



DEPARTMENT OF
**FAMILY AND
COMMUNITY
SERVICES**



THE UNIVERSITY OF
MELBOURNE

HILDA PROJECT TECHNICAL PAPER SERIES NO. 3/02, September 2002

The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Weighting

Nicole Watson and Tim R.L. Fry

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



Contents

CONTENTS.....	0
INTRODUCTION	1
HOUSEHOLD WEIGHTS	4
DESIGN WEIGHTS.....	4
ADJUSTMENTS TO DATA INTERNAL TO HILDA SURVEY	4
Internal data available	4
Model describing response	5
ADJUSTMENTS TO DATA EXTERNAL TO HILDA SURVEY	6
PERSON WEIGHTS	8
ENUMERATED AND RESPONDING PERSONS	8
WEIGHTS AT THE START OF THE WEIGHTING PROCESS FOR PERSONS.....	8
ADJUSTMENTS TO DATA INTERNAL TO HILDA SURVEY – RESPONDING PERSON WEIGHT ONLY	9
Internal data available	9
Model describing response	9
ADJUSTMENTS TO DATA EXTERNAL TO HILDA SURVEY.....	10
COMPARISON TO ACCEPTED POPULATION ESTIMATES	12
WEIGHTS PROVIDED IN THE WAVE 1 DATASETS.....	18
HOUSEHOLD FILE.....	18
PERSON FILE	19
REFERENCES	20
APPENDIX 1 – TECHNICAL REFERENCE GROUP MEMBERSHIP.....	21
APPENDIX 2 – PROBABILITY OF SELECTION FOR HOUSEHOLDS AND PERSONS	22
APPENDIX 3 – MODELS FOR PREDICTING RESPONSE TO THE HILDA SURVEY.....	24
APPENDIX 4 – EFFECT OF ADJUSTMENTS ON WEIGHTS	28

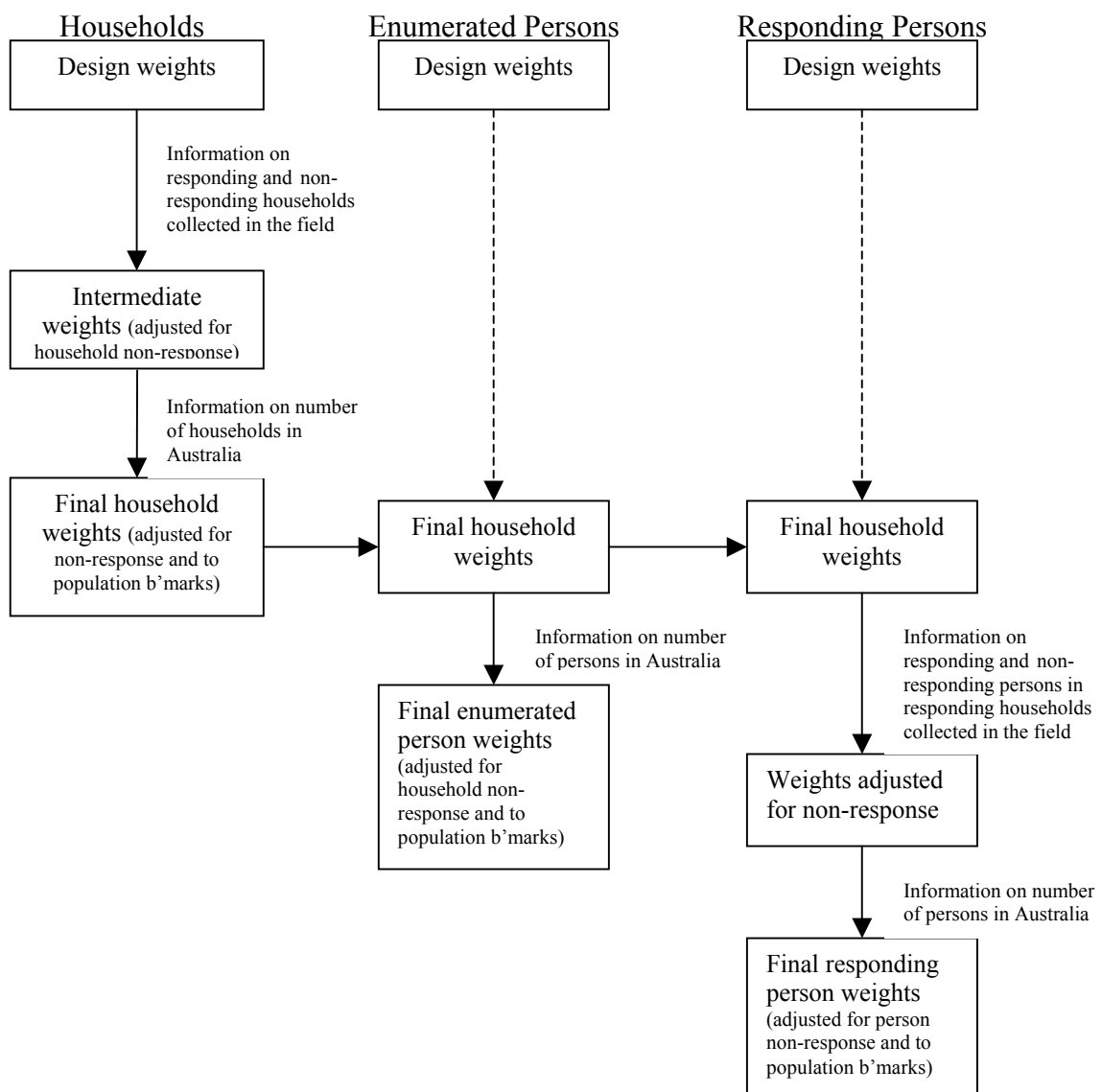
Introduction

This paper details the methodology used to calculate the weights for the Wave 1 sample of the Household, Income and Labour Dynamics in Australia (HILDA) Survey. The initial (or design) weights are derived from the probability of selecting the households into the sample. These household weights are initially adjusted according to information collected about all selected households (both responding and non-responding) and further adjusted so that weighted household estimates from the HILDA Survey match several known household-level benchmarks.

The person-level weights are based on the household-level weights, with adjustments made based on information collected about all the people listed in the responding households. These weights are also adjusted to ensure that the weighted person estimates match several known person-level benchmarks.

Diagram 1 shows the process involved in constructing the household and person weights. Each of these steps are discussed in more detail later in this paper.

Diagram 1: Weighting process



The weighting methodology was developed with input from a Technical Reference Group that included representatives from the Australian Bureau of Statistics, Department of Family and Community Services, Department of Employment and Workplace Relations, and Data Analysis Australia (see Appendix 1). The authors are appreciative of the suggestions and comments received from this group, together with other individuals from the Australian Bureau of Statistics who were consulted from time to time.

The Technical Reference Group agreed to the following broad principles that guided the weighting strategy used for the first wave of the HILDA Survey:

- The weights should be considered as expansion factors permitting the scaling of the sample to the population. Hence the sum of the weights should accurately match known population parameters such as the total population and the total number of households.
- The weights should adjust for unequal probabilities of inclusion in the survey, to redress any potential sampling biases. In many cases this will mean that weights are inversely proportional to the probability of inclusion.
- Where the survey is complex, resulting in data at several levels, weights for different levels of the survey should be consistent with each other.
- The design of weighting should aim to give a dataset with broad application, while at the same time accepting that some variables will need to be treated as more important than others.
- Care should be taken to match adjustment models and external constraints imposed on survey estimates to supportable reduction in estimate bias. Insupportable reduction is indicated by excessively low or high weights, or marked clustering of weights about imposed bounds.

These principles were largely based on the *HILDA Discussion Paper 3/01* (Henstridge 2001).

Before describing the weighting process in detail, it is worth clarifying at this point the scope and coverage of the HILDA Survey. The scope of the survey is the population that the survey purports to represent, and the coverage of the survey is the population from which the sample was selected. Essentially for the HILDA Survey, the following apply:

- Scope – Households (and usual residents) in private dwellings, excluding remote parts of the Northern Territory.
- Coverage – Households (and usual residents) in private dwellings, excluding remote parts of Australia.

For the most part, these scope and coverage rules are similar to those adopted by the Monthly Population Survey supplements conducted by the ABS.¹ The *HILDA Technical Paper 1/02* (Watson and Wooden 2002) provides detailed definitions of usual residents and households.

As will be discussed later in the paper, the household and person weights will be adjusted so that the sum of the weights match the population in scope. As a result, the total number of households and persons estimated from the HILDA Survey using the final weights will match the Australian population in private dwellings, excluding remote parts of the Northern Territory.

¹ The HILDA Survey includes children at boarding schools or halls of residences into their non-term time household along with military personnel living in non-private dwellings, whereas the ABS supplementary surveys usually exclude these people.

Household Weights

Design Weights

The probability of selecting each household into the sample was determined by the sample design employed for the HILDA Survey. Users not familiar with the sample design are directed to the *HILDA Technical Paper 1/02* (Watson and Wooden 2002).

Households were selected into the HILDA Survey using a multi-stage area-based design. The probability of selecting a particular household was dependent on the following:

- the probability of selecting the Census Collection District (CD);
- the probability of selecting the block given the CD was selected;
- the probability of selecting the dwelling given the block was selected; and
- the probability of selecting the household given the dwelling was selected.

The derivations for these probabilities of selection are provided in Appendix 2. Note that the block stage of selection was only used for five extremely large rural CDs where it was not practical to list the entire CD.

The design weight for a household was calculated as the inverse of the probability of selecting that household.

The main source of variability in the design weights stems from the difference between the expected number of dwellings in the CD (based on the 1996 Census count) and the actual number of dwellings identified in the CD (based on the full listing by HILDA interviewers in 2001). There were 12 CDs where the current dwelling count was more than 20 per cent lower than the Census count, and 52 CDs where the current dwelling count was more than 40 per cent higher than the Census count. Such changes in the dwelling counts were not unexpected given the five-year gap between the 1996 Census and the 2001 listing.

Adjustments to Data Internal to HILDA Survey

Internal data available

The data internal to the HILDA Survey that can be used to adjust the design weight comes from two sources. First, the HILDA interviewer collected some basic information about all households selected – both responding and non-responding. Second, we have information from the 1996 Census about the selected CDs that describes the neighbourhood to which the dwellings belong.

The information collected by the HILDA interviewers for all selected households includes:

- dwelling type;
- external condition of the dwelling;
- security features of the dwelling; and
- proportion of high-rise buildings in the area.

The information known about the CD selected includes:

- geographic location;
- proportion of different family and household types;
- median age of persons in the area;
- density of area (population per square kilometer);
- proportion of people speaking a language other than English;
- proportion of different dwelling types;
- proportion of people with different employment statuses;
- average household size; and
- median weekly household income.

This information was used to create a model for predicting response that was then used to adjust the design weights.

Model describing response

The model of household response was constructed using logistic regression. The final model is detailed in Appendix 3 (Table A3.1 and A3.2).

Households that had locked gates, dangerous dogs, security doors and ‘no junk mail’ signs were less likely to respond. Locked gates and dangerous dogs would have limited the interviewer’s access to the dwelling. Security doors would have limited communication between the interviewer and the householder, making it difficult for the interviewer to convince the household to take part in the survey. We suspect that people with ‘no junk mail’ signs are generally less receptive to requests on their time, such as taking part in a survey. The model also suggests that households protected by a security guard, doorman, on-site manager or gate keeper were less likely to respond, though we are unable to accurately estimate the effect of such guards due to the small number of such households in the HILDA sample.

People living in dwellings in poor condition were less likely to respond, suggesting that the small financial incentive to participate in the HILDA survey did not completely offset the lower response expected from poorer households.²

Response differed by State, and within State by metropolitan and rural areas. This is not uncommon in voluntary surveys conducted in Australia (Bednall, Cavenett and Shaw 2000). Sydney, in particular, had lower response rates than expected and this may be attributed to people in Sydney being more time pressured than people in other places and, therefore, less willing to take part in what were relatively lengthy interviews.

A number of variables describing the neighbourhood did not significantly add to the predicting power of the model and were dropped. Among these were variables describing the types of families living in the area, the median age of the population in the area, the average household size, and the type of dwellings in the area.

² Fully responding households (where all eligible people provided an interview) received \$50, and partially responding households (where at least one eligible person provided an interview) received \$20.

There were a few neighbourhood variables that did contribute to the model. The greater the density of the area, the less likely the households were to respond – that is households in the inner city areas were less likely to respond than those in the outer city areas. Response was lower in areas that had a high proportion of people speaking a language other than English (though this effect is not significant at the 5% level) and this is probably attributable to both a greater chance of a language barrier and a higher degree of suspicion of surveys in such areas. Response was higher in areas with a greater proportion of unemployed people or people not in the labour force. It is expected that this is a result of being able to contact these people more readily and of them being more receptive to the offered incentive. Gaining the participation of people living in more affluent areas was harder than other areas and the incentive would have had minimal impact.

The model was used to calculate the probability that each household in the selected sample had of responding, based on their dwelling and neighbourhood characteristics. There were 23 households that were missing one or more of the variables used in the model to predict response. Solely for the purposes of assigning the probability of response, these variables were imputed using information from another household that they were most like.

The design weights were then adjusted by a factor that was the inverse to the probability that the household was likely to respond. That is, the less likely a household was to respond, the higher the weight it was given. This may be thought of as increasing the contribution of households that have characteristics similar to other households that did not respond. Conversely, the more likely a household was to respond, the lower the weight boost it was given.

Adjustments to Data External to HILDA Survey

The household weights, adjusted for the probability of response, were further calibrated to population benchmarks. This means that estimates produced from the HILDA Survey will match the selected population benchmarks. Two sets of benchmarks were used:

- Household benchmark 1:- Number of households by State and part of State. For NSW, Vic, Qld, SA and WA, the part of State variable separated the metropolitan area from the rest of the State. For Tas, NT, and ACT, part of State was not used.
- Household benchmark 2:- Household type (based on number of adults and children) by broad geographic areas. There were nine household types combining one, two, and three or more adults (aged 15 and over) with zero, one and two or more children (aged under 15). The broad geographical areas included Sydney, Melbourne, Brisbane, ACT combined with rural NSW, WA combined with SA, Tas combined with rural Vic, NT combined with rural Qld.

The benchmarks were obtained from the Australian Bureau of Statistics (as a special data service) and relate to the estimated number of households in Australia as at 30

September 2001. The benchmarks excluded households in remote areas of NT and included only those households in private dwellings.

The household weights were simultaneously calibrated to both sets of benchmarks using GREGWT (a SAS macro developed in the Statistical Services area of the ABS).³ Appendix 4 provides some information on how the weights were changed through the various adjustments made.

³ The GREGWT macro performs generalized regression weighting as described by Stukel, Hidioglou and Sarndal (1996).

Person Weights

Enumerated and Responding Persons

Within a household, not everyone who was eligible for interview actually provided an interview (though some information about these individuals was collected during the household interview). Of the 7682 households participating in Wave 1, 10.5 per cent had at least one eligible person who did not complete an interview.⁴

As a result, two weights have been provided at the person level. Each person who is a usual member of a responding household is assigned an enumerated person weight (this includes respondents, non-respondents, and children). Each person providing a personal interview is assigned a responding person weight. Information about the non-respondents in responding households is used, together with benchmark information, to construct the weights for the responding persons. Essentially, this spreads the weight of the non-responding people across the responding people who are similar to them.

An alternative to these two person weights would be to impute all of the person-level information for the non-responding people, and simply use the enumerated person weight. The likely method of imputation would be hot deck, which takes the record of a similar person and inserts this into the non-responding person's record. The responding person's data would appear twice – once for themselves and once for the non-respondent. This could also be viewed as adding the non-responding person's weight to the responding person's weight as the respondent now makes a greater contribution to the estimates. Imputing non-respondents would complete the information for a partially responding household, though not all of this information would have originated from this household.

The rest of this section describes how the person weights were constructed.

Weights at the Start of the Weighting Process for Persons

Rather than using the household design weight as the starting person weight in the weighting process, the final household weight was used to encourage consistency between the person-level weights and the household weight.⁵

An alternative to using the household weight as the basis of the person weight and then making adjustments at the person level would have been to undertake the weighting of the household simultaneously with the weighting for the enumerated persons. This approach was investigated, but it greatly limited the detail that could be

⁴ At the person level, this translates to 7.7 percent of all eligible people in the 7682 households that did not complete a person interview. (That is, 1158 of the 15,127 eligible people in the 7682 households did not complete a person interview.)

⁵ Note that the design weight for the person is the same as the design weight for the household since the probability of selection of a person coincides with that of the household in which they reside.

applied in the benchmarking process and resulted in large changes to the weights of some households, particularly those with many household members.

Adjustments to Data Internal to HILDA Survey – Responding Person Weight Only

As noted earlier, information about the non-respondents in responding households was used to adjust the weight of the responding persons. A model of response to the personal interview was constructed and used to make the first adjustment to the responding person weight. Therefore, respondents who were like non-respondents had their weight increased.

This step in the process for constructing the responding person weight was not relevant to the enumerated person weight.

Internal data available

The data internal to the HILDA Survey that could be used to adjust the person weight comes from the household interview. The following information was considered in the construction of the model for predicting response at the person level:

- geographic location;
- sex;
- age group;
- marital status;
- labour force status;
- english ability;
- presence of long term health condition;
- number of adults and number of children in the household;
- housing tenure; and
- dwelling structure.

Model describing response

The model of person-level response was constructed using logistic regression. This model was restricted to people aged 15 or over in responding households with two or more eligible persons.⁶ The final model is detailed in Appendix 3 (Table A3.3 and A3.4).

This analysis indicates that the people who did not provide an interview were more likely to be:

- living in Sydney;
- employed full time;
- male;
- aged 20 to 44;

⁶ In responding households where there was only one person aged 15 or over, they did, by definition, respond.

- living in households with three or more adults;
- without children in the household;
- not married or in defacto relationship; and
- not able to speak English well and prefer to speak a language other English at home.

The model was used to calculate the probability that each person had of responding, based on their characteristics. There were 12 people that were missing one or more of the variables used in the model to predict response. Solely for the purposes of assigning the probability of response, these variables were imputed using the information from another person that they were most like.

The person weights were then adjusted by a factor that was the inverse of the probability that the person was likely to respond, given that their household responded. That is, the less likely a person was to respond, the higher the weight they were given. This may be thought of as increasing the contribution of people who have characteristics similar to the non-respondents. The weights for the people in households where there was only one eligible person were left unchanged.

To avoid the problem of significantly increasing the weights when the model is based on a relatively small number of cases at the tail of the distribution, the amount that the weights were adjusted by was limited to 1.8 times the person weight being adjusted. This affected 36 cases out of the 13,969 responding people.

Adjustments to Data External to HILDA Survey

The final step in calculating the enumerated person weights and the responding person weights was to calibrate them to known population benchmarks. This means that estimates produced from the HILDA Survey will match the selected population benchmarks. Two sets of benchmarks were used:

- Person benchmark 1:- Number of people by State, part of State, sex and age. For NSW, Vic, Qld, SA and WA, the part of State variable separated the metropolitan area from the rest of the State. For Tas, NT, and ACT, part of State was not used. The age categories used were:
 - 0-4, 5-9, 10-14, 15-19, 20-24, 25-34, 35-44, 45-54, 55-65, 65-75, 75+ in NSW, Vic, Qld, Adelaide and Perth;
 - 0-4, 5-9, 10-14, 15-24, 25-34, 35-44, 45-54, 55-65, 65+ in rural SA, rural WA and Tas;
 - 0-14, 15-34, 35+ in NT; and
 - 0-9, 10-14, 15-24, 25-34, 35-44, 45-54, 55+ in ACT;
- Person benchmark 2:- Number of people by labour force status and State. The labour force status included the following categories: under 15, employed, unemployed and not in the labour force. For NT and ACT, the unemployed and not in the labour force categories were collapsed.

The benchmarks were obtained from the Australian Bureau of Statistics (as a special data service). The first set of person benchmarks relate to the estimated number of

residents in Australia as at 30 September 2001. The second set of person benchmarks was obtained from the Labour Force Survey, with an average calculated across four months from August to December 2001. The second set of person benchmarks was proportionally adjusted so that the total number of people in each State matched the estimated residential population in the first set of person benchmarks. Both sets of benchmarks excluded people living in remote areas of NT and those living in non-private dwellings.

The enumerated person weights were simultaneously calibrated to both sets of benchmarks using GREGWT. The responding person weights (after being adjusted for the probability of response) underwent the same calibration procedure. The final person weights were constrained to be within a third and three times the design weight for each person. This ensured that the weights did not stray excessively from the underlying design weights. This limit only affected 11 of the 13,969 responding person weights and did not affect the enumerated person weights at all.

Appendix 4 provides some information on how the weights changed through the various adjustments made.

Comparison to Accepted Population Estimates

To consider how well the HILDA sample represents the population, the estimates produced from the HILDA sample can be compared to official population estimates. The following three tables provide a series of estimates from the HILDA Survey – the first table is at the household level and the second two tables are at the person level. The second table is for all people aged 15 and over listed in the responding households and the third table is for all people aged 15 and over who provided a personal interview.

These estimates make use of a number of weights. These are:

- equal weights (equivalent to unweighted);
- design weights inflated to account for global household non-response;
- weights following the non-response adjustment based on data internal to the survey; and
- weights following the adjustment to external benchmarks.

Table 1 indicates that the unweighted sample under-represents people living in Sydney and over-represents people living in the rural parts of New South Wales, Victoria, and South Australia. This problem was largely corrected when adjustments to the weights were made based on a household's probability of responding.

The average number of people per household in the HILDA Survey is consistent with the ABS.

The broad housing tenure categories of owner, renter and other are consistent with those from the ABS. Homeowners without a mortgage appear to be under-represented and homeowners with a mortgage appear to be over-represented. It is believed that this difference in the types of homeowners is a result of the questions asked and how the instruments were delivered. The 2001 Census (from which we take the ABS figures) was conducted using a self-complete form. The accompanying booklet instructed respondents to count loans secured against their home as a mortgage. We believe that this booklet would not have been consulted consistently across all households. The questions about housing tenure in the HILDA Survey are interviewer delivered and households were specifically asked about mortgages to pay for their home and about loans secured against their home. The apparent differences for other renters can be explained by a small percentage (0.3%) of renters in the Census who did not provide sufficient information to classify their type of landlord.

Dwelling structure is also consistent with the ABS figures, with the exception of other dwelling types. It is expected that the HILDA sample has a smaller proportion of other dwellings due to the exclusion of remote parts of Australia (where such dwellings are more likely to be found). The Northern Territory contributes more than half of the number of 'other' dwellings to the Australian level figure in the Census. Approximately twenty percent of the dwellings in the Northern Territory are in the remote areas that were excluded from the HILDA sample.

Table 1: Selected Wave 1 Household Characteristics and Population Counts from the ABS for 2001 Compared (private dwellings)

	ABS ^a	HILDA ^b			
		Unweighted	Design weights	Weights adj for non-resp	Weights adj for non-resp and ERH b'marks
Area of usual residence					
Sydney	0.204	0.175**	0.169**	0.212	0.204
Rest New South Wales	0.126	0.137**	0.146**	0.132	0.126
Melbourne	0.178	0.174	0.167*	0.174	0.178
Rest Victoria	0.068	0.077**	0.075	0.067	0.068
Brisbane	0.087	0.084	0.088	0.088	0.087
Rest Queensland	0.104	0.109	0.118	0.106	0.104
Adelaide	0.061	0.063	0.061	0.060	0.061
Rest South Australia	0.022	0.030*	0.023	0.018**	0.022
Perth	0.075	0.074	0.077	0.074	0.075
Rest Western Australia	0.026	0.027	0.028	0.024	0.026
Tasmania	0.026	0.03*	0.028	0.026	0.026
Northern Territory	0.007	0.005*	0.005	0.007	0.007
Australian Capital Territory	0.017	0.015	0.015*	0.013**	0.017
Housing tenure^c					
Owner	0.695	0.684	0.683	0.676	0.680
With mortgage	0.278	0.331**	0.335**	0.332**	0.327**
Without mortgage	0.417	0.351	0.346	0.341	0.351
Renter	0.276	0.288	0.290	0.298*	0.294
State/territory housing authority	0.047	0.049	0.048	0.048	0.048
Other	0.225	0.239	0.242*	0.250**	0.246*
Other	0.029	0.029	0.028	0.027	0.026
Mean number of usual residents in the household	2.568	2.593	2.602	2.602	2.581
Dwelling structure^d					
Separate house	0.753	0.773	0.773	0.755	0.758
Semi-detached	0.089	0.100	0.100	0.099	0.098
Flat	0.131	0.117	0.117	0.137	0.135
Other	0.019	0.010**	0.009**	0.008**	0.008**

Notes: ** and * denotes significantly different from the ABS population estimate at the 99% and 95% confidence levels respectively.

a The ABS figures for the usual area of residence exclude remote parts of NT.

b All HILDA estimates exclude dwellings in remote areas of NT.

c The 4.7% of dwellings in the Census for which housing tenure was not stated were assumed to be spread across tenure types in the same proportion as known tenure types.

d Excludes a small proportion of cases where the dwelling structure type was not able to be classified.

Source: ABS data are from 2001 Census (Cat. No. 2001.0), with the exception of the usual area of residence figures which are unpublished Estimated Residential Household data for September 2001 provided by the Demography Section.

The person-level estimates show a similar picture to the household-level estimates by State and metropolitan and rural areas, as detailed in Table 2.⁷ Correcting by the household's probability of response and adjusting to household-level benchmarks still leaves those aged 15-19 over-represented in the enumerated sample. Adjusting by the age and sex distribution of those in private dwellings does correct for this.

The HILDA sample over-represent married people. The size of this discrepancy is reduced somewhat through the weighting process, but a 3.3 percentage point gap still exists. Part of the gap may reflect differences in the way de facto partnerships are defined and treated in the HILDA Survey compared with the ABS Monthly Population Survey.

Table 3 shows that the responding sample displays similar characteristics to the enumerated sample, though the differences are slightly more marked. We are able to compare a greater range of estimates from the responding sample to ABS estimates as we have much more information about the people who provided a personal interview.

The observed differences in country of birth and employment status between the unweighted HILDA sample and the ABS estimates are corrected through the weighting process.

Own account workers are under-sampled in the HILDA responding sample. It is expected that this is a result of such people being more time pressured than other people and therefore less able to participate in an interview. They may also have been harder to contact than other people due to their irregular work schedule.

Overall, the HILDA sample largely reflects what we know about the population from other sources. The main concern identified through this comparison process is the apparent over-sampling of married people. Differences by geographic location, age and sex are expected and can be corrected for through the weighting process.

⁷ The figures presented in Table 2 have been restricted to people aged 15 years and over to aid comparisons between enumerated persons (shown in Table 2) and responding persons (shown in Table 3).

Table 2: Selected Wave 1 Enumerated Person Characteristics and Population Estimates from the ABS for 2001 Compared (persons aged 15 years or over)

	ABS ^a	HILDA ^b			Weights adj to ERP and LFS b'marks
		Unweighted	Design weights	Final HH weights	
Area of usual residence					
Sydney	0.217	0.185**	0.180**	0.217	0.217
Rest New South Wales	0.121	0.132**	0.141**	0.121	0.121
Melbourne	0.186	0.182	0.175	0.184	0.186
Rest Victoria	0.066	0.075**	0.073	0.065	0.066
Brisbane	0.086	0.083	0.088	0.086	0.085
Rest Queensland	0.100	0.105	0.114	0.100	0.100
Adelaide	0.058	0.060	0.059	0.059	0.058
Rest South Australia	0.020	0.030**	0.022	0.021	0.020
Perth	0.073	0.071	0.074	0.072	0.074
Rest Western Australia	0.025	0.026	0.027	0.024	0.025
Tasmania	0.024	0.029*	0.027	0.024	0.024
Northern Territory	0.007	0.005	0.005	0.008	0.007
Australian Capital Territory	0.016	0.016	0.016	0.017	0.016
Sex					
Males	0.495	0.486**	0.485**	0.487**	0.495
Females	0.505	0.514**	0.515**	0.513**	0.505
Age group					
15-19	0.086	0.088	0.090	0.097**	0.084
20-24	0.090	0.078**	0.078**	0.085	0.089
25-34	0.191	0.186	0.187	0.184	0.191
35-44	0.194	0.213**	0.215**	0.200	0.195
45-54	0.174	0.174	0.171	0.176	0.175
55-64	0.120	0.119	0.117	0.119	0.120
65+	0.146	0.142	0.140	0.139	0.146
Marital status					
Married	0.587	0.627**	0.627**	0.615**	0.620**
Not married	0.413	0.373**	0.373**	0.385**	0.380**

Notes: ** and * denotes significantly different from the ABS population estimate at the 99% and 95% confidence levels respectively.

a The ABS estimates for area of usual residence, sex, and age apply to people in private dwellings, excluding those in remote parts of NT. The ABS estimates for marital status apply to civilians.

b The HILDA estimates apply to people in private dwellings, excluding those in remote parts of NT.

Source: ABS data are unpublished Estimated Residential Population data for September 2001 provided by the Demography Section, with the exception of marital status which is from the Monthly Population Survey for October 2001 (Cat. No. 6203.0).

Table 3: Selected Wave 1 Responding Person Characteristics and Population Estimates from the ABS for 2001 Compared (persons aged 15 years or over)

	HILDA ^b					
	ABS ^a	Unweighted	Design weights	Final HH weights	Weights adjusted for non-resp	Weights adj for non-resp and to ERP and LFS b'marks
Area of usual residence						
Sydney	0.217	0.173**	0.169**	0.203	0.216	0.217
Rest New South Wales	0.121	0.137**	0.145**	0.125	0.120	0.121
Melbourne	0.186	0.180	0.173*	0.183	0.184	0.186
Rest Victoria	0.066	0.077**	0.075*	0.067	0.065	0.066
Brisbane	0.086	0.083	0.088	0.087	0.087	0.086
Rest Queensland	0.100	0.106**	0.115*	0.102	0.100	0.100
Adelaide	0.058	0.062	0.061	0.062	0.059	0.058
Rest South Australia	0.020	0.031**	0.024*	0.023	0.022	0.020
Perth	0.073	0.072	0.075	0.073	0.073	0.073
Rest Western Australia	0.025	0.027	0.028	0.025	0.024	0.025
Tasmania	0.024	0.029	0.028	0.025	0.024	0.024
Northern Territory	0.007	0.005	0.005	0.008	0.008	0.007
Australian Capital Territory	0.016	0.017	0.016	0.018	0.017	0.016
Sex						
Males	0.495	0.474**	0.474**	0.475**	0.487	0.495
Females	0.505	0.526**	0.526**	0.525**	0.513	0.505
Age group						
15-19	0.086	0.085	0.087	0.093*	0.097**	0.088
20-24	0.090	0.073**	0.074**	0.079**	0.084	0.088
25-34	0.191	0.186	0.187	0.183	0.184	0.191
35-44	0.194	0.215**	0.217**	0.203	0.200	0.194
45-54	0.174	0.173	0.171	0.176	0.177	0.174
55-64	0.120	0.121	0.120	0.123	0.120	0.120
65+	0.146	0.147	0.144	0.143	0.138	0.145
Marital status						
Married	0.587	0.633**	0.634**	0.623**	0.615**	0.618**
Not married	0.413	0.367**	0.366**	0.377**	0.385**	0.382**
Indigenous status						
Indigenous	0.017	0.019	0.018	0.017	0.017	0.017
Non-indigenous	0.983	0.981	0.982	0.983	0.983	0.983
Birthplace						
Born in Australia	0.724	0.745**	0.744*	0.729	0.722	0.721
Born outside Australia	0.276	0.255**	0.256**	0.271	0.278	0.279
Main English-speaking countries	0.102	0.109	0.109	0.111*	0.109	0.109
Other countries	0.175	0.145**	0.147**	0.160	0.170	0.170

Table 3 (cont'd)

	HILDA ^b					
	ABS ^a	Unweighted	Design weights	Final HH weights	Weights adjusted for non-resp	Weights adj for non-resp and to ERP and LFS b'marks
Labour force status^c						
Employed	0.607	0.610	0.611	0.615	0.623*	0.610
Full-time	0.431	0.415**	0.417*	0.420	0.431	0.428
Part-time	0.177	0.195**	0.195**	0.195**	0.191**	0.182
Unemployed	0.043	0.044	0.044	0.044	0.044	0.043
Not in the labour force	0.349	0.346	0.345	0.341	0.333*	0.347
Employment status in main job (employed persons only)						
Employee	0.860	0.864	0.870	0.875*	0.877**	0.876*
Employer	0.036	0.041	0.039	0.037	0.037	0.037
Own account worker	0.100	0.086**	0.084**	0.081**	0.079**	0.080**
Contributing family worker	0.004	0.008**	0.008**	0.007**	0.007*	0.007*

Notes: ** and * denotes significantly different from the ABS population estimate at the 99% and 95% confidence levels respectively.

a The ABS estimates for area of usual residence, sex, and age apply to people in private dwellings, excluding those in remote parts of NT. The ABS estimates and labour force status apply to civilians in private dwellings, excluding those in remote parts of NT. The estimates for country of birth apply to civilians in private dwellings. The estimates for marital status, indigenous status, employment status apply to civilians.

b The HILDA estimates apply to persons in private dwellings, excluding those in remote parts of NT.

c We vary from the usual ABS definition in defining full-time work solely from the basis of usual hours worked (rather than on a combination of usual hours and actual hours worked). The ABS labour force status estimates for full-time and part-time workers have been adjusted to reflect usual hours worked based on information from the October 2001 Labour Force Survey.

Source: ABS data for area of usual residence, sex and age group are unpublished Estimated Resident Population data for September 2001, provided by the Demography Section. The data for labour force status are unpublished data from the Monthly Population Survey for October 2001, provided by the Labour Force Section. The remainder of the ABS data are from the Monthly Population Survey (Cat. No. 6203.0). Marital status and birthplace are from the October 2001 issue, and indigenous status, and employment status are from the August 2001 issue.

Weights Provided in the Wave 1 Datasets

Household File

The weights provided on the Household file include:

- Household population weight – accounts for the probability of selection and household non-response (as described in detail on pages 4-7). The sum of these weights is 7,404,297, being the total number of households in Australia, excluding those in remote parts of NT. Fully and partially responding households have been assigned a household population weight. You would use this weight to produce population estimates at the household level and in analysis in statistical packages that expect the weights to sum to the population.
- Household sample weight – household population weight scaled back so that the sum of the weights is the same as the number of responding households (that is, 7682). You would use this weight in analysis at the household level using statistical packages that expect the weights to sum to the sample size rather than the population.
- Household design weight – accounts for the different probability of selection arising from the sample design that each household may have had. The household design weight has been crudely adjusted for non-response by assuming non-respondents are the same as respondents. A common factor has been applied to the weights so that they sum to 7,404,297, being the total number of households in Australia, excluding those in remote parts of NT. The household design weight has only been given for fully and partially responding households. The design weight has been included on the file for users wishing to make their own adjustments for non-response.
- Enumerated person population weight – accounts for the probability of selection and household non-response (as described in detail on pages 8-11). The enumerated person population weight is provided for every person listed as part of fully and partially responding households. There are up to twelve weights provided on the Household file – one for each person listed as part of the household. The sum of the enumerated person weights is 19,013,602, being the number of people in private dwellings, excluding remote parts of NT. All enumerated persons in fully and partially responding households have been given an enumerated person population weight. You would use these weights in analysis of enumerated persons, such as in an analysis of children. No attempt has been made to reconcile enumerated person estimates with respondent person estimates, although the estimates should be comparable and not too divergent.

Person File

The weights provided on the Person file include:

- Responding person population weight – accounts for the probability of selection, household and person non-response. The sum of the responding person population weights is 15,119,487, being the number of people aged 15 and over in private dwellings, excluding remote parts of NT. The responding person population weight is provided for every responding person in fully and partially responding households – that is, those completing a personal interview). You would use this weight to produce population estimates at the person level and in analysis at the person level using statistical packages that expect the weights to sum to the population.
- Responding person sample weight – responding person population weight rescaled such that the sum of the weights matches the number of people providing a personal interview (that is, 13,969). You would use this weight in analysis at the person level using statistical packages that expect the weights to sum to the sample size rather than the population.
- Person design weight – accounts for the probability of selection arising from the sample design (and reflects the household design weight as every person in the household was selected into the sample). The person design weights have been crudely adjusted for non-response by assuming that non-responding persons are the same as responding persons. A common factor has been applied to the weights so that they sum to 15,119,487, being the total number of people aged 15 and over in private dwellings, excluding remote parts of NT. The person design weight has only been provided for people completing a personal interview. The design weight has been included on the file for users wishing to make their own adjustments for non-response.

References

Bednall, D., Cavenett, P. and Shaw, M. (2000), 'Response Rates in Australian Market Research', Paper presented at the Market Research Society of Australia Annual Conference, Sydney, October.

Henstridge, J. (2001), 'The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Weighting and Imputation', HILDA Discussion Paper Series No. 3/01 (<http://www.melbourneinstitute.com/hilda/hdps.html>).

Stukel, Hidioglou and Sarndal (1996), 'Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization', *Survey Methodology*, vol. 22, no. 2, pp. 117-125.

Watson and Wooden (2002), 'Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology', HILDA Technical Paper Series No. 1/02 (<http://www.melbourneinstitute.com/hilda/hdps.html>).

Appendix 1 – Technical Reference Group Membership

The Technical Discussion Group constructed to provide technical input to the weighting methodology included the following people:

Peter Boal, Department of Employment and Workplace Relations

Dr Tim Fry, Melbourne Institute of Applied Economic and Social Research

Dr John Henstridge, Data Analysis Australia

Stephen Horn, Department of Family and Community Services

Michael Meagher (replaced by Frank Yu), Australian Bureau of Statistics

Nicole Watson, Melbourne Institute of Applied Economic and Social Research

Frank Yu (replacing Michael Meagher), Australian Bureau of Statistics

Appendix 2 – Probability of Selection for Households and Persons

Household selection probability

The probability of selecting a household is given by:

$$\begin{aligned} P(\text{select household}) = & P(\text{select CD}) \\ & * P(\text{select block} \mid \text{select CD}) \\ & * P(\text{select dwelling} \mid \text{select block}) \\ & * P(\text{select household} \mid \text{select dwelling}) \end{aligned}$$

Probability of selecting the CD

The CDs were selected with probability proportional to their size as calculated by the number of dwellings in the CD. A systematic sample was selected by applying a random starting point and a skip to the list of CDs that were stratified by State and major statistical region and then sorted by statistical sub-division, section of State and serpentine ordering.

CDs that were in remote parts of Australia and CDs without a physical location were excluded from the sampling frame. A total of 488 CDs were selected.

The probability of selecting the CD is calculated as:

$$P(\text{select CD}) = \frac{(488 * \text{number expected dwellings in CD})}{\text{number expected dwellings in CDs on frame}}$$

The 1996 Census dwelling counts were used as the expected dwelling counts in each CD on the sampling frame.

Probability of selecting the block given the CD was selected

For CDs where there was no stage of block selection, the CD is assumed to have one block that encompasses the entire CD. For five rural CDs with large areas, an extra stage of block selection was introduced so the entire CD did not have to be listed.

The probability of selecting the block, given the CD was selected, is calculated as:

$$\begin{aligned} P(\text{select block} \mid \text{select CD}) \\ = & 1, && \text{if no block stage of selection} \\ = & \frac{\text{number blocks selected}}{\text{number blocks in CD}}, && \text{if block stage of selection} \end{aligned}$$

Probability of selecting the dwelling given the block was selected

A skip was applied through the list of dwellings in the selected blocks using a random starting point. A skip of 5 was used in urban CDs (that is, in metropolitan areas and rural centres) and a skip of 2 was used in non-urban CDs.

The probability of selecting the dwelling, given the block was selected, is calculated as:

$$\begin{aligned} P(\text{select dwelling} \mid \text{select block}) \\ = & \frac{\text{Total number of dwellings selected}}{\text{number actual dwellings in CD}}, && \text{if no block stage of selection} \end{aligned}$$

$$= \frac{\text{Total number of dwellings selected}}{\text{number actual dwellings in selected blocks, if block stage of selection}}$$

Probability of selecting the household given the dwelling was selected

All households in the dwelling were selected if there were three or fewer households. Where there were four or more households, a random sample of three households was taken.

The probability of selecting the household, given the dwelling was selected, is calculated as:

$$\begin{aligned} &P(\text{select household} \mid \text{select dwelling}) \\ &= 1, \quad \text{if three or fewer households in the dwelling} \\ &= 3 / \text{number households in dwelling,} \\ &\quad \text{if four or more households in the dwelling} \end{aligned}$$

Person selection probability

All household members are considered part of the HILDA sample, though only those aged 15 years and older are interviewed. The probability of selecting an individual given the household was selected is 1. Therefore, the probability of selecting a person is the same as the probability of selecting the household in which they live.

Appendix 3 – Models for Predicting Response to the HILDA Survey

Table A3.1: Logistic Regression Model to Predict Household-level Response

Variable	Odds Ratio	Standard Error	P-value
Security features			
Locked gate – no intercom access	0.6436	0.0747	0.0000
Locked gate – intercom access	0.6753	0.0735	0.0000
Security guard/doorman/on-site manager/gatekeeper	0.7184	0.1713	0.1650
Security door	0.8897	0.0424	0.0140
No trespassing sign	1.2484	0.4654	0.5520
Beware of dog sign	0.9998	0.1358	0.9990
Evidence of a dangerous dog	0.7807	0.1100	0.0790
No junk mail sign/no hawkers sign	0.7851	0.0799	0.0170
Neighbourhood watch sign	1.0091	0.0969	0.9250
Bars on windows	1.0675	0.1165	0.5500
Dwelling type (base category is separate house)			
Semi-detached	1.1232	0.0717	0.0690
Flat/unit/apartment - 2 storey or less	1.0479	0.0825	0.5520
Flat/unit/apartment - 3 storey or more	0.7934	0.1015	0.0700
Other dwelling - caravan, tent, cabin, etc	1.4758	0.4657	0.2170
External condition of dwelling (base category is 'Very good/excellent')			
Good	1.0249	0.0555	0.6500
Average	0.9373	0.0554	0.2730
Poor	0.7254	0.0727	0.0010
Very poor/almost derelict	0.4869	0.1216	0.0040
Highrises in area (base category is no highrises)			
A lot - more than 50% of structures	0.7534	0.1384	0.1230
A fair bit - more than 20% of structures	1.3646	0.2705	0.1170
One or two such structures	0.9416	0.1675	0.7350
Geographic location (base category is Melbourne)			
Sydney	0.7428	0.0598	0.0000
Brisbane	1.0596	0.0958	0.5220
Adelaide	1.1096	0.1025	0.2600
Perth	1.2836	0.1461	0.0280
Tasmania	1.1686	0.1972	0.3560
Northern Territory	0.4756	0.0943	0.0000
Australian Capital Territory	1.6212	0.2559	0.0020
Rural New South Wales	1.3076	0.1270	0.0060
Rural Victoria	1.3485	0.1793	0.0250
Rural Queensland	1.4343	0.1639	0.0020
Rural South Australia	2.0287	0.5724	0.0120
Rural Western Australia	1.6989	0.3277	0.0060

Table A3.1 (cont'd)

Variable	Odds Ratio	Standard Error	P-value
Neighbourhood characteristics			
Population density (population per km ²)	0.99996	0.0000	0.0050
Proportion speaking language other than English	0.7546	0.1461	0.1460
Proportion of people not in labour force	1.2413	0.4720	0.5700
Proportion of people unemployed	4.2117	4.5463	0.1830
Median weekly household income (base category \$500-699)			
\$100-299	0.8589	0.1935	0.4990
\$300-499	0.9669	0.0751	0.6650
\$700-999	0.9777	0.0712	0.7570
\$1000-1499	0.7823	0.0959	0.0450
\$1500-1999	0.7162	0.4508	0.5960

Table A3.2: Model Selection Process in Prediction of Household-level Response

	Likelihood ratio	df	Model compared to	χ^2	df diff	P-value
1. Original model	-7261.7	52	-	-	-	-
2. Remove proportion of different family types	-7262.4	47	1	1.30	5	0.9353
3. Remove median age	-7262.8	46	2	0.74	1	0.3912
4. Remove average household size	-7263.0	45	3	0.44	1	0.4433
5. Remove proportion of different dwelling types	-7264.7	42	4	3.23	3	0.3575
[NOTE – The following variables were not removed from the final response model.]						
6. Remove proportion of different employment characteristics	-7266.13	40	5	2.93	2	0.2307
7. Remove proportion of people speaking a language other than English	-7267.0	39	6	1.68	1	0.1953
8. Remove proportion of highrises in area	-7270.1	36	7	6.20	3	0.1024

Table A3.3: Logistic Regression Model to Predict Person-level Response, for Persons Aged 15+ in Responding Households with Two or More Eligible Persons

Variable	Odds Ratio	Standard Error	P-value
Geographic location (base category is Melbourne)			
Sydney	0.5832	0.0545	0.0000
Brisbane	0.8953	0.1196	0.4080
Adelaide	1.8716	0.3562	0.0010
Perth	0.9711	0.1406	0.8390
Tasmania	1.1226	0.2663	0.6260
Northern Territory	0.7627	0.3202	0.5190
Australian Capital Territory	1.3113	0.3809	0.3510
Rural New South Wales	1.5055	0.2047	0.0030
Rural Victoria	1.0597	0.1621	0.7050
Rural Queensland	1.0254	0.1371	0.8510
Rural South Australia	2.0399	0.5846	0.0130
Rural Western Australia	1.3207	0.3416	0.2820
Labour force status (base category is employed full time)			
Employed part time	1.8396	0.1849	0.0000
Unemployed	1.9771	0.3287	0.0000
Not in labour force	1.6507	0.1583	0.0000
Sex (base category is male)			
Female	1.6044	0.1126	0.0000
Age group (base category is 15-19)			
20-24	0.7303	0.0964	0.0170
25-34	0.7132	0.0953	0.0110
35-44	0.7447	0.1058	0.0380
45-54	0.8328	0.1206	0.2070
55-64	0.8761	0.1460	0.4270
65+	0.7750	0.1357	0.1460
Number of adults in HH (base category is two adults)			
Three or more adults	0.4038	0.0311	0.0000
Number of children in HH (base category is zero children)			
One child	1.3402	0.1314	0.0030
Two or more children	1.2563	0.1220	0.0190
Marital status (base category is not married or defacto)			
Married or defacto	1.9973	0.1856	0.0000
English ability (base category is only speaks English at home)			
Well or very well	0.6492	0.0584	0.0000
Not well	0.2613	0.0422	0.0000
Not at all well	0.0804	0.0234	0.0000

Table A3.3 (cont'd)

Variable	Odds Ratio	Standard Error	P-value
Dwelling type (base category is separate house)			
Semi-detached	1.0594	0.1409	0.6640
Flat/unit/appartment	1.4081	0.1971	0.0140
Other dwelling - caravan, tent, cabin, etc	0.5832	0.2301	0.1720

Table A3.4: Model Selection Process in Prediction of Person-level Response

	Likelihood ratio	df	Model compared to	χ^2	df diff	P-value
1. Original model	-3413.0	35	-	-	-	-
2. Remove long term health condition	-3413.0	34	1	0.01	1	0.9396
3. Remove housing tenure	-3413.6	32	2	1.07	2	0.5860
[NOTE – The following variables were not removed from the final response model.]						
4 Remove age	-3418.6	28	3	10.03	6	0.1235
5 Remove dwelling type	-3422.2	23	4	7.17	3	0.0666
6 Remove number of children in the household	-3428.4	21	5	12.41	2	0.0020

Appendix 4 – Effect of Adjustments on Weights

Table A4.1: Distribution of the Weights

	Mean	Min	Q1	Median	Q3	Max
Household-level						
Design weight	964	150	867	927	1002	2609
Adjusted for probability of response	964	176	835	908	1020	3646
Adjusted to benchmarks	964	166	802	913	1052	3820
Enumerated person-level						
Final household weight for all enumerated persons (adjusted to total persons)	955	165	769	906	1057	3801
Adjusted to benchmarks (ERP only)	955	141	753	906	1084	3992
Adjusted to benchmarks (ERP and LFS)	955	134	749	907	1086	3914
Responding person-level						
Final household weight for all responding persons (adjusted to total person 15+)	1082	183	896	1028	1191	4218
Adjusted for probability of response	1082	173	858	996	1204	6005
Adjusted to benchmarks (ERP only)	1082	144	833	1000	1226	6300
Adjusted to benchmarks (ERP and LFS)	1082	137	830	1001	1232	6276

Table A4.1: Distribution of the percentage change in the weights

	Mean	Min	Q1	Median	Q3	Max
Household-level (compared to design weight)						
Adjusted for probability of response	0.1	-25.2	-11.1	-4.6	5.5	167.3
Adjusted to benchmarks	0.4	-39.7	-11.8	-3.1	9.3	180.2
Enumerated person-level (compared to design weight)						
Adjusted to benchmarks (ERP only)	0.5	-55.4	-17.4	-2.3	13.7	204.8
Adjusted to benchmarks (ERP and LFS)	0.6	-58.3	-18.0	-2.2	14.2	197.8
Responding person-level (compared to design weight)						
Adjusted for probability of response	0.4	-44.7	-16.6	-7.0	11.2	238.0
Adjusted to benchmarks (ERP only)	0.4	-56.8	-19.4	-5.7	13.0	200.0
Adjusted to benchmarks (ERP and LFS)	0.5	-63.9	-20.1	-5.3	13.5	200.0

Note: The percentage change is calculated as $100 * (\text{new weight} - \text{design weight}) / \text{design weight}$.
The design weight of the contributing units has been adjusted to sum to the total of the relevant population.