rather than the individual needs to be classified into the various population overlap categories. This problem reduces to one of identifying households that contain some, all or no recent arrivals in the top-up sample and in the main sample. Households in the main sample that are now in very remote parts of Australia or those in non-private dwellings are excluded from the cross-sectional weights in the benchmarking step so these can also be classified at this stage. People who have died or moved overseas in the main sample have already

[16] This group of recent arrivals is restricted to those individuals who form a household with a person who lived in Australia in 2001.

[17] There have been 0.4 per cent of our wave 1 sample who we have observed to have moved overseas between 2001 and 2006 and back again by 2011. Applying this rate to the active temporary sample members in the ongoing sample in wave 11 would suggest there may be around 10 individuals who may have been temporarily living overseas in 2001 when we selected the original sample who may have returned and have joined our sampled households as temporary sample members.

Figure 8: Changes in the population from 2001 to 2011



* Excludes 90,000 in very remote parts of Australia, and 390,000 people in non-private dwellings

\# Excludes 80,000 in very remote parts of Australia, and 330,000 people in non-private dwellings

Dashed line indicates the population overlap represented by the ongoing sample and the top-up sample in 2011

been excluded from respondents and non-respondents in the domain identification step. In households with non-responding adults, it has been assumed that they would be classified in the same way as the respondents in that household.

Table 8 provides a breakdown of these different household types in the wave 11 responding sample. There are 85 households in the main sample that have moved into institutions and 28 that have moved into very remote parts of Australia. These will be excluded from the cross-sectional weights. In the remainder of the main sample, we have 97.6 per cent of households that do not contain any recent arrivals, 2.1 per cent that contain some recent arrivals, and 0.3 that only contain recent arrivals.[18] In comparison, in the top-up sample, we have 87.5 per cent of households without any recent arrivals, 4.2 per cent that contain some, and 8.3 per cent that contain all recent arrivals.

Table 8: Classification of the wave 11 responding households representing population overlap groups

	Main sample	*Top-up sample*	*Action in combining samples*
Moved into institution (non-private dwellings)	85	-	Excluded from cross-sectional weights
Moved into very remote parts of Australia	28	-	Excluded from cross-sectional weights
Contains no recent arrivals	7104	1884	Integrate
Contains some recent arrivals	149	90	Integrate
Contains all recent arrivals	24	179	Weight only top-up sample
Total households	7390	2153	

Combine samples

The main and top-up samples can be integrated by either combining estimates or pooling samples. Essentially we are deciding on a factor θ which specifies how much emphasis to give cases from each sample that represents the overlapping portion of the population. Four options were evaluated together with two options that keep the samples separated (see Appendix 3 for details of the comparison). The methods were assessed in terms of minimizing the bias in the range of estimates considered and reducing the variance of these estimates. The method that provides the best improvement to the estimates is the one that pools the samples and estimates the probability of selection and response in each sample.

For the portion of households that contain no or some recent arrivals, the pooled weight for household *i* is given by:

$$w_{11h,pool} = \frac{1}{p_{iA} + p_{iB}}$$

where p_{iA} is the probability of selection and response in the wave 11 main sample, and p_{iB} is the probability of selection and response in the wave 11 top-up sample. For the main sample, the probability of selection and response in the main sample is the inverse of the interim weight after the adjustment for new entrants and non-response ($p_{iA} = 1/w_{11h,interim}$). For the top-up sample, the probability of selection and response in the top-up sample is the inverse of the design weight adjusted for non-response ($p_{iB} = 1/w_{11h,adj}$). As we do not know the probability of selection and response for a household in the sample that we do not observe them in, it is estimated via a regression model (in the same way that is used when adjusting the household weights for the new entrants that have joined the sample). That is, for households in the main sample, the probability of response and selection in

[18] Due to the changes to the following rules to convert recent arrivals into PSMs, we have followed a small number of recent arrivals when they moved out from living with other PSMs and into their own household.

the top-up sample for a household with the same household characteristics and individual characteristics of the household representative person is estimated based on a model of the sample and response probabilities in the top-up sample, giving $\hat{p}_{iB}$. The same process is done for the top-up sample to predict the sampling and response probability for a household with the same characteristics would have had in the main sample. The adjusted-R^2 for the model of probabilities in the main sample is 0.210 and for the top-up sample it is 0.176.

For the households that only contain recent arrivals in the top-up sample, their weight following the response adjustment to the design weight ($\boldsymbol{w_{11h,adj}}$) remains unchanged by this integration step, whereas the households in the main sample that contain only recent arrivals are given zero weight. This latter decision is because the households with only recent arrivals in the main sample are quite a unique subset of all households in the population that only contain recent arrivals. They only became part of the main sample because they lived with someone who was selected in original sample. While the inclusion of these people into the main sample helps reduce bias until a top-up sample is added, we do not have a way to identify those households in the top-up sample to which they would be most similar. It is therefore cleaner to leave them aside in the wave 11 cross-sectional weights.

Caution in using cross-section weights to produce a series of estimates

As mentioned earlier, users of the HILDA data should be mindful of the potential for bias in estimates that are associated with country of birth and year of arrival to Australia. The inclusion of the top-up sample in the wave 11 cross-sectional weights reduces this bias, but in a potentially dramatic fashion. For example, the proportion of people aged 15 and over that are born overseas and arrived in 2001 or later is shown in Figure 9 (this is the same as figure 4 but now includes wave 11). The HILDA estimate for 2010 is 0.02 and in 2011 it jumps to 0.10 using the combined sample which is much closer to the Labour Force Survey estimate. The proportion of people born in Australia also makes a corresponding marked drop between the 2010 estimate and the 2011 estimate (see Figure 10).

Figure 9: Proportion born overseas and arrived in 2001 or later (aged 15+), years 2001 to 2011

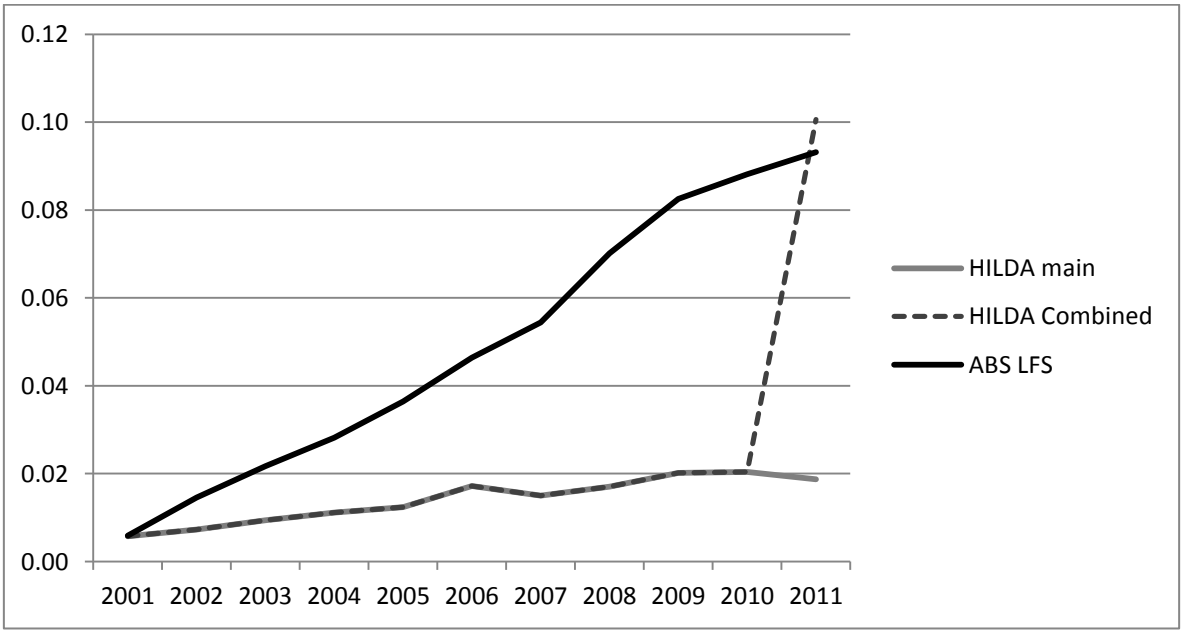
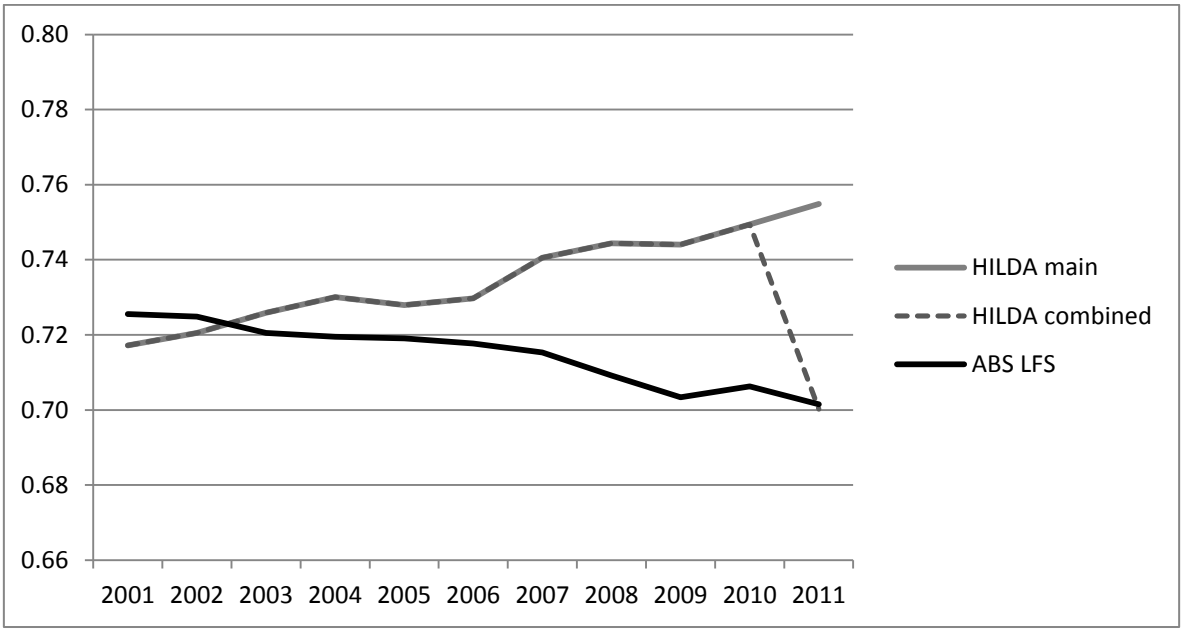


Figure 10: Proportion born in Australia (aged 15+), years 2001 to 2011



For Release 11, three additional weighting variables are provided on the datasets to allow users to see what impact the top-up sample has on the cross-sectional estimates. These weights are:

- *khhwthm* – cross-sectional wave 11 household weight that includes only the main sample
- *khhwtem* – cross-sectional wave 11 enumerated person weight that includes only the main sample
- *khhwtrpm* – cross-sectional wave 11 responding person weight that includes only the main sample

We have also provided a top-up indicator in the household file (*khhtuh*) and person files (*khhtup*).

Comparison of HILDA and ABS cross-sectional estimates for 2011

Table 9 provides a comparison of a range of cross-sectional estimates from the main HILDA sample, the combined sample and several ABS surveys for 2011. The ABS estimates come from the monthly Labour Force Survey and one of it's supplementary surveys (Survey of Education and Work). In addition to the estimate and standard error, the root mean square error is also provided. The RMSE gives a measure of the quality of an estimate ($\hat{Y}$) that considers both the bias in the estimate and the variability in the estimate. It is calculated as:

$$RMSE(\hat{Y}) = \sqrt{bias(\hat{Y})^2 + var(\hat{Y})} = \sqrt{bias(\hat{Y})^2 + SE(\hat{Y})^2}$$

The bias is taken as the difference between the relevant HILDA estimate ($\hat{Y}$) and the ABS estimate ($\hat{Y}_{ABS}$):

$$bias(\hat{Y}) = \hat{Y} - \hat{Y}_{ABS}$$

A lower RMSE is better than a higher one and the lowest RMSE of the estimates from the main sample or the combined sample is in bold in Table 9. The estimates include family type, relationship in household (for both enumerated and responding persons), highest level of education, country of birth, year of arrival, indigenous status, and for those employed we consider part time worker, usual hours worked, occupation, industry and employment status.

For many of the estimates considered in Table 9, combining the main sample with the top-up sample provides an improved estimate compared to just using the main sample. Most of this gain is in improving the standard errors of the estimates, though there are substantial gains from reducing the bias with the combined sample for estimates of country of birth and year of recent arrival.

There are two variables for which the combined estimate is further away from the ABS estimate than the estimate from the main HILDA sample: hours worked and highest level of education. For these two variables there are differences in the collection methodology or questions asked that may limit the validity of these comparisons. The Labour Force Survey obtains information about all adults in the household from any responsible adult whereas the HILDA Survey interviews each adult in the household. Wooden, Wilkins and McGuinness (2007) shows that probably for this reason the HILDA estimates on hours worked align more closely with the ABS Survey of Employment Arrangements and Superannuation, where all adults are interviewed, than the Labour Force Survey. To some extent this collection methodology will also impact on the highest level of education information collected with qualifications not being known to the responsible adult in the household. Further the education questions are quite different between the HILDA Survey and the Labour Force Survey. Respondents to the HILDA Survey are asked to recount all of their education qualifications in their first interview and this is updated over time with subsequent education activity reported in later interviews. The ABS question in the Labour Force Survey asks for the highest level of education. It is possible that the respondent filters out some less important or less relevant qualifications when answering the more aggregated question used by the ABS. There is also some suggestion of this in the HILDA Survey, with wave 1 respondents and wave 11 top-up respondents aged 15-64 showing fewer Certificate III or IV and fewer graduate diplomas or certificates than respondents aged 15-64 in other waves. Nevertheless, the differences between estimates from the main sample and the combined sample for these two variables are generally less than 1 percentage point.

Table 9: Comparison of cross-sectional estimates (%) from main and combined HILDA samples with ABS for 2011

	Estimate			*Standard Error*		*RMSE*	
Characteristic	*Main*	*Comb.*	*ABS*	*Main*	*Comb.*	*Main*	*Comb.*
Family-level variables							
Family type (as proportion of all families, excludes lone persons and group households)							
Couple family	82.1	82.3	83.5	0.87	0.71	1.66	**1.45**
Couple family with dependent children	37.0	37.1	36.0	0.84	0.75	**1.28**	1.35
Couple family with children under 15	29.8	29.6	29.3	0.80	0.70	0.92	**0.78**
Lone parent family	16.1	15.5	14.8	0.81	0.61	1.52	**0.92**
Lone parent family with dependent children	9.6	9.8	9.9	0.53	0.46	0.62	**0.46**
Lone parent family with children under 15	7.0	7.3	7.7	0.42	0.38	0.80	**0.51**
Other families	1.8	2.2	1.6	0.36	0.36	**0.38**	0.67
Enumerated adult-level variables							
Relationship in household							
Couple with children < 15	21.6	21.4	22.0	0.62	0.54	**0.75**	0.78
Couple with dependent student (no child<15)	5.2	5.4	4.5	0.35	0.27	**0.83**	1.00
Couple with nondependent children	6.2	5.8	5.1	0.40	0.31	1.14	**0.71**
Couple without children	26.5	26.8	27.7	0.68	0.63	1.37	**1.06**
Lone parent with children<15	2.5	2.6	3.0	0.16	0.14	0.46	**0.35**
Lone parent with dependent student (no child<15)	0.9	0.9	0.8	0.11	0.09	0.15	**0.12**
Lone parent with nondependent children	2.4	2.1	1.7	0.22	0.14	0.71	**0.39**
Dependent student	7.7	7.6	7.1	0.31	0.25	0.61	**0.51**
Nondependent child	10.2	9.2	8.5	0.49	0.35	1.82	**0.80**
Other family member	3.0	3.2	2.6	0.42	0.30	**0.59**	0.71
Unrelated to all HH members	2.1	3.2	5.2	0.20	0.47	3.10	**1.99**
Lone person	11.7	11.7	11.8	0.38	0.34	0.40	**0.38**
Responding person-level variables							
Relationship in household							
Couple with children < 15	21.7	21.6	22.0	0.63	0.56	**0.67**	0.70
Couple with dependent student (no child<15)	5.5	5.7	4.5	0.37	0.29	**1.13**	1.23
Couple with nondependent children	5.7	5.4	5.1	0.38	0.30	0.64	**0.39**
Couple without children	26.3	26.6	27.7	0.68	0.62	1.58	**1.28**
Lone parent with children<15	2.7	2.8	3.0	0.17	0.15	0.34	**0.25**
Lone parent with dependent student (no child<15)	1.0	1.0	0.8	0.12	0.10	0.24	**0.20**
Lone parent with nondependent children	2.6	2.2	1.7	0.26	0.16	0.95	**0.55**
Dependent student	8.7	8.6	7.1	0.36	0.29	1.57	**1.46**
Nondependent child	9.6	8.5	8.5	0.47	0.33	1.17	**0.33**
Other family member	2.7	3.2	2.6	0.37	0.30	**0.39**	0.64
Unrelated to all HH members	1.9	3.0	5.2	0.17	0.43	3.27	**2.25**
Lone person	11.6	11.6	11.8	0.37	0.34	0.42	**0.40**
Highest level of education (15-64 year olds)							
Postgraduate (masters or doctorate)	4.1	5.5	4.6	0.28	0.41	**0.54**	0.96

	Estimate			*Standard Error*		*RMSE*	
Characteristic	*Main*	*Comb.*	*ABS*	*Main*	*Comb.*	*Main*	*Comb.*
Grad diploma or grad certificate	5.0	5.3	2.1	0.29	0.28	**2.93**	3.19
Bachelor or honours	14.6	15.7	17	0.51	0.54	2.50	**1.39**
Advanced diploma or diploma	8.6	8.9	9.1	0.33	0.31	0.58	**0.38**
Cert IV or III	21.3	20.6	17.4	0.53	0.48	3.97	**3.20**
Year 12	19.0	17.9	20.6	0.50	0.42	**1.68**	2.73
Year 11 or below (inc Cert I, II, nfd)	27.2	26.1	29.1	0.64	0.56	**2.00**	3.08
Undetermined	0.1	0.1	NA	0.11	0.06		
Country of birth							
Australia	75.5	70.0	70.1	0.88	1.06	5.41	**1.07**
Main English speaking country	8.8	10.8	10.8	0.38	0.56	1.98	**0.57**
Other country	15.7	19.2	19.1	0.89	1.07	3.51	**1.07**
Year of arrival (if born overseas)							
Before 1971	27.9	22.3	24.0	1.45	1.23	4.13	**2.11**
1971-1980	14.2	11.2	11.6	1.09	0.77	2.75	**0.90**
1981-1990	25.7	17.0	16.8	1.67	1.10	9.10	**1.12**
1991-2000	24.6	16.5	16.4	1.98	1.15	8.47	**1.15**
2001-2005	4.0	9.7	9.8	0.57	1.10	5.83	**1.11**
2005-2010	3.5	20.4	18.7	0.51	1.98	15.23	**2.57**
2011	0.1	3.0	2.7	0.05	0.63	2.53	**0.71**
Indigenous	2.5	2.2	2.1	0.29	0.20	0.49	**0.24**
Employed persons							
Part time worker	32.0	32.8	30.6	0.65	0.73	**1.54**	2.38
Usual hours worked							
0	0.1	0.0	0.2	0.02	0.02	**0.18**	0.19
1-15	12.4	12.7	11.6	0.41	0.38	**0.87**	1.13
16-29	13.5	14.0	13.0	0.49	0.60	**0.75**	1.21
30-34	6.0	6.1	5.7	0.30	0.26	**0.39**	0.44
35-39	19.2	19.6	23.4	0.58	0.54	4.21	**3.83**
40	16.9	16.3	19.7	0.51	0.46	**2.88**	3.42
41-44	4.2	4.3	3.2	0.28	0.27	**1.05**	1.08
45-49	9.2	9.1	7.1	0.44	0.39	2.18	**2.06**
50-59	11.6	10.9	9.3	0.46	0.38	2.31	**1.69**
60 or more	6.9	6.9	6.7	0.34	0.33	0.40	**0.39**
Occupation							
Managers	13.0	13.2	13.0	0.53	0.46	0.53	**0.52**
Professionals	23.3	23.5	21.6	0.76	0.78	**1.94**	2.13
Technicians and trade workers	14.7	13.7	14.2	0.50	0.42	**0.66**	0.68
Community and personal service workers	9.5	10.0	9.7	0.38	0.36	0.47	**0.43**
Clerical and administrative workers	15.4	14.8	15.1	0.47	0.41	0.57	**0.50**
Sales workers	9.1	9.3	9.4	0.39	0.35	0.49	**0.35**
Machinery operators and drivers	6.0	5.8	6.8	0.35	0.29	**0.91**	1.04
Labourers	9.1	9.7	10.2	0.50	0.61	1.29	**0.83**
Industry							
Agriculture, forestry and fishing	2.4	2.6	2.8	0.35	0.35	0.53	**0.39**

Characteristic	Estimate			Standard Error		RMSE	
	Main	*Comb.*	*ABS*	*Main*	*Comb.*	*Main*	*Comb.*
Mining	1.9	1.9	2.0	0.25	0.22	0.26	**0.23**
Manufacturing	8.4	8.2	8.3	0.48	0.36	0.48	**0.38**
Electricity, gas, water and waste services	1.1	1.0	1.2	0.15	0.11	**0.20**	0.28
Construction	8.4	8.4	9.1	0.44	0.39	0.85	**0.78**
Wholesale trade	3.1	3.4	3.6	0.24	0.22	0.59	**0.29**
Retail trade	10.7	10.6	10.8	0.48	0.41	0.49	**0.45**
Accommodation and food services	6.0	6.1	6.9	0.36	0.33	0.92	**0.82**
Transport, postal and warehousing	5.0	4.7	5.1	0.34	0.27	**0.39**	0.49
Information media and telecommunications	1.8	2.0	1.8	0.20	0.18	**0.21**	0.27
Financial and insurance services	4.1	4.0	3.8	0.34	0.28	0.45	**0.35**
Rental, hiring and real estate services	1.4	1.3	1.7	0.20	0.13	**0.34**	0.48
Professional, scientific and technical services	8.2	8.5	7.7	0.37	0.37	**0.59**	0.83
Administrative and support services	3.4	3.4	3.6	0.38	0.32	0.44	**0.38**
Public administration and safety	6.7	6.3	6.5	0.36	0.30	0.42	**0.35**
Education and training	9.6	9.3	7.6	0.41	0.35	1.99	**1.70**
Health care and social assistance	12.1	12.5	11.7	0.47	0.42	**0.61**	0.94
Arts and recreational services	2.0	1.8	1.8	0.17	0.15	0.24	**0.15**
Other services	3.9	4.0	4.0	0.27	0.25	0.28	**0.26**
Employment status							
Employee	90.5	90.4	89.2	0.47	0.40	1.36	**1.28**
Employer	2.2	2.1	2.9	0.20	0.17	**0.77**	0.81
Own account worker	7.1	7.2	7.8	0.41	0.35	0.86	**0.77**
Contributing family member	0.3	0.3	0.0	0.06	0.06	**0.23**	0.27

Note: ABS estimates for country of birth, year of arrival and indigenous status exclude institutionalised population, otherwise the estimates apply to all civilians aged 15 and over. HILDA estimates also for aged 15 and over including the defence force but excluding institutionalised population and very remote parts of Australia.

ABS sources: i) Family type is from ABS Cat.No. 6224.0.55.001 *Labour Force, Australia: Labour Force Status and Other Characteristic of Families*, June 2011. ii) Relationship in household, country of birth, year of arrival and usual hours worked is from ABS Cat.No. 6291.0.55.001 *Labour Force, Australia, Detailed - Electronic Delivery*, September 2011. iii) Highest level of education is from ABS Cat.No. 62270DO001_201105 *Education and Work, Australia*, May 2011. Indigenous status is from ABS Cat.No. 62870DO001_2011 *Labour Force Characteristics of Aboriginal and Torres Straight Islander Australians*, 2011. iv) Occupation, industry and employment status is from ABS Cat.No. 6291.0.55.003 *Labour Force, Australia, Detailed, Quarterly*, August 2001.

Weights in the HILDA data release

Weights provided

Table 10 provides a list of the weights provided on the data files. The longitudinal weights provided on the enumerated person, responding person, and combined files are the ones users will most likely to use. Other longitudinal weights are provided on the *Longitudinal Weights File*.

Replicate weights have been provided for users to calculate standard errors that take into account the complex sample design of the HILDA Survey. These weights can be used by the SAS GREGWT macro, the STATA 'svy jackknife' commands (more detail is provided in the section below on *Calculating Standard Errors*), or you can write your own routine to use these weights. Weights for 45 replicate groups are provided.

As noted earlier, the weights may change from release to release because of updates to the benchmarks from time to time but also for two other reasons. Firstly, corrections may be made to age and sex variables when these are confirmed with individuals in subsequent interviews. And secondly, duplicate or excluded people in the sample may be identified after the release (very occasionally).

Table 10: Weights provided in the HILDA datasets

Population	*File*	*Weight*	*Replicate Weights*
Longitudinal weights			
Responding persons			
Continuous balanced panel from wave 1 to wave "_"	Responding person file; Combined file	_lnwtrp	_rwln1 to _rwln45
Continuous balanced panel from wave *t1* to wave *tn*	Longitudinal weights file	wlr*t1_tn*	wlr*t1_tn1* to wlr*t1_tn45*
Paired balanced panel for wave *t1* and *tn*	Longitudinal weights file	wlr*t1tn*	wlr*t1tn1* to wlr*t1tn45*
Balanced panel for retirement module waves (3, 7, 11)	Longitudinal weights file	wlrc__k	wlrc__k*1* to wlrc__k*45*
Balanced panel for fertility module waves (waves 5, 8, 11)	Longitudinal weights file	wlre__k	wlre__k*1* to wlre__k*45*
Enumerated persons			
Continuous balanced panel from wave 1 to wave "_"	Enumerated person file; Combined file	_lnwte	_rwlne1 to _rwlne45
Continuous balanced panel from wave *t1* to wave *tn*	Longitudinal weights file	wle*t1_tn*	wle*t1_tn1* to wle*t1_tn45*
Paired balanced panel for wave *t1* and *tn*	Longitudinal weights file	wle*t1tn*	wle*t1tn1* to wle*t1tn45*
Balanced panel for wealth module waves (2, 6, 10)	Longitudinal weights file	wleb__j	wleb__j*1* to wleb__j*45*
Cross-sectional weights*			
Households	Household file; Combined file	_hhwth	_rwh1 to _rwh45
Responding persons	Responding person file; Combined file	_hhwtrp	_rwrp1 to _rwrp45
Enumerated persons	Enumerated person file; Combined file	_hhwte	_rwe1 to _rwe45

Note: "_", "*t1*" and "*tn*" indicate wave letters (*a* for wave 1, *b* for wave 2, *c* for wave 3, etc). The Longitudinal Replicate Weights File is available on request. Please email hilda-inquiries@unimelb.edu.au.
* *khhwthm, khhwtrpm,* and *khhwtem* have also been provided for Release 11. These weights exclude the top-up sample in Wave 11 so users can check how different their cross-sectional estimates would be without the top-up sample.

Plan for inclusion of the top-up sample into the weights in future releases

As there are already a broad range of weights provided for the HILDA data, the wave 11 top-up sample will be integrated into the weights from wave 11 onwards as shown in Table 11. The cross-sectional weights are denoted by a wave number (1, 2, etc) whereas the longitudinal weights are denoted by a wave combination: using a '-' to indicate a continuous panel, or an '&' to denote a panel for a pair of waves. Further some three or more wave non-continuous combinations are provided for wave combinations that carry the same special module (such as wealth, fertility, retirement, and eventually health and human capital). For a particular release, the weights listed under that release and all prior releases would be included in the data files. The weights that include the top-up sample are shown in bold. The top-up sample will be included in the cross-sectional weights from wave 11 onwards and in the longitudinal weights that start in wave 11 or later.

The wave 11 cross-sectional weights excluding the top-up sample are also provided for users who wish to examine the impact of the top-up sample on the weights. Such weights will not be provided after wave 11.

Table 11: Inclusion of the top-up sample in the cross-sectional and longitudinal weights (in bold)

	Release													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cross-section	1	2	3	4	5	6	7	8	9	10	**11**	**12**	**13**	**14**
Longitudinal														
Continuous		1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13	1-14
			2-3	2-4	2-5	2-6	2-7	2-8	2-9	2-10	2-11	2-12	2-13	2-14
														…
											10-11	10-12	10-13	10-14
												11-12	**11-13**	**11-14**
													12-13	**12-14**
														13-14
Pair		1&2	1&3	1&4	1&5	1&6	1&7	1&8	1&9	1&10	1&11	1&12	1&13	1&14
			2&3	2&4	2&5	2&6	2&7	2&8	2&9	2&10	2&11	2&12	2&13	2&14
				…	…	…	…	…	…	…	…	…	…	…
											10&11	10&12	10&13	10&14
												11&12	**11&13**	**11&14**
													12&13	**12&14**
														13&14
4-year cycle														
Wealth										2,6,10				2,6,10,14
Retirement											3,7,11			
3-year cycle														
Fertility											5,8,11			

Which weight to use

For some users, the array of weights on the dataset may seem confusing. This section provides examples of when it would be appropriate to use the different types of weights.

If you want to make inferences about the Australian population from frequencies or cross-tabulations of the HILDA sample then you will need to use weights. If you are only using information collected during the wave 4 interviews (either at the household level or person level) then you would most likely use the wave 4 cross-section weights. If you want to infer how people have changed across the five years between waves 1 and 6, then you would use the longitudinal weights for the balanced panel from wave 1 to 6.

The following five examples show how the various weights may be used to answer questions about the population:

- What is the average salary of professionals in 2003? This is a question that can only be answered from the responding person file using the cross-section responding person weight for wave 3. We would identify those reportedly working in professional occupations and take the weighted average of their wages and salaries.

- How many people live in poor households in 2002? We are interested in the number of individuals with a certain household characteristic, such as having low equivalised disposable household incomes. We would use the cross-section enumerated person weight for wave 2 and count the number of enumerated people in households with poorest 10 per cent of equivalised household incomes. (We do not need to restrict our attention to responding persons only as total household incomes are available for all households after the imputation process. We also want to include children in this analysis and not just limit our analysis to those aged 15 year or older.)
- For how many years have people been poor between 2001 and 2006? We might define the 'poorest' 10 per cent of households as having the lowest equivalised household incomes in each wave. We could then calculated how many years people were poor between wave 1 and wave 6, and apply the longitudinal enumerated person weight (*flnwte* or equivalently *wlea_f*) for those people enumerated every wave between wave 1 and 6.
- What proportion of people have changed their employment status between 2002 and 2007? This question can only be answered by considering the responding persons in both waves. We would use the longitudinal responding person weight for the pair of waves extracted from the Longitudinal Weight File (*wlrbg*) and construct a weighted cross-tabulation of the employment status of respondents in wave 2 against the employment status of respondents in wave 7.

When constructing regression models, the researcher needs to be aware of the sample design and non-response issues underlying the data and will need to take account of this in some way.

Calculating standard errors

The HILDA Survey has a complex survey design that needs to be taken into account when calculating standard errors. It is clustered, stratified and the weights are not all equal. Applying weights will correct point estimates, but appropriate standard errors and confidence intervals will not be calculated unless the stratification and the clustering are taken into account. Some options available for calculating appropriate standard errors and confidence intervals include:

- Use of 'svy' commands in STATA – Stata has a set of survey commands that deal with complex survey designs. Using the 'svyset' commands, the clustering, stratification and weights can be assigned. You can request the standard errors be calculated using the Jackknife method using 'svy jackknife' and the replicate weights. Various statistical procedures are available within the suite of 'svy' commands including means, proportions, tabulations, linear regression, logistic regression, probit models and a number of other commands.
- Use of SAS procedures SURVEYMEANS, SURVEYREG, SURVEYFREQ and SURVEYLOGISTIC (the last two only in SAS Version 9 onwards). The SAS procedures provide standard errors via the Taylor Series approximation. SAS does not have a built in feature to handle replicates weights, however, a SAS macro has been provided by one of our users in the program library.
- Use of GREGWT macro in SAS – Some users within FaHCSIA, ABS and other organisations may have access to the GREGWT macro (written by the ABS Methodology Division) that can be used to construct various population estimates. The macro uses the jackknife method to estimate standard errors using the replicate weights.
- Use of the SPSS add-on module "SPSS Complex Samples" (available from SPSS Release 12). The add-on module produces standard errors via the Taylor Series approximation. SPSS does not have a built in feature to handle replicates weights.
- Standard Error Tables – Based on the wave 1 data, approximate standard errors have been constructed for a range of estimates (see Horn, 2004). Similar tables for later waves have not been produced.

A user guide for calculating the standard errors in HILDA is provided as part of our technical paper series, see Hayes (2008). Example code is provided in SAS, SPSS and STATA. Note however that the name of the sample design variables have changed: *xhhraid* refers to the randomised area id and *xhhstrat* refers to the wave 1 proxy stratification. Also, the multiplier you need to use in the Jackknife method is 44/45 (i.e., 0.977778).

To assist you in the calculation of appropriate standard errors, the wave 1 area (cluster), and proxy stratification variables have been included all files. These are listed in Table 12 and need to be specified for the standard error calculations using the Taylor Series approximation method as suggested above. Any new entrants to the household are assigned to the same sample design information as the permanent sample member they join.

Table 12: Sample design variables

Variable	*Description*	*Design element*
xhhraid	DV: randomised area id	Cluster
xhhstrat	DV: original strata	Proxy stratification

References

De Leeuw, E. and De Heer, W. (2002), Trends in household survey response: a longitudinal and international comparison. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 41-54. John Wiley and Sons, New York.

Hayes, C. (2008), 'HILDA Standard Errors: User Guide', HILDA Project Technical Paper Series No. 2/08, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Horn, S. (2004), 'Guide to Standard Errors for Cross Section Estimates', HILDA Project Technical Paper Series No. 2/04, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Kaminska, O, and Lynn, P. (2012), 'Combining refreshment or boost samples with an existing panel sample: challenges and solutions', Background paper for the 2012 Panel Survey Methods Workshop, 4-5 July 2012, Melbourne.

LaRoche, S. (2003), 'Longitudinal and cross-sectional weighting of the Survey of Labour and Income Dynamics', Income Research Paper Series, Cat.No. 75F0002MIE, No. 007, Statistics Canada.

O'Muircheartaigh, C. and Pedlow, S. (2002), 'Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97', *ASA Proceedings of the Joint Statistical Meetings*, pp. 2557-2562, American Statistical Association, Alexandria, VA.

Spiess, M., and Rendtel, U. (2000), 'Combining an ongoing panel with a new cross-sectional sample', DIW Discussion Paper Series, No. 198, DIW, Berlin.

Stukel, Hidiroglou and Sarndal (1996), 'Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linerization', *Survey Methodology*, Vol. 22, No. 2, pp. 117-125.

Summerfield, M., Freidin, S., Hahn, M., Ittak, P., Li, N., Macalalad, Watson, N., Wilkins, R., Wooden, M. (2012), 'HILDA User Manual – Release 11', Melbourne Institute of Applied Economic and Social Research.

Watson, N. (2006), 'Options for a top-up sample to the HILDA Survey', *Proceedings of the ACSPRI Social Science Methodology Conference*, University of Sydney, Australia, 10-13 December 2006.

Watson, N. (2011), 'Methodology for the HILDA top-up sample', *HILDA Technical Paper Series*, No. 1/11, Melbourne Institute of Applied Economic and Social Research.

Watson, N., and Fry, T. (2002), 'The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Weighting', *HILDA Technical Paper Series*, No. 3/02, Melbourne Institute of Applied Economic and Social Research.

Watson, N., and Wooden, M. (2002), 'The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology', *HILDA Technical Paper Series*, No. 1/02, Melbourne Institute of Applied Economic and Social Research.

Watson, N., and Wooden, M. (2004), 'Sample Attrition in the HILDA Survey', *Australian Journal of Labour Economics*, vol. 7, June, pp. 293-308.

Watson, N., and Wooden, M. (2009), 'Identifying Factors Affecting Longitudinal Survey Response', in P. Lynn (ed.), *Methodology of Longitudinal Surveys*, John Wiley and Sons, Chichester, pp. 157-181.

Watson, N., and Wooden, M. (2011), 'Re-engaging with Survey Non-respondents: The BHPS, SOEP and HILDA Survey Experience', HILDA Project Discussion Paper Series No. 1/11, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Wooden, M, Wilkins, R., and McGuinness, S. (2007), 'Minimum wages and the working poor'', *Economic Papers*, Vol. 26, No. 4, pp. 295-307.

Appendix 1: Comparison of design-adjusted HILDA estimates to ABS estimates

Table A1.1: Characteristics of wave 1 and wave 11 top-up samples compared to 2001 and 2011 ABS Census[a]

	HILDA wave 1				HILDA wave 11 top-up			
	Selected dwell.	Resp. HH	P(resp and non-resp HH same)	ABS Census 2001	Selected dwell.	Resp. HH	P(resp and non-resp HH same)	ABS Census 2011
Area								
Sydney	21.0	16.9		20.3	19.4	17.3		19.6
Rest of NSW	13.4	14.6		12.8	11.6	13.0		12.2
Melbourne	17.6	16.7		17.6	18.8	17.0		18.4
Rest of Vic	6.7	7.5		6.9	6.2	6.5		6.6
Brisbane	8.8	8.8		8.5	9.9	9.8		9.4
Rest of Qld	10.4	11.8		10.6	10.9	11.1		10.5
Adelaide	6.1	6.1		6.1	6.2	7.5		6.1
Rest of SA	1.9	2.3		2.2	2.0	2.1		1.9
Perth	7.3	7.7		7.2	7.0	7.0		8.1
Rest of WA	2.4	2.8		2.6	2.7	3.2		2.2
Tasmania	2.6	2.8		2.6	3.3	3.7		2.5
Northern Territory	0.6	0.5		0.9	0.4	0.4		0.8
ACT	1.4	1.5	<.0001	1.6	1.6	1.4	<.0001	1.7
Dwelling type[b]								
Separate house	76.4	77.8		75.3	73.6	73.4		75.6
Semi-detached	9.8	10.1		8.9	10.2	10.9		9.9
Flat	13.4	11.8		13.1	15.1	14.6		13.6
Other	0.4	0.4	<.0001	1.9	1.0	1.1	0.534	0.9
Condition of dwelling								
Very good / excellent	33.8	33.8			35.3	34.4		
Good	35.6	36.2			35.3	35.5		
Average	25.2	25.0			25.4	25.8		
Poor	4.9	4.6			3.9	4.1		
Very poor / almost derelict	0.5	0.4	0.031		0.2	0.2	0.589	
Security features								
Locked gate (w/o intercom)	3.3	2.8	<.0001		4.2	3.7	0.019	
Locked door / gate (w intercom)	5.6	4.1	<.0001		8.2	6.7	0.001	
Security guard, doorman etc	1.3	1.1	0.326		1.1	0.7	0.003	
Bars on windows	4.9	4.9	0.742		1.9	1.9	0.545	
Security door	37.0	37.4	0.156		26.2	25.7	0.532	
No trespassing sign	0.5	0.5	0.510					
Beware of dog sign	2.2	2.2	0.905					
Evidence of dangerous dog	2.5	2.3	0.233		1.1	0.9	0.196	
No junk mail sign	3.5	3.3	0.009		5.3	5.2	0.858	
Neighbourhood watch sign	3.7	3.6	0.728					
Rails / ramp access					0.9	0.9	0.950	
Roller shutters					2.4	2.8	0.199	
High-rise buildings in area								
A lot - more than 50%	1.7	1.0						
A fair bit - 20-50%	1.0	0.8						
One or two	2.4	2.1						
None at all	94.9	96.1	<.0001					

	HILDA wave 1				HILDA wave 11 top-up			
	Selected dwell.	Resp. HH	P(resp and non-resp HH same)	ABS Census 2001	Selected dwell.	Resp. HH	P(resp and non-resp HH same)	ABS Census 2011
Likely dwelling contains children								
Very likely					5.2	6.3		
Likely					8.5	8.5		
Unlikely					20.9	20.2		
Very unlikely					14.7	15.8		
Cannot tell from observation					50.7	49.2	0.001	
Overgrown / unkept garden								
Yes					10.2	11.1		
No					80.0	80.4		
No obvious garden					9.8	8.5	0.010	
Type of road								
Not on main / major road					79.0	78.9		
Main road - single lane					14.6	15.5		
Main road - two or more lanes					6.4	5.6	0.030	

Note: a. HILDA estimates are weighted to adjust for variation in probability of selection (i.e. by the design weight).
b. Excludes small portion of cases where dwelling structure is not able to be classified.

Table A1.2: Selected wave 1 and wave 11 top-up individual characteristics compared to 2001 and 2011 ABS Labour Force Survey estimates[a]

	HILDA wave 1				HILDA wave 11 top-up			
	Enum. adults	Resp. adults	P(resp and non-resp HH same)	ABS LFS 2001	Enum. adults	Resp. adults	P(resp and non-resp HH same)	ABS LFS 2011
Area								
Sydney	18.2	16.9		21.5	17.7	17.4		20.4
Rest of NSW	14.0	14.5		12.2	13.3	13.5		12.0
Melbourne	17.7	17.4		18.4	18.0	17.8		18.4
Rest of Vic	7.3	7.5		6.7	6.3	6.3		6.7
Brisbane	8.8	8.8		8.6	9.6	9.8		8.9
Rest of Qld	11.3	11.4		10	10.3	10.2		11.1
Adelaide	5.8	6.0		5.8	6.9	7.3		5.5
Rest of SA	2.2	2.4		2	2.1	2.2		2.0
Perth	7.4	7.5		7.3	6.9	6.3		7.7
Rest of WA	2.7	2.7		2.5	3.4	3.5		2.6
Tasmania	2.7	2.7		2.4	3.5	3.7		2.2
Northern Territory	0.5	0.5		0.9	0.4	0.5		0.9
ACT	1.6	1.6	<.0001	1.6	1.5	1.5	<.0001	1.6
Sex								
Male	47.6	46.2		49.3	48.4	47.8		49.3
Female	52.4	53.8	<.0001	50.7	51.6	52.2	0.006	50.7
Age (years) at 30 Sept								
15-19	11.2	10.9		8.8	8.8	8.5		8.1
20-24	8.4	7.8		8.9	8.7	8.3		9.0
25-34	18.8	18.7		18.7	16.8	17.1		17.8
35-44	21.1	21.3		19	18.4	18.6		17.3
45-54	16.6	16.6		17.1	15.9	15.8		16.6
55-64	11.1	11.4		11.8	14.2	14.5		14.1
65 or over	12.8	13.3	<.0001	15.6	17.1	17.3	<.0001	17.1
Marital status								
Married (including defacto)	62.8	63.8		58.7	62.7	63.2		58.2
Not married	37.2	36.2	<.0001	41.3	37.3	36.8	0.030	41.8
Relationship in household								
Couple with children < 15	29.6	30.6		26.3	28.4	29.2		24.9
Couple with dependent student (no child<15)	4.5	4.4		5.1	5.3	5.2		5.1
Couple with nondependent children	5.6	5.4		6.5	4.9	4.5		5.8
Couple without children	30.1	31.5		30.5	33.4	34.6		31.4
Lone parent with children<15	3.2	3.5		3.7	3.7	4.0		3.4
Lone parent with dependent student (no child<15)	0.6	0.6		0.7	0.8	0.8		0.9
Lone parent with nondependent children	1.3	1.3		1.8	1.4	1.4		1.9
Dependent student	8.3	8.5		7.7	7.9	7.9		8.1
Nondependent child	10.0	8.3		9.3	8.0	6.8		9.6
Other family member	2.9	2.5		2.6	3.0	2.9		3.0
Unrelated to all HH members	3.8	3.4	<.0001	6.0	2.9	2.7	<.0001	5.9
Lone person in household	10.2	11.2		12.0	13.1	14.0		11.8
Indigenous status								
Indigenous		1.8		1.7		2.5		2.1

	HILDA wave 1				HILDA wave 11 top-up			
	Enum. adults	Resp. adults	P(resp and non-resp HH same)	ABS LFS 2001	Enum. adults	Resp. adults	P(resp and non-resp HH same)	ABS LFS 2011
Non-indigenous		98.2		98.3		97.5		97.9
Birthplace								
Born in Australia		74.6		72.4		68.7		70.1
Main English speaking country		10.7		10.2		13.0		10.8
Other country		14.7		17.5		18.3		19.1
Labour force status								
Employed								
Full-time		41.2		42.1		41.4		44.3
Part-time		20.2		17.4		20.9		18.3
Unemployed		4.5		4.3		3.6		3.4
Not in the Labour force		34.2		36.3		34.0		34.0
Employment status in main job (employed persons only)								
Employee		87.3		86		90.3		89.2
Employer		3.8		3.6		2.0		2.9
Own account worker		8.1		10		7.3		7.8
Contributing family worker		0.8		0.4		0.4		0.0

Appendix 2: Response models

Table A2.1: Logistic regression model of household-level response, waves 1 and wave 11 top-up compared

Variable	Wave 1		Wave 11 Top-up Model A		Wave 11 Top-up Model B	
	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
Security features						
Locked gate – no intercom access	0.641	<.0001	0.606	0.0097	0.576	0.0034
Locked gate – intercom access	0.686	0.000	0.559	0.0004	0.541	<.0001
Security guard/doorman/on-site manager/gatekeeper	0.792	0.270	0.47	0.1257	0.579	0.2623
Security door	0.904	0.025	0.769	0.0109	0.811	0.0296
No trespassing sign	1.24	0.478	-	-	-	-
Beware of dog sign	1.003	0.983	-	-	-	-
Evidence of a dangerous dog	0.789	0.067	0.607	0.1675	0.608	0.1648
No junk mail sign/no hawkers sign	0.781	0.019	0.974	0.885	0.904	0.5745
Neighbourhood watch sign	1.003	0.980	-	-	-	-
Bars on windows	1.066	0.513	1.031	0.9133	0.964	0.8922
Grab rails / ramp access / other mods	-	-	0.748	0.4964	-	-
Roller shutters	-	-	1.272	0.3822	-	-
Dwelling type (base=separate house)						
Semi-detached	1.117	0.122	1.352	0.0398	1.343	0.0332
Flat/unit/apartment – 1-2 storey	1.067	0.406	-	-	-	-
Flat/unit/apartment – 3+ storey	0.785	0.050	-	-	-	-
Other dwelling (caravan, tent, etc)	1.458	0.192	-	-	-	-
Flat/unit/apartment/other	-	-	1.433	0.0143	1.288	0.0549
External condition of dwelling (base='Very good/excellent')						
Good	1.014	0.783	1.036	0.7224	1.043	0.6676
Average	0.934	0.213	1.082	0.4902	1.093	0.4137
Poor	0.713	0.000	1.288	0.3037	1.513	0.0706
Very poor/almost derelict	0.483	0.006	0.438	0.2626	0.464	0.2937
Household likely contains children < 15 (base=Very likely)						
Likely	-	-	0.38	0.0002	-	-
Unlikely	-	-	0.322	<.0001	-	-
Very unlikely	-	-	0.429	0.0006	-	-
Cannot tell from observation	-	-	0.391	<.0001	-	-
Unkempt/overgrown garden (base=Yes)						
No	-	-	0.742	0.047	-	-
No obvious garden	-	-	0.635	0.0201	-	-
Type of road (base=not main road)						
Main road, single lane	-	-	1.039	0.7464	-	-
Main road, two or more lanes	-	-	0.826	0.2508	-	-
Highrises in area (base=no highrises)						
A lot - more than 50%	0.719	0.077	-	-	-	-
A fair bit - more than 20%	1.351	0.156	-	-	-	-
One or two such structures	0.891	0.362	-	-	-	-
Geographic location (base=Melbourne)						
Sydney	0.775	0.000	1.129	0.4294	1.041	0.7625
Brisbane	1.041	0.653	0.927	0.6806	0.956	0.7908
Adelaide	1.076	0.450	2.654	<.0001	2.642	<.0001
Perth	1.254	0.015	1.106	0.6074	1.108	0.5723
Tasmania	1.039	0.792	-	-	-	-
Northern Territory	0.54	0.011	-	-	-	-
Australian Capital Territory	1.492	0.031	-	-	-	-
Rural New South Wales	1.274	0.005	-	-	-	-

Variable	Wave 1		Wave 11 Top-up Model A		Wave 11 Top-up Model B	
	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
Rural Victoria	1.301	0.012	-	-	-	-
Rural Queensland	1.444	0.000	-	-	-	-
Rural South Australia	1.927	0.000	-	-	-	-
Rural Western Australia	1.657	0.002	-	-	-	-
Tas, NT, ACT	-	-	1.509	0.0662	1.597	0.0422
Rural NSW, Vic, Qld, SA, WA	-	-	1.4	0.0267	1.344	0.045
Neighbourhood characteristics						
Population density (per km^2)	1	0.002	1.000	0.7247	1	0.5428
Proportion speaking language other than English	0.762	0.084	0.419	0.0238	0.443	0.0025
Proportion of people not in labour force	0.91	0.709	-	-	-	-
Proportion of people unemployed	1.62	0.616	-	-	-	-
Median weekly household income >=$1200	-	-	1.295	0.0438	1.273	0.0532
Median age	-	-	0.976	0.0366	-	-
Average household size	-	-	0.695	0.1481	-	-
Proportion of flats	-	-	0.339	0.0447	-	-
SEIFA index of education and occupation	0.994	0.653	-	-	-	-
SEIFA index of education and occupation squared	1.000	0.566	-	-	-	-
SEIFA index of advantage	0.995	0.807	-	-	-	-
SEIFA index of advantage squared	1.000	0.820	-	-	-	-
SEIFA index of economic advantage	1.003	0.785	0.976	0.073	-	-
SEIFA index of economic advantage squared	1.000	0.662	1.000	0.0958	-	-
SEIFA index of disadvantage	-	-	1.016	0.3005	-	-
SEIFA index of disadvantage squared	-	-	1.000	0.3042	-	-

Table A2.2: Logistic regression model of individual-level response in 2+ adult households, waves 1 and wave 11 top-up compared

	Wave 1		Wave 11 Top-up	
Variable	Odds Ratio	P-value	Odds Ratio	P-value
Geographic location (base=Melbourne)				
Sydney	0.594	<.0001	0.84	0.3867
Brisbane	0.901	0.435	0.923	0.7807
Adelaide	1.83	0.001	8.687	0.003
Perth	0.985	0.918	0.325	<.0001
Tasmania	1.117	0.640	-	-
Northern Territory	0.752	0.499	-	-
Australian Capital Territory	1.306	0.360	-	-
Rural New South Wales	1.427	0.008	-	-
Rural Victoria	1.066	0.675	-	-
Rural Queensland	1.021	0.876	-	-
Rural South Australia	2.032	0.013	-	-
Rural Western Australia	1.288	0.312	-	-
Tas, NT, ACT	-	-	3.605	0.0343
Rural NSW, Vic, Qld, SA, WA	-	-	1.017	0.93
Labour force status (base=emp. full time)				
Employed part time	1.838	<.0001	1.698	0.0077
Unemployed	1.968	<.0001	1.165	0.6802
Not in labour force	1.703	<.0001	1.526	0.0252
Female	1.597	<.0001	1.207	0.1704
Age group (base=15-19)				
20-24	0.724	0.014	0.788	0.3632
25-34	0.711	0.011	1.046	0.8789
35-44	0.743	0.035	0.761	0.3649
45-54	0.831	0.198	0.766	0.3671
55-64	0.874	0.418	0.834	0.5723
65+	0.73	0.069	0.494	0.029
Three or more adults in HH (compared to 2)	0.4	<.0001	0.313	<.0001
Number of children in HH (base= zero children)				
One child	1.348	0.002	1.808	0.0035
Two or more children	1.254	0.020	0.944	0.769
Married or defacto	2.025	<.0001	1.841	0.0014
English ability (base=only speaks English at home)				
Well or very well	12.578	<.0001	0.848	0.3453
Not well	8.226	<.0001	0.307	0.0001
Not at all well	3.365	0.000	0.186	0.0012
Dwelling type (base=separate house)				
Semi-detached	1.04	0.762	0.777	0.2544
Flat/unit/apartment – 1-2 storey	1.556	0.016	-	-
Flat/unit/apartment – 3+ storey	1.207	0.343	-	-
Other dwelling - caravan, tent, cabin, etc	0.854	0.798	-	-
Flat/unit/apartment/other	-	-	1.347	0.2591

Table A2.3: Logistic regression model of response for balanced panel of enumerated persons and respondents between waves 1 and 2

Variable	Balanced panel of enum persons (ivwd wave1)		Balanced panel of enum persons (not ivwd wave 1)		Balanced panel of responding persons	
	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
Wave 1 person characteristics						
Female	1.049	0.606	1.66	0.251	1.027	0.6742
Age	1.067	<.0001	0.928	0.284	1.082	<.0001
Age squared	0.999	<.0001	1.001	0.397	0.999	<.0001
Female aged 65 or over	0.897	0.612	16.563	0.284	1.016	0.9203
Marital status (base category married)						
De facto	0.831	0.1763	-	-	0.823	0.0539
Separated	1.632	0.1054	-	-	1.228	0.3636
Divorced	1.565	0.1057	-	-	1.122	0.5626
Widowed	2.195	0.0144	-	-	1.805	0.0163
Never married	1.493	0.125	-	-	1.244	0.1946
Ability in speaking English (base category English only language spoken)						
Speaks English well or very well	0.74	0.0524	-	-	0.883	0.2714
Speaks English not well	0.577	0.0443	-	-	0.768	0.2227
Speaks English not at all	1.478	0.6552	-	-	0.756	0.5471
Employment status and hours (base category not in labour force)						
Unemployed	0.698	0.0411	-	-	0.866	0.2712
Employed less than 25 hrs pw	1.121	0.4478	-	-	1.067	0.5126
Employed 25 to 34 hrs pw	0.794	0.1953	-	-	0.748	0.023
Employed 35 to 44 hrs pw	0.725	0.0124	-	-	0.721	0.0003
Employed 45 to 54 hrs pw	0.923	0.6232	-	-	0.857	0.1826
Employed 55 or more hrs pw	0.74	0.0872	-	-	0.7	0.0038
Number of children have	1.005	0.8801	-	-	0.999	0.9678
Country of birth (base category Australia)						
Main English speaking country	0.807	0.0819	-	-	0.811	0.0218
Main non-English speaking country	0.83	0.2008	-	-	0.848	0.1254
Highest level of education achieved (base category yr12 or below)						
Certificate or diploma	0.993	0.939	-	-	1.119	0.093
Bachelor or post-graduate	2.02	<.0001	-	-	1.892	<.0001
Relationship in household (base category couple with child under 15)						
Couple with dependent student	0.612	0.3164	0.074	0.042	0.762	0.2221
Couple with non-dependent child	0.741	0.4619	0.896	0.919	0.932	0.7535
Couple without children	0.552	0.1093	0.097	0.051	0.967	0.8926
Lone parent with child under 15	0.659	0.3431	0.475	0.678	0.612	0.0918
Lone parent with dependent child	0.473	0.263	-	-	0.752	0.5031
Lone parent with non-dependent child	1.568	0.3413	0.058	0.069	1.321	0.401
Other family member	0.357	0.006	0.261	0.254	0.588	0.0253
Lone person	0.612	0.3348	-	-	0.981	0.9612
Unrelated to all HH members	0.295	0.0013	0.123	0.114	0.507	0.0091
Health status (base category excellent)						
Very Good	1.122	0.2807	-	-	1.106	0.1831
Good	1.184	0.1396	-	-	1.04	0.6231
Fair	1.234	0.141	-	-	1.086	0.4269
Poor	0.768	0.1605	-	-	0.811	0.1542
Likelihood of moving (base category not likely to move)						
Not sure if moving	1.055	0.6716	-	-	1.035	0.7122
Likely or very likely to move	1.366	0.0055	-	-	1.207	0.0223

Variable	Balanced panel of enum persons (ivwd wave1)		Balanced panel of enum persons (not ivwd wave 1)		Balanced panel of responding persons	
	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
Number of times moved in last 10 yrs (base category no moves)						
Moved 1 to 2 times in last 10 yrs	0.794	0.0505	-	-	0.893	0.1548
Moved 3 to 4 times in last 10 yrs	0.801	0.0791	-	-	1.023	0.7962
Moved 5 to 9 times in last 10 yrs	0.946	0.6975	-	-	1.049	0.6248
Moved 10 or more times in last 10 yrs	0.66	0.0232	-	-	0.776	0.0649
Moved unknown number of times in last 10 yrs	0.115	0.0487	-	-	0.824	0.8403
Length of PQ ivw in w1	1.003	0.4374	-	-	0.999	0.7111
Length of PQ ivw unknown	1.347	0.3718	-	-	0.932	0.7597
Whether completed SCQ in W1	0.466	<.0001	-	-	0.448	<.0001
Whether reference person in HH	1.329	0.0842	-	-	1.038	0.5825
Wave 1 interview situation						
Respondent's cooperation was fair, poor or very poor	-	-	-	-	0.476	<.0001
Interview was assisted	-	-	-	-	1.033	0.683
English was a problem as it was a second language	-	-	-	-	0.848	0.2711
Eyesight was a problem	-	-	-	-	1.417	0.2825
Hearing was a problem	-	-	-	-	1.005	0.9825
Other language problems occurred	-	-	-	-	0.76	0.3443
Reading was a problem	-	-	-	-	0.982	0.9252
Respondent was somewhat or very suspicious of interview	-	-	-	-	0.611	<.0001
Respondent's understanding was fair, poor or very poor	-	-	-	-	0.869	0.2528
Other adults influenced the interview	-	-	-	-	0.9	0.0636
Wave 1 household characteristics						
Location (base category Sydney)						
Rural NSW	0.865	0.4263	1.797	0.501	1.011	0.9335
Melbourne	0.826	0.132	0.093	0.001	1.043	0.631
Rural Vic	0.783	0.2576	2.281	0.488	0.842	0.2657
Brisbane	1.004	0.9801	0.482	0.297	1.208	0.1115
Rural Qld	1.319	0.1606	1.36	0.743	1.403	0.0182
Adelaide	1.06	0.7555	0.076	0.042	1.281	0.0662
Rural SA	1.005	0.9884	0.511	0.635	1.432	0.1208
Perth	1.049	0.7793	0.383	0.256	1.388	0.0087
Rural WA	0.836	0.5465	2.19	0.667	0.896	0.6101
Tas	0.774	0.3616	3.446	0.452	0.698	0.0679
NT	5.16	0.0433	<0.001	0.986	5.643	0.0068
ACT	1.253	0.5462	-	-	1.267	0.3359
Remoteness Area (base category major cities)						
Inner regional	1.182	0.2904	0.131	0.019	1.347	0.0095
Outer regional	0.861	0.4175	0.575	0.553	0.936	0.6241
Remote	1.001	0.9978	1.013	0.992	1.116	0.649
SEIFA index of disadvantage (base category is lowest decile – most disadvantaged)						
Second decile	0.702	0.0284	1.008	0.991	0.713	0.004
Third decile	0.714	0.0355	0.668	0.597	0.768	0.0232
Fourth decile	0.719	0.0682	0.915	0.917	0.709	0.0089
Fifth decile	0.901	0.5634	0.551	0.519	0.874	0.2941
Sixth decile	0.949	0.7688	0.561	0.514	0.939	0.6167
Seventh decile	0.982	0.919	1.263	0.784	0.883	0.3293
Eighth decile	1.056	0.7694	0.482	0.420	1.019	0.8852
Ninth decile	1.042	0.8238	1.406	0.675	1.082	0.5478

Variable	Balanced panel of enum persons (ivwd wave1)		Balanced panel of enum persons (not ivwd wave 1)		Balanced panel of responding persons	
	Odds Ratio	P-value	Odds Ratio	P-value	Odds Ratio	P-value
Tenth decile (least disadvantaged)	0.759	0.1422	0.787	0.763	0.795	0.0826
Dwelling type (base category separate house)						
Semi-detached	0.94	0.6523	0.569	0.473	1.024	0.8311
Apartment less than 3 storeys	0.755	0.0642	0.464	0.479	0.759	0.0233
Apartment 3 storeys or more	0.615	0.0143	0.597	0.648	0.789	0.1484
Dwelling unknown	0.635	0.6987	-	-	0.883	0.9167
Dwelling condition (base category excellent)						
Good	0.881	0.1865	0.49	0.176	0.954	0.4806
Average	1.081	0.4789	0.444	0.141	1.123	0.1345
Poor	1.014	0.938	0.633	0.584	1.077	0.5991
Very poor/almost derelict	0.765	0.5952	-	-	0.944	0.8901
Condition unknown	0.929	0.9538	-	-	0.224	0.0465
Number of bedrooms per person in HH	0.797	0.0078	0.242	0.065	0.785	0.001
Number of calls made to HH in w1	0.953	0.0011	0.972	0.705	0.938	<.0001
Whether HH was partly responding in W1	0.334	<.0001	3.16	0.246	0.386	<.0001
Number of adults in HH (base category two adults)						
One adult in HH in w1	0.836	0.5809	0.161	0.157	1.211	0.4963
Three or more adults in HH in w1	0.765	0.0974	2.413	0.192	0.766	0.0149
Number of children in HH (base category zero children)						
One child in HH in w1	1.025	0.9551	0.501	0.564	0.964	0.8999
Two or more children in HH in w1	1.277	0.6015	1.745	0.687	1.202	0.5486
Household type (base category couple without children)						
Couple with children under 15	0.424	0.0843	0.336	0.469	0.731	0.3454
Couple with dependent student	0.773	0.638	0.583	0.723	0.951	0.8557
Couple with non-dependent child	0.83	0.6941	0.169	0.198	0.862	0.5939
Lone parent with children under 15	0.427	0.0882	1.43	0.801	0.777	0.4578
Lone parent with dependent child	0.781	0.6952	0.128	0.253	0.925	0.8156
Lone parent with non-dependent child	0.299	0.0151	0.273	0.357	0.651	0.1509
Multifamily HH	0.475	0.0891	2.26	0.569	0.698	0.2226
Housing tenure (base category own/rent-buy)						
Rent	0.803	0.0421	0.913	0.849	0.84	0.0303
Rent free	1.279	0.3809	<0.001	0.976	1.405	0.1249
Known whether benefit recipient in HH in w1	0.776	0.012	-	-	0.782	0.0004
HH income for last financial year	1	0.3063	1	0.011	1	0.0551
Wave 2 household characteristics						
HH split in wave 2	1.488	0.0038	0.353	0.059	1.759	<.0001
Whether moved between w1 and w2	0.391	<.0001	0.414	0.113	0.452	<.0001

Appendix 3: Comparison of integration options for wave 11 cross-sectional weights

Integration methods

There are two main ways to integrate two independent surveys together: i) by combining the estimates, or ii) by pooling the samples (O'Muircheartaigh and Pedlow, 2002).

Combining the estimates involves taking a weighted average of the estimates from the two samples. Let's say we have sample A and sample B and we are interested in estimates of a total of a variable of interest *Y*. The combined estimate would be:

$$\hat{Y}_{combined} = \theta\hat{Y}_A + (1-\theta)\hat{Y}_B$$

where θ is between 0 and 1. O'Muircheartaigh and Pedlow (2002) find that when the samples are independent, the optimal choice of θ which minimises the variance of $\hat{Y}$ is:

$$\theta = \frac{\frac{n_A}{deff_A}}{\frac{n_A}{deff_A} + \frac{n_B}{deff_B}}$$

where n_A and n_B are the number of elements in each sample, and $deff_A$ and $deff_B$ are the design effects for the estimate $\hat{Y}$ in each sample.[19] Where the total is a weighted estimate of observed element in the sample, the weights for each element will therefore be:

$$w_{combined,i} = \begin{cases} \theta w_{Ai} & if\ i \epsilon S_A \\ (1-\theta) w_{Bi} & if\ i \epsilon S_B \end{cases}$$

That is, for all elements in sample A we take a fraction θ of the original weight and (1-θ) of the weight in sample B. θ can therefore be thought of as a panel allocation factor. As the design effects can be different depending on the variable of interest, we evaluate a number of choices for θ.

When pooling the samples, we need to know (or estimate) the probability of selection and response that each element had in either sample A or B. The pooled weight is given by:

$$w_{pooled,i} = \frac{1}{p_{iA} + p_{iB}}$$

where p_{iA} and p_{iB} is the probability of selection and response in sample A and B respectively. As we do not observe the probability of selection and response in the sample the element was selected into, we need to estimate this via a modelling process.

Four integration options were evaluated together with two options that keep the samples separate. Referring to the main sample as sample A and the top-up sample as sample B, the options are:

1. Set θ=1 to give the main sample only.
2. Set θ=0 to give the top-up sample only.
3. To assign θ based on the relative sample size of the two samples.
4. To assign θ based on optimising one particular estimate, chosen to be the average proportion of people aged 15 and over in the households.
5. To assign θ based on an average of θs that are optimal for a range of estimates.
6. Pooling the samples and estimate the probability of selection and response in each sample.

[19] For the integration of samples within the German Socio-Economic Panel, Spiess and Rendtel (2000) use this formula to define a range within which the optimal θ will lie for any given estimate. They then choose a particular convenient value for θ from this range.

Panel allocation and adjustment factors

A range of estimates were considered for determining the optimal value of θ, some at the household level and others at the person level, as shown in Table A3.1. These estimates are restricted to the part of the sample representing the overlapping population (being households with some or no recent arrivals that are not living in institutions or very remote parts of Australia).[20]

For option 3, the panel allocation factor is set based on the proportion of households in the main sample, being θ=0.786.

For option 4, the panel allocation factor is determined by the optimal θ for one particular estimate. This estimate was chosen to be the average number of adults in the household as this is very similar to the measure used in combining the rotating panels in the Survey of Labour and Income Dynamics (LaRoche, 2003). The panel allocation factor for option 4 is 0.793.

For option 5, we take the average of the optimal θ across a range of variables in Table A2.1, resulting in a panel allocation factor of 0.785. Reassuringly, the range of optimal θ is not particularly large, so we can be reasonably confident that the choice we make for θ within this range will not be greatly detrimental to other estimates.

Table A3.1: Optimal theta for key variables

	Main (Sample A)		Top-up (Sample B)		
SEs using PSU and strata	Deff	Sample	Deff	Sample	θ
HH level variables					
Number adults in HH	2.10	7,253	2.18	1,974	0.793
Total gross FY household income	2.22	7,253	3.07	1,974	0.835
Own dwelling	2.12	7,232	2.73	1,970	0.826
Income support reliant	2.79	7,253	3.43	1,974	0.819
Lone parent HH with dependants	1.17	7,253	1.27	1,974	0.800
Person-level variables					
Wages and salaries FY	1.84	8,647	1.76	2,268	0.785
Usual hours worked	1.53	8,623	1.31	2,262	0.766
Has permanent job	1.53	7,318	1.09	1,897	0.733
Supervisor	1.53	8,647	1.33	2,266	0.768
Job satisfaction	1.92	8,643	1.22	2,267	0.707
Married/defacto	2.50	13,404	1.91	3,655	0.737
Number of children had	1.85	13,415	2.18	3,656	0.812
Life satisifaction	2.53	13,415	2.26	3,651	0.767
Has university degree	2.71	13,413	3.65	3,654	0.831
Long term health condition	3.01	13,419	3.08	3,657	0.789
Average					0.785
Proportion of HH in sample A					0.786

For option 6, the weight of each household is adjusted for an estimated probability of selection and response in the alternative sample. This is obtained by the following process:

- Calculate the probability of selection and response for each household in the sample they were selected into. For the main sample, the probability of selection and response in the main sample is the inverse of the interim weight after the adjustment for new entrants and non-response ($p_{iA} = 1/w_{11h,interim}$). For the top-up sample, the probability of selection and response in the top-up sample is the inverse of the design weight adjusted for non-response ($p_{iB} = 1/w_{11h,adj}$).

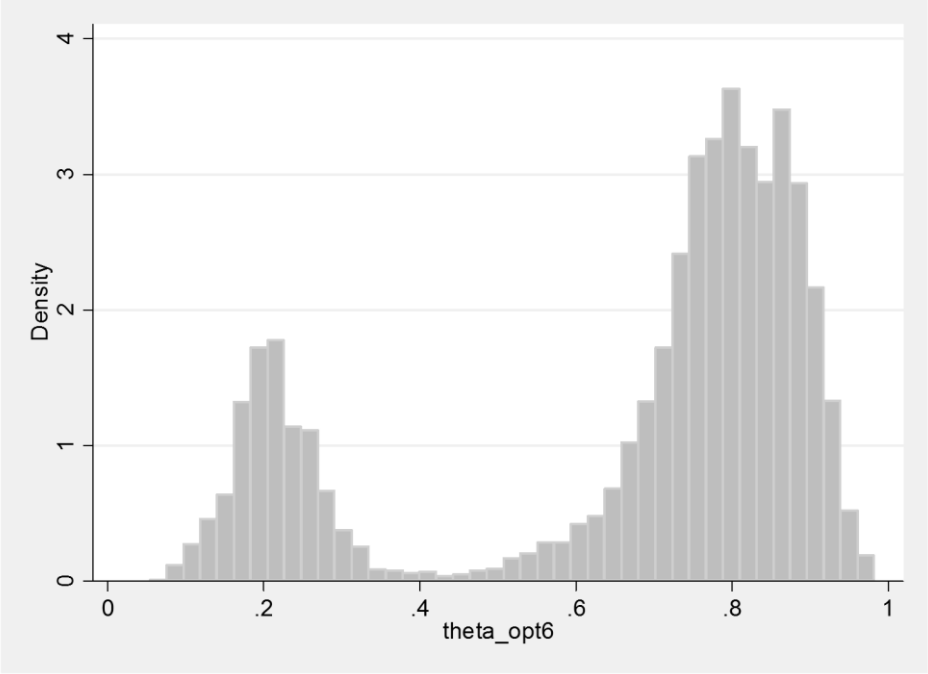
[20] This excludes households containing all recent arrivals.

- Estimate the probability of selection and response for each household in the sample they were not selected into. This is estimated via a regression model (in the same way that is used when adjusting the household weights for the new entrants that have joined the sample).[21] That is, for households in the main sample, the probability of response and selection in the top-up sample for a household with the same household characteristics and individual characteristics of the household representative person is estimated based on a model of the sample and response probabilities in the top-up sample, giving $\hat{p}_{iB}$. The same process is done for the top-up sample to predict the sampling and response probability that a household with the same characteristics would have had in the main sample, giving $\hat{p}_{iA}$. The adjusted-R^2 for the model of probabilities in the main sample is 0.210 and for the top-up sample it is 0.176.
- Calculate the pooled weights:

$$w_{pooled,i} = \begin{cases} \dfrac{1}{p_{iA} + \hat{p}_{iB}} & if\ i \in S_A \\ \dfrac{1}{\hat{p}_{iA} + p_{iB}} & if\ i \in S_B \end{cases}$$

For option 6, we can reformulate the pooled weight to identify the adjustment factor that is multiplied by the original weight in each sample. These adjustment factors are shown in Figure A3.1. The factors on the left hand side are for the top-up sample and those on the right hand side are for the main sample. The average adjustment factor for the main sample is very close to those calculated under the options to combine the samples.

Figure A3.1 Adjustment factor for weights when pooling samples (option 6)



[21] This method has also been proposed by Kaminska and Lynn (2012) for potentially integrating the British Household Panel Survey with the much larger and more recent UK Household Longitudinal Survey 'Understanding Society'.

Evaluation of integration options

A summary of the distribution of the weights under the 6 options is provided in Table A2.2. There is very little difference in the distribution of the weights between the three options to combine the samples (options 3, 4, and 5). There is sizeable reduction in the variability in the weights when we pool the samples (options 6). This finding is consistent with O'Muircheartaigh and Pedlow (2002) who show that pooling the samples produces much less variable weights when the selection probabilities are quite different between the two samples.

Table A3.2 Distribution of the weights, options 1-6 compared

Variable	N	Mean	Std Dev	Min	Max	Quartile range
Household weights						
1. Main sample	7390	1162	862	0	14368	660
2. Top-up sample	2153	3989	1600	851	15741	1870
3. Combine on sample size	9543	900	689	0	11930	475
4. Combine on one optimal theta	9543	900	692	0	11998	475
5. Combine on average optimal theta	9543	900	689	0	11920	476
6. Pool samples	9543	900	590	0	8701	392
Enumerated person weights						
1. Main sample	17953	1231	961	0	14368	708
2. Top-up sample	5451	4056	1701	851	15741	2033
3. Combine on sample size	23404	945	755	0	11930	517
4. Combine on one optimal theta	23404	945	757	0	11998	513
5. Combine on average optimal theta	23404	945	754	0	11920	515
6. Pool samples	23404	945	648	0	8701	431
Responding person weights						
1. Main sample	13603	1317	1054	0	17905	800
2. Top-up sample	4009	4467	2035	733	18438	2404
3. Combine on sample size	17612	1017	840	0	16232	582
4. Combine on one optimal theta	17612	1017	843	0	16316	581
5. Combine on average optimal theta	17612	1017	839	0	16219	581
6. Pool samples	17612	1017	721	0	10528	502

More importantly, we now consider what impact these six integration options have on various estimates. A common measure of the quality of an estimate ($\hat{Y}$) that considers both the bias in the estimate and the variability in the estimate is the root mean square error:

$$RMSE(\hat{Y}) = \sqrt{bias(\hat{Y})^2 + var(\hat{Y})}$$

The bias is taken as the difference between the relevant HILDA estimate ($\hat{Y}$) and the ABS estimate ($\hat{Y}_{ABS}$):

$$bias(\hat{Y}) = \hat{Y} - \hat{Y}_{ABS}$$

Table A3.3 provides the estimates from the six options together with the ABS estimate and the root mean square error for these six options. The option with the lowest root mean square error is indicated in bold and is the best estimate. The estimates include family type, relationship in household (for both enumerated and responding persons), highest level of education, country of birth, year of arrival, indigenous status, and for those employed we consider whether part time worker, usual hours worked, occupation, industry and employment status.

The method that provides the lowest RMSE on the majority of occasions is option 6 where we have pooled the estimates. Most of the time, this comes about via reduced variability in the estimates. Figure A3.2 shows the percentage change in the standard errors of the estimates under option 6 (pooling the samples) and option 3 (combining the estimates based on sample size). For almost all

estimates, there is a reduction in the standard error with option 6. While not shown here, there is almost no difference in the standard errors of the estimates between the three options to combine estimates (options 3, 4 and 5).

There are two variables for which the estimate for option 6 is further away from the ABS estimate than the estimate from the main HILDA sample (option 1). These variables are hours worked and highest level of education. These differences may stem from differences in the collection methodology or questions asked which will in turn limit the validity of these comparisons. The Labour Force Survey obtains information about all adults in the household from any responsible adult whereas the HILDA Survey interviews each adult in the household. Wooden, Wilkins and McGuinness (2007) shows that probably for this reason the HILDA estimates on hours worked align more closely with the ABS Survey of Employment Arrangements and Superannuation, where all adults are interviewed, than the Labour Force Survey. To some extent this collection methodology will also impact on the highest level of education information collected with qualifications not being known to the responsible adult in the household. Further the questions asked about education are quite different between the HILDA Survey and the Labour Force Survey. Respondents to the HILDA Survey are asked to recount all of their education qualifications in their first interview and this is updated over time with subsequent education activity reported in later interviews. The ABS question in the Labour Force Survey asks for the highest level of education. It is possible that the respondent filters out some less important or less relevant qualifications when answering the more aggregated question used by the ABS. There is also some suggestion of this in the HILDA Survey, with wave 1 respondents and wave 11 top-up respondents aged 15-64 showing fewer Certificate III or IV and fewer graduate diplomas or certificates than respondents aged 15-64 in other waves. Nevertheless, the differences between the estimates from the main sample and the combined sample for these two variables is generally less than 1 percentage point.

Table A3.3 Estimates and root mean squared error, options 1-6 compared

	Estimate							*RMSE*					
Characteristic	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*	*ABS*	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*
Family-level variables													
Family type (as proportion of all families, excludes lone persons and group households)													
Couple family	82.1	82.0	82.3	82.3	82.3	82.3	83.5	1.66	1.96	1.46	1.46	1.46	**1.45**
Couple family with dependent children	37.0	36.6	37.2	37.2	37.2	37.1	36.0	**1.28**	1.55	1.38	1.39	1.38	1.35
Couple family with children under 15	29.8	29.1	29.7	29.7	29.7	29.6	29.3	0.92	1.30	0.85	0.85	0.85	**0.78**
Lone parent family	16.1	15.2	15.5	15.5	15.5	15.5	14.8	1.52	1.18	0.93	0.93	0.93	**0.92**
Lone parent family with dependent children	9.6	10.3	9.7	9.7	9.7	9.8	9.9	0.62	1.02	0.51	0.51	0.51	**0.46**
Lone parent family with children under 15	7.0	8.1	7.2	7.2	7.2	7.3	7.7	0.80	0.91	0.58	0.58	0.58	**0.51**
Other families	1.8	2.8	2.3	2.2	2.3	2.2	1.6	**0.38**	1.32	0.71	0.70	0.71	0.67
Enumerated adult-level variables													
Relationship in household													
Couple with children < 15	21.6	21.1	21.5	21.5	21.5	21.4	22.0	**0.75**	1.37	0.76	0.76	0.76	0.78
Couple with dependent student (no child<15)	5.2	5.4	5.4	5.4	5.4	5.4	4.5	**0.83**	1.09	0.94	0.94	0.94	1.00
Couple with nondependent children	6.2	6.0	5.8	5.8	5.8	5.8	5.1	1.14	1.15	0.73	0.73	0.73	**0.71**
Couple without children	26.5	26.9	26.8	26.8	26.8	26.8	27.7	1.37	1.46	1.09	1.09	1.09	**1.06**
Lone parent with children<15	2.5	2.9	2.6	2.6	2.6	2.6	3.0	0.46	**0.31**	0.38	0.38	0.38	0.35
Lone parent with dependent student (no child<15)	0.9	0.8	0.9	0.9	0.9	0.9	0.8	0.15	0.17	0.11	0.11	**0.11**	0.12
Lone parent with nondependent children	2.4	1.8	2.1	2.1	2.1	2.1	1.7	0.71	**0.28**	0.43	0.43	0.43	0.39
Dependent student	7.7	7.2	7.5	7.5	7.5	7.6	7.1	0.61	**0.48**	0.48	0.49	0.48	0.51
Nondependent child	10.2	8.9	9.4	9.4	9.4	9.2	8.5	1.82	0.85	0.98	0.98	0.98	**0.80**
Other family member	3.0	3.6	3.3	3.3	3.3	3.2	2.6	**0.59**	1.08	0.78	0.78	0.78	0.71
Unrelated to all HH members	2.1	3.6	3.1	3.1	3.1	3.2	5.2	3.10	**1.69**	2.16	2.16	2.16	1.99
Lone person	11.7	11.7	11.7	11.7	11.7	11.7	11.8	0.40	0.88	0.39	0.39	0.39	**0.38**
Responding person-level variables													
Relationship in household													
Couple with children < 15	21.7	21.3	21.6	21.7	21.6	21.6	22.0	0.67	1.23	0.68	**0.67**	0.68	0.70
Couple with dependent student (no child<15)	5.5	5.5	5.6	5.6	5.6	5.7	4.5	1.13	**1.12**	1.19	1.20	1.19	1.23
Couple with nondependent children	5.7	5.8	5.4	5.4	5.4	5.4	5.1	0.64	0.92	0.40	0.40	0.40	**0.39**
Couple without children	26.3	26.7	26.5	26.5	26.5	26.6	27.7	1.58	1.58	1.32	1.32	1.32	**1.28**

	Estimate							*RMSE*					
Characteristic	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*	*ABS*	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*
Lone parent with children<15	2.7	3.1	2.7	2.7	2.7	2.8	3.0	0.34	0.34	0.28	0.28	0.28	**0.25**
Lone parent with dependent student (no child<15)	1.0	0.9	1.0	1.0	1.0	1.0	0.8	0.24	0.18	0.18	0.18	**0.18**	0.20
Lone parent with nondependent children	2.6	1.8	2.3	2.3	2.3	2.2	1.7	0.95	**0.32**	0.62	0.62	0.62	0.55
Dependent student	8.7	8.0	8.6	8.6	8.5	8.6	7.1	1.57	**0.97**	1.44	1.45	1.44	1.46
Nondependent child	9.6	8.1	8.7	8.7	8.7	8.5	8.5	1.17	0.74	0.42	0.42	0.42	**0.33**
Other family member	2.7	3.8	3.2	3.2	3.2	3.2	2.6	**0.39**	1.31	0.66	0.66	0.66	0.64
Unrelated to all HH members	1.9	3.6	2.8	2.8	2.8	3.0	5.2	3.27	**1.74**	2.39	2.40	2.39	2.25
Lone person	11.6	11.6	11.6	11.6	11.6	11.6	11.8	0.42	0.90	0.41	0.41	0.41	**0.40**
Highest level of education (15-64 year olds)													
Postgraduate (masters or doctorate)	4.1	5.8	5.4	5.4	5.4	5.5	4.6	**0.54**	1.30	0.86	0.86	0.86	0.96
Grad diploma or grad certificate	5.0	4.9	5.1	5.1	5.1	5.3	2.1	2.93	**2.81**	3.06	3.06	3.06	3.19
Bachelor or honours	14.6	16.5	15.7	15.7	15.7	15.7	17	2.50	**1.09**	1.39	1.40	1.39	1.39
Advanced diploma or diploma	8.6	9.5	8.9	8.9	8.9	8.9	9.1	0.58	0.78	0.39	0.39	0.39	**0.38**
Cert IV or III	21.3	22.6	20.7	20.7	20.7	20.6	17.4	3.97	5.33	3.37	3.35	3.37	**3.20**
Year 12	19.0	16.3	18.0	18.0	18.0	17.9	20.6	**1.68**	4.33	2.62	2.61	2.62	2.73
Year 11 or below (inc Cert I, II, nfd)	27.2	24.3	26.0	26.0	26.0	26.1	29.1	**2.00**	4.94	3.17	3.16	3.18	3.08
Undetermined	0.1	0.1	0.2	0.2	0.2	0.1	-						
Country of birth													
Australia	75.5	68.3	69.5	69.5	69.5	70.0	70.1	5.41	2.38	1.22	1.21	1.22	**1.07**
Main English speaking country	8.8	12.2	10.8	10.8	10.8	10.8	10.8	1.98	1.67	0.56	**0.56**	0.56	0.57
Other country	15.7	19.5	19.7	19.7	19.7	19.2	19.1	3.51	1.71	1.22	1.22	1.22	**1.07**
Year of arrival (if born overseas)													
Before 1971	27.9	23.4	22.4	22.4	22.4	22.3	24.0	4.13	2.15	2.05	2.06	**2.05**	2.11
1971-1980	14.2	13.2	11.3	11.3	11.3	11.2	11.6	2.75	2.06	0.88	0.89	**0.88**	0.90
1981-1990	25.7	12.0	17.6	17.6	17.6	17.0	16.8	9.10	4.87	1.45	1.48	1.44	**1.12**
1991-2000	24.6	15.9	17.3	17.4	17.3	16.5	16.4	8.47	1.65	1.60	1.61	1.60	**1.15**
2001-2005	4.0	11.6	9.2	9.2	9.2	9.7	9.8	5.83	2.18	1.19	1.20	1.19	**1.11**
2005-2010	3.5	20.9	19.3	19.3	19.3	20.4	18.7	15.23	2.83	1.98	**1.98**	1.98	2.57
2011	0.1	3.0	2.8	2.8	2.8	3.0	2.7	2.53	0.69	0.63	0.63	**0.63**	0.71
Indigenous	2.5	2.6	2.2	2.2	2.2	2.2	2.1	0.49	0.71	0.26	0.26	0.26	**0.24**
Employed persons													

	Estimate							*RMSE*					
Characteristic	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*	*ABS*	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*
Part time worker	32.0	33.9	32.8	32.8	32.8	32.8	30.6	**1.54**	3.49	2.31	2.30	2.31	2.38
Usual hours worked													
0	0.1	-	0.0	0.0	0.0	0.0	0.2	**0.18**	-	0.20	0.20	0.20	0.19
1-15	12.4	13.3	12.6	12.6	12.6	12.7	11.6	**0.87**	1.83	1.05	1.04	1.05	1.13
16-29	13.5	14.4	14.1	14.1	14.1	14.0	13.0	**0.75**	1.80	1.31	1.31	1.31	1.21
30-34	6.0	6.2	6.0	6.0	6.0	6.1	5.7	0.39	0.69	0.37	**0.37**	0.37	0.44
35-39	19.2	18.3	19.4	19.4	19.4	19.6	23.4	4.21	5.23	4.02	4.01	4.02	**3.83**
40	16.9	15.0	16.4	16.4	16.4	16.3	19.7	**2.88**	4.81	3.35	3.34	3.35	3.42
41-44	4.2	4.4	4.3	4.3	4.3	4.3	3.2	**1.05**	1.27	1.14	1.14	1.14	1.08
45-49	9.2	9.3	9.2	9.1	9.2	9.1	7.1	2.18	2.37	2.13	2.12	2.13	**2.06**
50-59	11.6	11.6	11.1	11.1	11.1	10.9	9.3	2.31	2.36	1.83	1.82	1.83	**1.69**
60 or more	6.9	7.6	6.8	6.8	6.8	6.9	6.7	0.40	1.18	0.37	**0.36**	0.37	0.39
Occupation													
Managers	13.0	13.7	13.1	13.1	13.1	13.2	13.0	0.53	1.09	0.51	**0.51**	0.51	0.52
Professionals	23.3	22.2	23.3	23.3	23.3	23.5	21.6	1.94	**1.58**	1.92	1.93	1.92	2.13
Technicians and trade workers	14.7	13.6	13.9	13.9	13.9	13.7	14.2	0.66	1.08	0.53	**0.53**	0.53	0.68
Community and personal service workers	9.5	11.1	9.9	9.8	9.9	10.0	9.7	0.47	1.50	0.38	**0.37**	0.38	0.43
Clerical and administrative workers	15.4	14.4	14.8	14.8	14.8	14.8	15.1	0.57	1.02	0.48	**0.48**	0.48	0.50
Sales workers	9.1	9.6	9.3	9.3	9.3	9.3	9.4	0.49	0.72	0.36	0.37	0.36	**0.35**
Machinery operators and drivers	6.0	5.8	5.9	5.9	5.9	5.8	6.8	**0.91**	1.16	0.97	0.97	0.97	1.04
Labourers	9.1	9.6	9.7	9.7	9.7	9.7	10.2	1.29	1.03	0.83	0.83	0.83	**0.83**
Industry													
Agriculture, forestry and fishing	2.4	2.0	2.5	2.5	2.5	2.6	2.8	0.53	0.97	0.45	0.45	0.45	**0.39**
Mining	1.9	1.9	1.9	1.9	1.9	1.9	2.0	0.26	0.44	0.24	0.24	0.24	**0.23**
Manufacturing	8.4	8.6	8.4	8.3	8.4	8.2	8.3	0.48	0.77	0.40	0.40	0.40	**0.38**
Electricity, gas, water and waste services	1.1	0.9	1.0	1.0	1.0	1.0	1.2	**0.20**	0.47	0.27	0.27	0.27	0.28
Construction	8.4	9.8	8.4	8.4	8.4	8.4	9.1	0.85	1.05	0.82	0.83	0.82	**0.78**
Wholesale trade	3.1	4.1	3.3	3.3	3.3	3.4	3.6	0.59	0.66	0.33	0.33	0.33	**0.29**
Retail trade	10.7	10.6	10.7	10.7	10.7	10.6	10.8	0.49	0.77	0.44	0.44	**0.44**	0.45
Accommodation and food services	6.0	6.0	6.1	6.1	6.1	6.1	6.9	0.92	1.08	0.85	0.85	0.85	**0.82**
Transport, postal and warehousing	5.0	4.7	4.8	4.8	4.8	4.7	5.1	**0.39**	0.68	0.43	0.43	0.44	0.49

	Estimate							*RMSE*					
Characteristic	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*	*ABS*	*Opt1*	*Opt2*	*Opt3*	*Opt4*	*Opt5*	*Opt6*
Information media and telecommunications	1.8	2.2	2.0	2.0	2.0	2.0	1.8	**0.21**	0.53	0.30	0.29	0.30	0.27
Financial and insurance services	4.1	3.6	4.0	4.0	4.0	4.0	3.8	0.45	0.42	0.38	0.39	0.38	**0.35**
Rental, hiring and real estate services	1.4	0.9	1.3	1.3	1.3	1.3	1.7	**0.34**	0.81	0.44	0.44	0.44	0.48
Professional, scientific and technical services	8.2	9.2	8.5	8.5	8.5	8.5	7.7	**0.59**	1.63	0.83	0.83	0.83	0.83
Administrative and support services	3.4	2.8	3.5	3.5	3.5	3.4	3.6	0.44	0.86	0.40	0.40	0.40	**0.38**
Public administration and safety	6.7	5.0	6.3	6.3	6.3	6.3	6.5	0.42	1.56	0.37	0.36	0.37	**0.35**
Education and training	9.6	8.8	9.2	9.2	9.2	9.3	7.6	1.99	**1.29**	1.60	1.60	1.60	1.70
Health care and social assistance	12.1	13.2	12.4	12.4	12.4	12.5	11.7	**0.61**	1.71	0.87	0.86	0.87	0.94
Arts and recreational services	2.0	1.3	1.8	1.8	1.8	1.8	1.8	0.24	0.57	0.16	0.16	0.16	**0.15**
Other services	3.9	4.5	4.0	4.0	4.0	4.0	4.0	0.28	0.77	0.27	0.27	0.27	**0.26**
Employment status													
Employee	90.5	89.8	90.6	90.6	90.6	90.4	89.2	1.36	**1.03**	1.43	1.43	1.43	1.28
Employer	2.2	2.1	2.1	2.1	2.1	2.1	2.9	**0.77**	0.88	0.83	0.83	0.83	0.81
Own account worker	7.1	7.6	7.0	7.0	7.0	7.2	7.8	0.86	**0.72**	0.88	0.89	0.88	0.77
Contributing family member	0.3	0.5	0.3	0.3	0.3	0.3	0.0	**0.23**	0.46	0.26	0.26	0.26	0.27

Note: ABS estimates for relationship in household, country of birth, year of arrival and indigenous status exclude institutionalised population, otherwise the estimates apply to all civilians aged 15 and over. HILDA estimates also for aged 15 and over including the defence force but excluding institutionalised population and very remote parts of Australia.

ABS sources: i) Family type is from ABS Cat.No. 6224.0.55.001 *Labour Force, Australia: Labour Force Status and Other Characteristic of Families*, June 2011. ii) Relationship in household, country of birth, year of arrival and usual hours worked is from ABS Cat.No. 6291.0.55.001 *Labour Force, Australia, Detailed - Electronic Delivery*, September 2011. iii) Highest level of education is from ABS Cat.No. 62270DO001_201105 *Education and Work, Australia*, May 2011. Indigenous status is from ABS Cat.No. 62870DO001_2011 *Labour Force Characteristics of Aboriginal and Torres Straight Islander Australians*, 2011. iv) Occupation, industry and employment status is from ABS Cat.No. 6291.0.55.003 *Labour Force, Australia, Detailed, Quarterly*, August 2001.

Figure A3.2: Percentage change in standard error for pooled sample (option 6) compared to combined sample based on relative sample size (option 3)