DEPARTMENT OF
FAMILY AND
COMMUNITY
SERVICES

THE UNIVERSITY OF
MELBOURNE

HILDA PROJECT DISCUSSION PAPER SERIES
NO. 3/01, July 2001

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

John Henstridge
Data Analysis Australia

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

MELBOURNE INSTITUTE
of Applied Economic and Social Research

ACER

Australian Institute
of Family Studies

Introduction

The HILDA Survey aims to collect data from a representative sample of Australian households and residents. However no survey can ensure that this collection process is perfect and this is especially the case with a long-term panel survey where, even if the original sample is accurately representative of the population, there is likely drift over time. Hence there is a need to adjust the data so that it continues to accurately reflect the population of households and persons being surveyed.

The adjustment takes two forms:

- weighting of the data to overcome differences in the likelihood of various households and individuals being in the sample; and
- imputation of missing values where the data is incomplete.

Weights are a common issue in survey analysis. Alternative terms that are more descriptive are *expansion factor* and *population weight*, reflecting the use of weights to scale up from the sample to the whole population. If a sample was ideal, all the weights might be equal to the ratio of the population size to the sample size. However in reality they will be unequal to compensate for sample design constraints, operational difficulties and, most importantly with a longitudinal survey, the way that the panel develops over time.

This document only considers weights in the context of expanding the sample or adjusting it to the population. Statisticians have a second meaning of the term *weight*, as a measure of how precise a variable might be or how much confidence might be placed in it. Such a weight could be termed an *analysis weight* and should be inversely proportional to the variance of the error or uncertainty of the variable and it bears no relationship to the weights discussed above. Usually each variable in the survey will have a different analysis weight so it is not possible to talk of a weight associated with the sample unit itself. When conducting an analysis where the task is to model or explain a variable collected by the survey, the analysis weight is usually the appropriate one to use.

Imputation usually involves replacing a missing quantity with a reasonable substitute that permits a reasonable analysis to be carried out without being misleading. It may also extend to replacing values that, while not missing, are considered extreme. The major applications foreseen in the HILDA Survey are:

- the estimation of attributes of persons identified as household members but not available for interview; and
- the estimation of attributes where the interviewee refuses to answer the survey question.

This discussion paper considers the proposed weighting and imputation procedures for the HILDA Survey and outlines a recommended approach. In doing so it refers to the experience of the British and German panel surveys.

Definitions

Panel surveys are complex and it is important to clarify definitions used. Confusion even extends to what constitutes the panel – is it the set of households or is it the set of members of the households? In practice this will depend upon the precise design of the survey and the scope rules.

Here we use definitions associated with the HILDA design.

Continuing Sample Member	A person who will continue to be in the panel as long as contact can be maintained. Abbreviated to CSM.
Temporary Sample Member	A person for whom information is collected while they satisfy certain conditions, such as sharing a household with a CSM. When they cease to satisfy those conditions they may be dropped from the panel. Abbreviated to TSM.

British Household Panel Survey

The British Household Panel Survey (BHPS) is managed by the Institute of Social and Economic Research at the University of Essex and has operated since 1991. The initial sample was an equi-probability sample of British households. Ten years on, the panel is still based upon this initial sample using two mechanisms for maintaining its size:

- Considerable effort is put into tracking panel members so that the annual survey waves suffer minimal attrition. In a country the size of Britain, it is possible to maintain contact and access panel members almost regardless of where they move. This maintenance of the panel database is considered so central that it is carried out by the research team, unlike the data collection that is outsourced.
- Data collection is primarily face-to-face.
- A limited ongoing sample recruitment takes place whereby, as households split, all derived households with an original sample individual are included in the sample, and newborn children of a panel member are automatically panel members.

The result of this is a longitudinal sample that has been able to maintain a relatively modest variation in the weights, giving high sampling efficiency.

German Socio-Economic Panel

The German Socio-Economic Panel (GSOEP) has been run since 1984 by the German Institute for Economic Research (Deutsches Institut fur Wirtschaftsforschung or DIW), an independent research institute in Berlin. The panel currently consists of approximately 20,000 persons in 12,000 households, with annual interviewing being primarily by telephone.

The panel is significantly different from the BHPS in that it is several distinct segments corresponding to recruitment waves. The need to view these segments that had quite distinct initial sampling procedures as a single panel has made the weighting process more complicated. For example, the initial sample only selected German citizens and a subsequent sample selected non-citizen residents. Yet another sample extended coverage to the former East Germany. In addition the GSOEP has had relatively high attrition rate and is older, meaning that the adjustments to weights required to adapt to differential attrition have needed to be relatively complex.

One result of this has been a relatively high diversity of the weights.

Principles

In considering weighting and imputation it is worth defining the principles against which procedures can be judged. The following have been identified:

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.

Ideally weights should not vary from a constant value more than can be avoided since this reduces the statistical efficiency. This is obvious if the situation of spending significant effort on collecting data for a particular unit and then giving it very low weight – much of that effort is effectively wasted.

Where certain analyses may be restricted to subsets of the population, specialised weights may be required. However such weights should be as consistent as possible with the principal weights since they should not lead to contradictory analyses.

Where the survey is complex resulting in data at several levels, weights for different levels of the survey should be consistent with each other.

Imputation should not introduce significant biases or distributional changes into the data or introduce significant extra variance to estimators.

The imputation procedure should, as far as possible, be data driven from the sample rather than making external assumptions about the likely structure of responses.

Imputation should not lead to important sample estimates being based too heavily upon imputed values.

These principles can not always be satisfied simultaneously and hence compromises must be made. For example, if a population segment has very low probability of inclusion then ideally it should have a high weight to compensate. However the high weight may appreciably reduce sample efficiency so a compromise weight may be chosen. This would, however, introduce a small bias in some sample estimates.

The issue of compromise is a complex one. In some texts this is discussed as if the survey had the aim of measuring only one quantity so a mathematical treatment of the trade off between bias and variability can be carried out. However, for a survey such as HILDA, the data is complex and will be used for a number of different purposes. This leads to another broad but difficult principle:

The design of weighting and imputation systems should aim to give a dataset with broad application, while at the same time emphasising some variable as more important than others.

Consideration of these principles begins with the design – good design should minimise issues of both weighting and imputation. For example, the use of equi-probability sampling leads to equal weights, a desirable statistical property.

Issues Affecting Weights

Initial sampling

The first factor to impact upon weights is the initial sample design that determines which households are in the panel in Wave 1. Current plans are that the sampling will follow a three-stage process whereby:

- Across Australia but excluding specified remote areas, 488 Census Collector Districts (CDs) are chosen with probability proportional to size as measured by the number of dwellings enumerated in the 1996 Census.
- Within each selected CD, dwellings are enumerated and then a number are chosen for contact. This number may depend upon the region to reflect experience with response rates and will endeavour to ensure that on average 16 households will respond.

Ideally this will result in every household in Australia outside of the specified remote areas having the same probability of selection. Hence the initial *design weights* may be uniform. However it is likely that small variations in design weights may be required to account for:

- the actual number of households enumerated in a CD being different from the number used in the process of selecting CDs; and
- response rates varying from those expected.

It will almost surely be found that non-response affects the distribution of household types and the age-gender distribution of persons. Since population figures will be available for 2001 from the Census to be conducted in August, adjustment or post-stratification to such population benchmarks should be considered.

While it is too early to predict the range of design weights, they are unlikely to vary by more than 20% from the average value.

Panel Life

Much of the literature on panel surveys relates to designs where the participants have a relatively short life in the panel and where there is significant panel renewal. An example of this is the Australian Labour Force Survey that consists of a panel of households. Each household completes monthly interviews but stays in the panel for only eight months. Such a survey has a very different reason for choosing a panel design – it is chosen to ensure highly stable trend estimates over short time periods, typically month to month. The panel means that the sampling units from one month to the next largely overlap and hence differences will be estimated with high accuracy. Data from such a panel is typically not analysed longitudinally.

The HILDA survey is quite different – its whole aim is to collect data for longitudinal analysis. Hence the time between waves is substantial – one year – and ideally the data will be available over a number of waves. The changes in household structure over that time are likely to be larger and the problem of attrition is much greater. Hence issues such as following rules and weighting methods have much greater importance.

Following rules

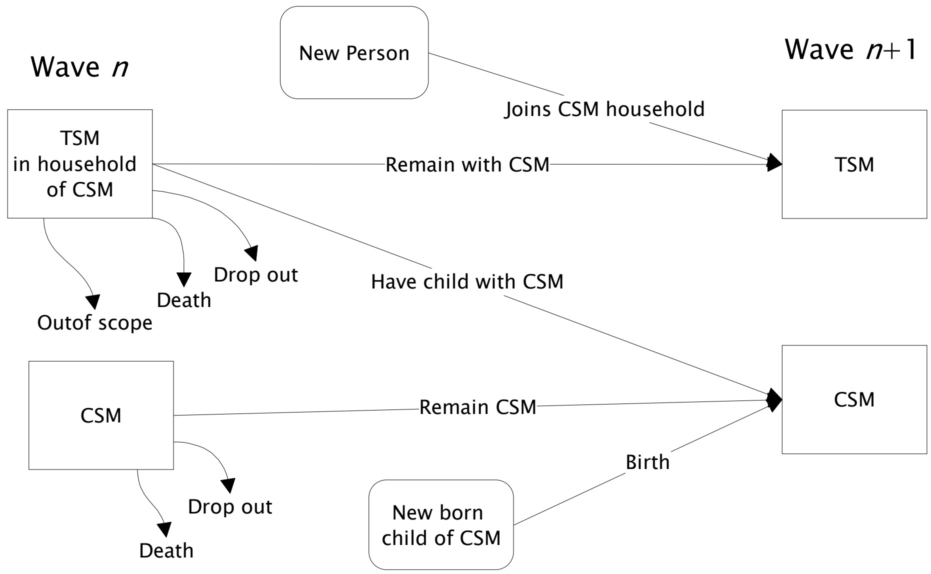
Between waves much can happen to the individuals in a household and, as a general principle, the HILDA following rules include in the sample *all* the individuals from Wave 1 and all other members of their current household. In detail, the rules state:

- All members of the Wave 1 households are defined to be Continuing Sample Members (CSMs).
- At any wave, all households containing a CSM are included in the sample, with members who are not CSMs being termed Temporary Sample Members (TSMs).
- Any children born to or adopted by CSMs will be CSMs.
- Parents of newly born CSMs will become CSMs if they are not already.

These are forward looking rules, aimed at defining who is in scope for the survey at each wave. It is best to view these following rules as *operational rules* rather than necessary sample definition rules. Differing analyses may need to restrict the sample in different ways and provided that the following rules are broad enough, this should not be a problem.

However for calculating weights at each wave, it is necessary to look backwards and to consider how each individual and household *could* come to be included so that a probability of inclusion can be calculated. It is also necessary to consider how the sample may change over time. The structure can be illustrated inFigure 1.

Figure 1. Possible transitions in the state of an individual between waves.



The probabilities of each event illustrated can be estimated from the data itself, recognising that they will depend upon the personal or household characteristics. Two issues of concern are:

- Is there are enough information on the non-panel members to estimate the probabilities of entering a household with a CSM? The difficulty is that the survey observes only those who join, not those who do not join. This issue is discussed further in Section 0.
- Will the sample change substantially over time? The following rules are taken from the BHPS where they appeared to keep the panel size reasonably stable. However there is no reason to believe that the structure of the panel remains the same.

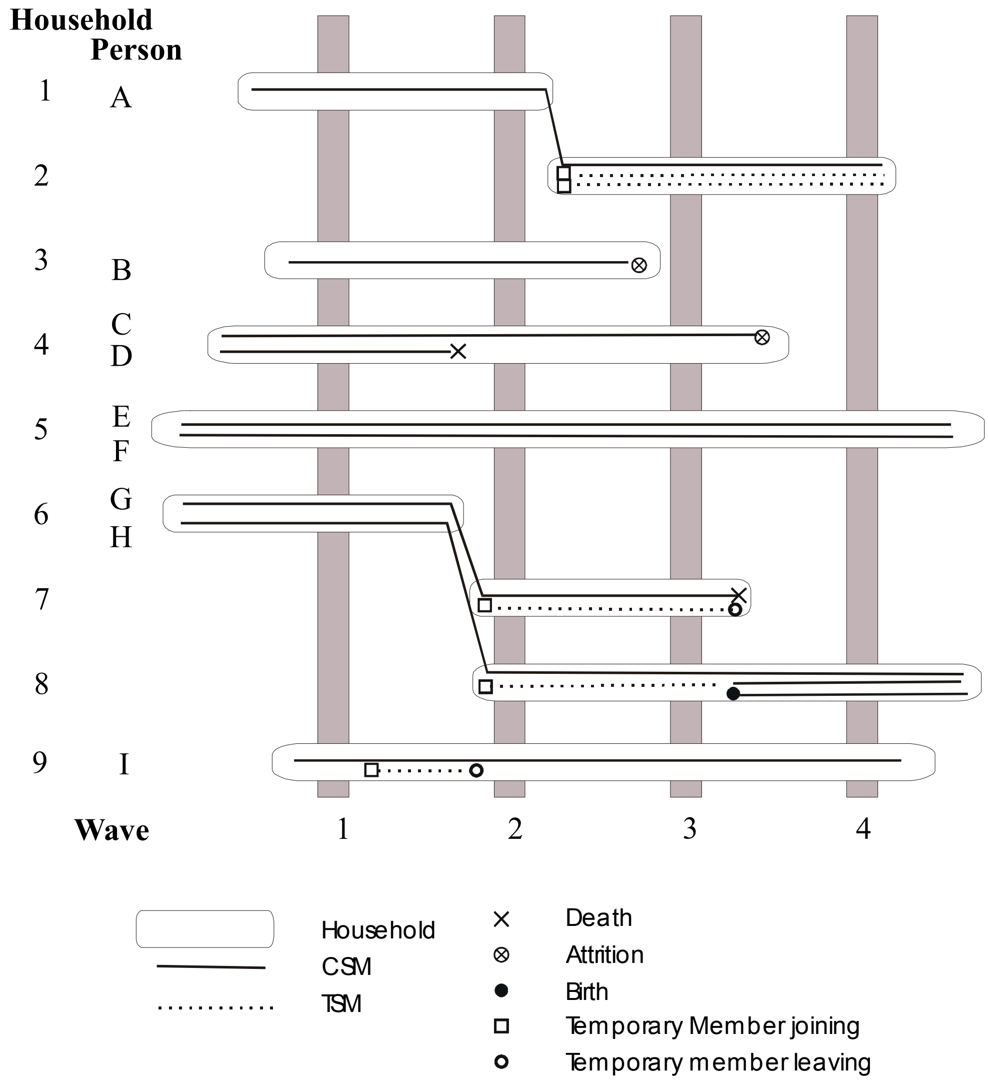
A diagram similar to that in Figure 1 can be produced for households but it is much harder to define the longitudinal structures. In effect, a household is in the panel for a given wave if any its members are CSMs, making the remainder TSMs. However it is not always easy to uniquely define what is the continuing household entity between waves. Ernst (1989) discussed this in formal detail and Folsam, LaVange and Williams (1989) provide a highly asymmetrical concept of principal predecessor and principal successor families. These differing definitions have less effect on rotating panel surveys where a household might only be in the sample for less than a year. For a long term survey it can have a significant effect.

A household panel example

The HILDA Survey is designed to supply data for a broad range of uses and hence the nature of future analysis that will be performed on it cannot be fully specified. Furthermore the panel structure gives rise to a potentially complex data set. To understand this it is worth following the example in Figure 2 below. Note that this example considers ‘households’ but the issue of just what is a household is not always straightforward – this is discussed in more detail in Section 0.

Here we have a panel observed over four waves, represented by the vertical bands. In Wave 1 six households are observed, with numbers 1, 3, 4, 5, 6 and 9 and those households consist of eight persons A, B, C, D, E, F, G, H and I, all of whom are CSMs.

Figure 2. An example of a panel maintained over four waves.



By the time of Wave 2 several changes have taken place:

- Household 4 has lost one member (D) who died but is still considered to be the same household.
- Household 6 has split with person G going to the new household 7 and H going to household 8. Each of these new households also has a TSM.
- Household 9 gained a member who left again before Wave 2. This person was not available for interview so only proxy information is available. It is valid for a survey design to not consider including such individuals and to restrict the scope to persons present in the household at the survey times. This is the approach taken in HILDA.

Hence at Wave 2 seven households are observed but only five of these were in the original sample. By Wave 3 further changes have taken place:

- Household 1 has undergone a change – the addition of two extra members – which mean that it is now considered another household, number 2.
- Household 3 has dropped out of the survey, a clear case of attrition.

Hence at this stage the panel is back to six households but only seven of the original persons. By Wave 4 further changes have taken place:

- Households 4 and 7 have dropped out, 4 due to a physical move that led to being uncontactable while person G in household 7 died. Since household seven is no longer followed, its temporary member is no longer considered part of the survey.
- In household 8 a child was born to the CSM and the TSM. The child is considered to be a CSM and the TSM converts to a CSM.

In the end only four households remain, only two of which were in Wave 1. Only four out of the original nine persons remain.

Various summaries can then be considered over the four waves:

		Wave 1	Wave 2	Wave 3	Wave 4
Households	Total	6	7	6	4
	Original	6	5	3	2
CSMs	Total	9	8	7	7
	Original	9	8	7	5
TSMs		0	2	4	2
Persons		9	10	11	9

It can be seen that the numbers of original households and members declines over time – they can never increase. Also the counting of original households over time depends upon the definition of a continuing household. The panel can be expected to settle down over time with a mix of CSMs and TSMs although the precise mix and size of the panel will depend on the balance of the events bringing persons into the panel and those removing them.

What is the population?

Before one can answer the question as to whether the panel is functioning properly and how it should be weighted, it is necessary to decide what population is being studied. Various options for this population are:

1. The cohort of individuals alive at Wave 1 – all individuals are tracked until they move out of scope or are lost through attrition, with no new members entering the sample;

2. The cohort of households existing at Wave 1 – all households are tracked until they move out of scope, cease to exist as the same household or are lost through attrition;
3. The population of individuals alive for the duration of the panel – in practice similar to 1 but restricting the sample to those successfully tracked through all waves;
4. The population of households that exist for the duration of the panel – in practice similar to 2 but restricting the sample to those successfully tracked through all waves;
5. Individuals who were alive at some stage in the life of the panel – all people surveyed at any stage in the life of the panel; or
6. Households that existed at some time during the life of the panel.

In any consideration of households it is necessary to have a definition of what constitutes the continued existence of a household between waves (termed here the previous wave and the following wave). Ernst (1989) gives several criteria that to some extent may be applied in combination:

- **Same householder or reference person**. This requires the definition of who the reference person should be – at one stage the concept of 'head of household' would have been used.
- **Same spouses**. If this is applied to households with married couples it is reasonably clear and the break up of such a couple means that the previous household terminates. Such a definition may be more difficult to apply in the case of de facto or same-sex relationships and group households. It has the advantage that it is symmetrical with respect to time.
- **No change**. The household must have precisely the same members in the two waves.
- **Reciprocal majority**. The majority of household members from the previous wave remain with the household *and* the majority of the members in the next wave were members in the previous wave.
- **Reciprocal plurality**. The household in the following wave has more members from the household in the previous wave than has any other household in the following wave and the household in the previous wave has more members in the household in the following wave than any other household in the previous wave.
- **Same type**. The household should remain of the same type, using a classification typically into single households, adult couples, adult couples with children etc. This must be used in conjunction with another criterion.

In the example of Section 0 the reciprocal majority rule was used. (Note that in the case of Household 4 a strict application of this rule might suggest that the death of one of the two persons in the household should lead to the end of that household and the creation of a new one. This might be logical in the case of a household splitting into two, but might be questioned where the change is due to death.) If the no change rule was applied there would have been eleven distinct households instead of nine. Such a difference could naturally affect the weights applied to households while at the same time it should logically leave the definition of individual weights unchanged.

Clearly the problem of defining a population of households is much more difficult than defining a population of persons. In addition, the appropriateness of a definition is likely to depend upon the questions being asked. For example,

- In some cases the analysis would better proceed using a concept of *family* rather than household;
- If the question relates to household stability, the population of all households that existed in the panel (option 6 above) becomes the best concept.

It is unrealistic to believe that one set of longitudinal household weights will satisfy all needs. Hence, initial priority is best given to assigning appropriate weights to persons. In this case we do not need to consider continuity of households between waves but only the affect of households on the probability that any individual is in the sample.

Panel evolution

When the panel is established in Wave 1 it will be, subject to minor adjustments, a representative sample of households and persons. Over time changes will take place as members are added to or leave the panel. It is not automatic that these changes allow the panel to remain representative.

This evolution can be illustrated by several hypothetical examples:

Example 1. A demographer's approach to a panel might be to consider only the females and to follow them through without loss. The only way of leaving the panel would be death and the only additions would be female offspring. Such a panel would remain representative of the female population, with the exception that it does not allow for migration. This approach is often preferred by demographers since maternal parentage is usually better defined than paternal parentage. The panel size would be stable since births would approximately match deaths in the same way that it does for the population.

Example 2. A panel in Example 1 may accurately follow females and could be extended by including, as TSMs, the males in the households defined by the females. This would be stable over time but would not be representative of males – for example those living in male-only households are excluded.

Example 3. If, as in the proposed HILDA panel, the CSMs include males and there are additional means by which new CSMs are generated – female births to the male CSMs and the conversion of the female partners of male CSMs to CSMs on the birth of a baby of either sex – the situation changes substantially. In the absence of attrition the panel would grow substantially, approximately growing by a factor of 2.5 with each generation, corresponding to a growth rate of about 3.7% per annum.

Example 4. If attrition is introduced at about 3.7% per annum to Example 3, the panel might stay at a stable size, with the excess new members numerically balancing the attrition. However the excess new members would enter at either age zero (a birth) or at child bearing age (a mother). The ages of attrition would not match this and hence the panel would evolve to a stable age distribution almost surely different from that of the population. If attrition is above 3.7% the panel would decline and if it is below

3.7% the panel would increase in size. In both cases, the stable age distribution would differ from that of the population and would depend upon a variety of factors but particularly the ages of attrition.

Example 5. (The actual proposed HILDA sample.) TSMs are added to the panel of Example 4. These are likely to further modify the age distribution although it is not clear whether the modification would be towards or away from the population age distribution. The ratio of TSMs to CSMs can be expected to settle to a constant after a period.

The appropriate theoretical methodology for understanding panel evolution is the Leslie matrix used in population biology, providing a means of calculating the changes in panel size and the development of stable age distributions. This does not appear to have been applied to panel studies.

The BHPS experience has been that their panel has remained relatively stable in size, suggesting a balance between attrition and recruitment. However, the above examples illustrate that the age-sex distribution is likely to have changed significantly. The GSOEP experience is different. They had more recruitment since they termed everyone who shared a household a CSM but higher attrition. The continual augmentation of the GSOEP sample to maintain the size also reduces drift in the age-sex distribution.

The importance of weights in a panel survey is that they are the only means of adjusting an evolving panel for the drift of the age distribution away from that of the population.

Types of Weights

Weights will always depend upon the type of analysis being carried out, but there will always be four basic types depending upon whether they relate to:

- persons or households; and
- cross-sectional or longitudinal studies.

The calculation of weights should consider the sample design for Wave 1 when the panel is recruited and the subsequent development of the panel.

Cross-sectional personal weights

These will be used when a single wave of data is used for a standard analysis. If household information is used, it is considered as an attribute of the persons in the household. Cross-sectional personal data is readily benchmarked (post-stratified) against external data such as the Census so that it has the opportunity to be quite accurate. These weights are the easiest to calculate.

Cross-sectional household weights

Cross-sectional household data is likewise relatively easy to use. In this case some personal data might be considered as an attribute of the household. A typical way of calculating a cross-sectional household weight might be to average the personal weights of its members – this leads to a good level of consistency between cross-sectional household and personal analyses.

Longitudinal personal weights

In many cases a longitudinal personal weight is restricted to the surviving part of the original Wave 1 sample. Obviously the surviving part of the sample decreases over successive waves so that a new set of longitudinal weights might be associated with each wave, with the weights associated with lost persons set to zero.

This approach to personal longitudinal weights will be acceptable initially but may become inefficient when significant numbers of TSMs are in the panel, many for extended periods.

Longitudinal household weights

Longitudinal household weights are less frequently used, largely because of:

- the difficulty in defining a longitudinal household; and
- the issue of differing households being defined in the panel for a limited time, resulting in a different set of weights for each possible interval of study.

These two issues suggest that the provision of longitudinal weights for the initial cohort of households is the most that is feasible. Even then, it is necessary to update the weights with each wave to adjust for attrition. Hence after the n^{th} wave, the Wave 1 data will have n sets of household weights.

Relationships between weights

At the cross-sectional level it is both possible and expected that the household and personal weights should be consistent with each other. For example, each household sample unit will have as an attribute the number of persons in the household. An estimate of the total population is the weighted sum of this attribute. The person data unit will have one record per person and the sum of the weights for these units will also be an estimate of the total population. These two estimates should agree.

This agreement is easy to achieve if all the persons in a household have the same weight. If weights within a household differ then a minimum requirement is that the household weight should equal the average of the personal weights. However in this latter case differences in estimates will arise for some estimates of subpopulations such as the total number of males.

Personal weights are complicated by the fact that some household members will be identified but may not be eligible, refuse or not be available for a personal interview. Hence all are enumerated but only a subset respond. This leads to the possible need for *enumerated person weights* (that are likely at Wave 1 to equal the household weights) and *responding person weights*. The latter will be zero for persons who do not respond. Ideally however the response rate would be 100% so that the two types of weight would be equal.

Response rates

At various stages in the survey, households or individuals may refuse to participate. In the case of Wave 1 households, these will be replaced in the sample by households in the same area (cluster) who are prepared to participate. At that and later stages there may be some individuals who do not respond to the detailed questionnaire.

If the unobserved characteristics of the refusing households or individuals are the same as those who participate, this is not a significant problem. Obviously the design attempts to achieve this by, for example, ensuring that Wave 1 refusals are replaced by households in the same cluster.

In practice however there will be relationships between the household or individual characteristics and the probability of refusal. Hence the weights should attempt to adjust for any such biases.

This is typically done by dividing the sample into response classes and weighting to ensure that each response class has appropriate final weight. For example, if non-response leads to single person elderly households being under represented in the sample, then those that are in the sample should be given higher weight.

Attrition

Attrition is a form of non-response but one that is more difficult to manage. It is a household or person permanently dropping out of the sample after Wave 1, either by choice or through the difficulty in contacting them. It creates particular problems in a longitudinal survey since it limits the period for which the household or individual is observed and means that different households or individuals are observed for different periods.

Like initial non-response, attrition is likely to be related to the characteristics of the households being studied. It is likely that factors that lead to attrition – family disruption, employment changes and relocations – are variables of significant interest in the survey creating a situation where persons of greatest interest may well be the most difficult to collect longitudinal data for.

Adjustment for attrition involves increasing the weights of remaining panel members. There are a number of ways of doing this but they all attempt to increase weights of members most like those who have left. The two most common methods are:

- To divide the sample into cells and to increase remaining weights in cells where attrition has occurred. Ideally the cells should reflect both reporting frameworks (so that major estimates are not affected by attrition) and structural aspects of the data so that the reweighting does not bias results.
- Modelling attrition using logistic regression. The variables used to explain attrition are likely to be similar to those used in defining cells in the former method - if they are exactly the same then the methods are identical. The logistic regression will then give an estimate of the probability of attrition for each unit. The weights of remaining units can be increased by an amount proportional to this probability.

Related to attrition is the problem of wave non-response – households or individuals that miss a wave but then re-enter the panel. Operationally they are a problem in that until they reappear they are likely to be treated as attrition. From an analysis viewpoint they are a problem since the longitudinal data is incomplete. Imputation may be required to infill the data.

Note that attrition is quite different from death, which is an observed natural event. Adjustment for attrition should not explicitly adjust for deaths, although adjustment to population benchmarks does do so.

Calculation Procedures

Wave 1 weights

With the exception of the non-response issue, Wave 1 weights are largely determined by the sample design. Since household surveys are well understood and survey teams are experienced in their implementation, it is likely that close to equal probability sampling can be achieved at the household level.

Weights after Wave 1 – A modelling approach

At each wave the weights attached to individuals should inversely reflect the probability that that individual had of being in the sample. In the case of original CSMs, this is clearly their Wave 1 weight appropriately modified by a factor inverse to their probability of *not* dropping out through attrition.

For new entrants to the panel after Wave 1, the weights are more complex. The proper starting point is Figure 1 that presents the means or events by which a non-sample member can join that panel. The situation is the simplest in Wave 2 since at that stage no TSM has had the opportunity to leave the panel.

The probability of a TSM being on the panel can be written in probability terms as:

P(TSM being in panel) = **P**(TSM joining a household)×**P**(that household in panel)

A statistical model is required for each of a number of types of TSM (such as age-gender classes) and probably type of household to give the first factor. The second factor is a relatively simple one, reflecting the types of households in the panel. In fact, it would appear that all the information required is collected by the panel, although it is not clear how accurate the models can be estimated. Since the household type is important the probability is probably better represented as:

$$P(\text{TSM in Panel}) = \sum_{\text{household types}} P(\text{TSM joins household of given type}) \times P(\text{Household of that type in panel})$$

The following calculation steps would be required:

Classify the Wave 1 persons into person types, and households into household types.

For each person type and each household type calculate the probability of such a person joining such a household, using the Wave 1 and Wave 2 data.

Use the Wave 1 persons as an estimate of the proportion of persons of each type in the population.

Use the Wave 1 households to estimate the probability of each household type being in the panel. This step is almost trivial since each household type should have had equal probability of selection into Wave 1.

Calculate the probability of selection for the TSM.

The difficult step here is the second – calculating the probabilities of joining the panel. The best source of such information is the panel itself since the panel already contains a sample that is representative of persons who might join the panel. For example, if we are interested in the characteristics of persons who might join a type of panel household, we can examine who in the panel joins another household. Hence the panel will collect the appropriate data to carry out the required analysis to develop a model for persons joining households in the panel. However, theoretical feasibility does not always imply practicality. The analysis requires careful attention to definitions and data quality and it would be unwise to assume that it will be possible, especially in the early years of the panel.

For waves after Wave 2 the same procedure is followed except that at Step 4 it will be necessary to use the Wave 1 households to estimate the true proportions and numbers of households in the population and compare this with the current wave to estimate the probability of panel inclusion for each household type.

For waves after Wave 1 the probability of a TSM being in the panel is also affected by the scope rules – it would be necessary to model the probability that they have been added to the panel.

A modelling approach can be applied at any level of detail, depending upon how many types of households and persons joining households are used. In the simplest case, all households are treated the same, as are all persons joining the household. The resulting weights are likely to be stable since they will be the same for all entrants, but clearly will have some biases. Having too many types will require more complicated analysis and more data and will result in less stable weights.

If a household has a high weight in Wave *n*, it is logical that the households related to it in Wave *n*+1 should also have a higher weight. In particular, if the household itself continues into the next wave with little change in its membership, then its weight should not change significantly. The relationship between households in successive waves is clearly measured by their overlapping members and the inheritance of weights is clearly through these members. Ernst (1989) outlines these principles and showed the interplay between optimal weighting systems and definitions of relationships between households. These systems have many attractive properties such as being unbiased but rapidly become complex in the presence of attrition.

Trimming of weights

After the first wave it is possible that the sample will include some individuals or households with a low probability of selection. For example, a TSM may have at Wave 1 been out of scope and hence the way of being in the sample in Wave 2 would only be through joining a panel household, an event of much lower probability than having being in the sample in Wave 1. Hence the ideal weight calculated as the inverse of the probability of sampling will be potentially large compared with those of other individuals.

It is useful to consider a simple measure of the sampling efficiency

$$E = \frac{\left(\sum w_i\right)^2}{n\sum w_i^2}$$

where w_i is the weight of unit i and n is the sample size. This efficiency is maximised by having all the weights equal and minimised by having all but one of the weights zero.[1]

To maintain sampling efficiency, potentially large weights are to be avoided. At the same time, departure from the inverse probability calculation introduces bias. Compromises may be required and it is a not uncommon process to trim the extreme weights, a process sometimes called 'Windsorisation'.

The BHPS limited weights to 2.5 times the average weight. This choice is reasonable, since introducing into the sample a unit with twice the weight of the otherwise almost equally weighted units, gives no overall improvement to sample accuracy. Units with higher weight actually detract from overall sample precision, although to exclude them may introduce bias.

Keeping units with such weights can be justified in terms of minimising bias, ensuring complete coverage and providing adequate sample for certain subsets. In this light, the BHPS threshold of 2.5 seems not unreasonable but it is lower than is commonly applied.

The GSOEP does not trim weights and has a consequent higher variability but no biases from this source. It is understood that weights vary by several orders of magnitude.

The key is to review any such trimming process using logistic regression to model the likelihood of a unit being trimmed to determine whether the trimming affects certain types of person or household more than others and whether the effect of trimming grows with each wave. Fortunately, the problem is unlikely to be significant for the first few waves provided that the Wave 1 sampling is effective.

Adjustment to population benchmarks

Irrespective of the above issues, there is usually a requirement to match the sample to population measures. For example, it may be considered necessary to ensure that the weights bring the age-gender breakdown of the sample in line with the population as measured by the Census so that the panel can be used as a valid cross-sectional information source. This step is sometimes termed *post-stratification*. The sequence in which these adjustments are made can have an impact upon the final weights – the processes do not commute. If the adjustments are minor then the order of adjustment is much less critical.

Note that if the panel attrition is modelled and adjusted for correctly, the subsequent adjustments to population norms should be minor. Hence it is important to correctly

[1] This measure does not consider all aspects of sample efficiency. It ignores the reduction in efficiency due to correlations within clusters of a sample.

address attrition. Conversely, if a simple approach is taken to modelling panel recruitment and attrition, the role of adjusting to benchmarks becomes more critical.

Estimation of standard errors

The survey should deliver estimates of a known accuracy. Hence there must be means of calculating standard errors for the most important survey estimates. The methods must account for the sample size, the sample design and the weights.

The following rules will mean that after Wave 1 some households in the sample will be related to each other, having derived from the same Wave 1 household. In addition, there will be a structure related to having several individuals in each household. These will give a complex correlation structure on the data in addition to the normal effects due to the cluster design.

The only feasible methods of calculating such standard errors collectively termed resampling procedures, and include the jackknife, bootstrapping and 'half sample' methods. (See for example Lehtonen and Pahkinen, 1992.)

For example, Rendtel (1991) applied a 'random groups' procedure to the GSOEP data to investigate the development of sampling errors over waves. With each wave, the panel will have a sample of households, each of which can be linked back to one household in Wave 1. Since households can and will drop out but none can enter except via a link to the Wave 1 sample, the number of Wave 1 households remaining relevant to the current panel can only diminish. This will lead to an increase in the correlation between panel households and a subsequent increase in standard errors. The level of this increase will depend upon the item being considered, being the greatest for items that might be 'inherited'. The example given by Rendtel suggests that over five years the standard error for estimates of political preference increased by 14% *due to this effect alone,* corresponding to a drop in sample efficiency of 24%.

These increases in standard errors can be reduced by minimising attrition. The only way to avoid the effect is through the addition of new households not related to the original ones.

Imputation of missing values

When a data value is missing three choices are available:

- it can be left as missing and subsequent analysis can be restricted to cases with only non-missing values;
- it can be left as missing and subsequent analysis can use models to account for the missing values; or
- it can be replaced in the data set by an estimate and subsequent analysis can proceed as if the estimate was an actual data value.

In most survey contexts the last approach is often taken.[2]

[2] The first can lead to a propagation of missing values – for example if one expenditure item is missing then so will total expenditure. The second can lead to inconsistent treatment and is inconvenient for any but the most experienced users of the data.

Generally imputation should ensure that the average of the imputed values should match the average of the responses for similar sampling units. This means either using the average itself (single value imputation) or some random quantity that has that average as its expectation.[3] The BHPS has used a form of 'hot deck' imputation where the missing value is replaced by a randomly selected response from similar sampling units. The purpose of random replacement is to ensure that the imputed value not only leads to correct estimates of means but also gives correct indications of variability. Quantities such as average are affected by weights and the imputation process should therefore use weighted averages or random selection proportional to weights.

Hence the only questions that really need to be addressed are what is meant by *similar sampling units* and what is the impact upon the data. The definition of similar sampling units should be based upon information available for all sampling units, but at the same time should be designed to best predict the missing values. The BHPS used a CHAID modelling approach to this but there could be others.

One major impact upon the data is that imputed values are not the same as real values. They are based upon the remainder of the data and hence do not really carry any unique information. This means that the survey standard errors should reflect the number of survey units with real values, not the total number.

Where values are imputed it is good practice to flag them as such in the data set. This increases the size of the database since most data fields will need an associated status field of at least one byte. However, failure to do so is likely to lead to misunderstandings in future use of the data, with the danger that studies using specialised subsets of the data being unduly dominated by imputed values.

Outlier detection and processing

Related to missing values are outliers – extreme values as responses to items. There are two reasons for considering these:

- they may represent errors in the data and should thus be corrected or set to missing; and
- while being correct responses, they may add unacceptable variability to estimates.

The first situation is relatively easy once an outlier is detected. If it is clearly incorrect then the rawest form of the data is checked. If there is no indication as to the true value it must be considered missing.

The second is much more difficult since it is rare in a survey that just one estimate is being produced. Hence the criteria of unacceptable variability are not clear, although a common principle is that no single data value should contribute more than a certain proportion of specified key estimates. For example, it might be decided that no household income value should contribute more than 10% of a state estimate of average household income.

[3] Some authors have suggested more robust measures such as the median but they almost invariably introduce biases in sample estimates.

Once the decision is made that an extreme value is unacceptable by such a criterion, it is usual practice to replace it with the most extreme acceptable value. Typically this introduces a bias through the removal of the 'excess contribution'. It is possible but not common practice to then reallocate this contribution to other sample units to maintain unbiasedness of estimates.

Commentary

What can weighting achieve?

Most of the theoretical literature on sample surveys including household panels assumes that there are a few well defined population parameters, the estimation of which is the primary aim of the survey. This can simplify the theory but does not always help in practice. In common with many surveys, HILDA has many aims and its data will be put to many uses, some of which cannot be imagined at this stage.

Hence an over-riding design principle must be that of flexibility and the need to not exclude alternative definitions, weights and analyses at an early stage. This is nowhere more apparent than in the definition of a household from one survey wave to the next. Ernst (1989) illustrates how differing definitions of a continuing household can dramatically affect both weighting and conclusions.

In addition, it is clear that the various weighting schemes that have been used by panel surveys can rarely adhere to the statistical ideals. For example, adjusting the panel in each wave to match broad population parameters will almost surely lead to small inconsistencies in longitudinal weights. Not adjusting will lead to cross-sectional data sets for each wave with known biases.

While it may appear less than rigorous, recognising that perfection is not achievable is probably worthwhile. The emphasis must be on getting workable sets of weights that can be used for most purposes and recognising that they have limitations. These could be termed the 'reference weights' and their calculation should be regarded as a key part of the survey task. Alternative weighting systems could be used but analysis based on them should clearly identify that the reference weights have not been used.

General principles again

Within the limitation of what a single set of weights can do, the weights should attempt to:

- Give proper weighting to each household in each wave, so that each wave can be used as a valid and representative cross-sectional data set. The definitional problem associated with continuing households should be irrelevant to this.
- Give proper longitudinal weighting to individuals, again avoiding the issue of households – there is no question as to whether an individual is the same one as in previous waves.[4]

[4] There will, however, be a question of scope as after the survey has been running for an extended period the use of the original sample to define scope for the longitudinal sample becomes inappropriate.

- Ensure that longitudinal household weights are consistent with the longitudinal individual weights.

These principles lead to a weighting system similar to the BHPS. However, several potential differences will remain, due to some extent upon quantitative rather than qualitative reasons.

The BHPS uses the share approach (well outlined by Lavallee and Hunter, 1992) with the added steps of post-stratification and trimming of weights. This depends upon their concept of Permanent Sample Members (PSMs) that is essentially the same as HILDA's CSM concept, however possibly different recruitment and retention rates may mean that the post-stratification step will have a different sized adjustment. It is worth noting that the post-stratification and trimming carried out by the BHPS conflicts with the unbiasedness concept in the share approach as described by Ernst.

Long-term maintenance

Continual 'topping up' of the panel is often considered in the light of correcting for attrition. However, it can have a second role that is just as important – ensuring that the panel does not significantly evolve over time to the extent that it is no longer representative of the current population. In addition the new sample members, being independent of the existing sample, would ensure that standard errors are constrained from growing.

Ideally these new households should be selected through a sampling scheme essentially the same as in Wave 1. This would ensure that the weights attached to them are firmly based on an independent and well understood sample design. Since they are statistically independent of the Wave 1 households, they will not contribute to the degradation of standard errors – rather they would reverse this effect.

The frequency of introducing new households is largely an issue of cost. Since HILDA uses a cluster sampling design for Wave 1, ideally the new households should use a similar approach. This should be considered for Wave 4 after a review of panel evolution. (See Section 0 below.)

Computation of weights and data management

The calculation of weights is intimately linked with the purposes to which the data will be put. It is likely that several weighting schemes will be trialed over the life of the project.

Hence it is essential that the weighting process be well documented and totally reproducible. It should be implemented as a set of computer algorithms, probably in the language of the database system chosen for the project, and developed to the stage where it will run without manual intervention. This may have implications for the choice of database system.

It is also probably worth having sufficient detail in the distribution file from the survey to enable this weighting or variants of it to be repeated. This is likely to lead to improved systems being developed and users considering weighting systems for specific questions. In this regard the note of Winsborough (1989) is pertinent.

Recommendations

The weighting approach of the BHPS is conservative relative to that of the GSOEP and their following rules are closer to those of HILDA. The following rules largely determine panel evolution and panel evolution largely determines the need for sophisticated weighting.

Note that this approach does not include model-based weighting as is done with the GSOEP. However, the information needed to fit models for weight estimation is not yet available and ideally should be based on several years HILDA data. In addition, the GSOEP approach has led to substantial variation in weights, resulting in reduced sample efficiency.

Recommendation 1. The principal weighting scheme for HILDA should follow the BHPS in having weights that are comprised of the initial Wave 1 sample weight with adjustments applied at each Wave to adjust the panel sample to population demographics.

The BHPS experience in weighting should be actively pursued, with a copy of the data being obtained so that the practical implications of their weighting procedure can be better appreciated. Many potential issues will be shown by the BHPS to be insignificant in practice. The user documentation for the BHPS only summarises the weighting and imputation issues – more detailed technical information should be obtained.

Recommendation 2. Initial household weights derived from the sample design should be applied to the individuals in Wave 1, with the same adjustment process to bring the sample to population demographics being used in all waves including Wave 1.

Wave 1 weighting is not a significant issue since it is a standard survey - the problems associated with a panel only become evident at Wave 2.

Recommendation 3. The population demographics applied to adjust weights should be the estimated resident population for each year by state/territory, metropolitan/non-metropolitan, age (ten year groups) and gender.

This weighting strategy will yield a dataset that accurately represents the population for a number of years. The adjustments will largely compensate for panel evolution. However it will not be perfect and hence it should be reviewed after three to five years to examine whether the weights continue to provide the best possible option.

Recommendation 4. The weighting system should be reviewed when data from three Waves is available for analysis. In particular, a model-based weighting system should be considered.

This review is critical to the ongoing development of the panel since the ability of a refined weighting procedure to correct for problems that may creep into the panel will affect decisions on panel maintenance procedures.

Recommendation 5. Imputation should be used to overcome missing values, with a 'hot deck' procedure to ensure maintenance of variation. The imputation classes should be defined through a statistical analysis of the Wave 1 data when available.

This approach matches the BHPS and will ensure that the sample remains representative, while allowing most analysis to proceed unhindered by missing values.

Recommendation 6. Outlier treatment should use the principle of the maximal allowable contribution of any single data record to key estimates. The criteria should be developed through an analysis of Wave 1 data.

Recommendation 7. Calculation procedures for weighting, imputation and outlier treatment should be implemented programatically to ensure correctness and objective application of procedures.

Only if the procedures can be reproduced can there be confidence that they are understood and implemented properly.

Recommendation 8. The distributed dataset should clearly flag all variables as being original data, imputed or treated as an outlier.

This is a database and documentation issue. The value of this approach will be seen later on when users of the data will have less familiarity with the way that it was collected.

Recommendation 9. Depending upon achieved attrition rates and costs, augmentation of the sample in subsequent waves should be considered as a means of maintaining the sampling validity of the panel over extended periods.

Sample augmentation does not have to be annual but should probably be conducted every four to five years. It is suggested that it be actively considered for Wave 4 after the review recommended above.

References

Ernst, L.R. (1989) 'Weighting issues for longitudinal household and family estimates', in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. (eds), *Panel Surveys,* John Wiley and Sons, New York.

Folsom, R., La Vange, L. and Williams, R. (1989) 'A probability sampling perspective on panel data analysis', in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. (eds), *Panel Surveys,* John Wiley and Sons, New York.

Lavallee, P. and Hunter, L. (1992) 'Weighting for the survey of labour and income dynamics', *Proc. Statist. Canada Symp. 92, Design and analysis of longitudinal surveys*.

Lehtonen, R. and Pahkinen, E.J. (1994) *Practical Methods for Design and Analysis of Complex Surveys*, Revised Edition, John Wiley and Sons, Chichester.

Rendtel, U. (1991) ‘Weighting procedures and sampling variance in household panels’, *Working Papers of the European Science Foundation Scientific Network on Household Panel Studies*, No. 11, Essex.

Winsborough, H.H. (1989) ‘Database management: discussion’, in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. (eds), *Panel Surveys,* John Wiley and Sons, New York.