Survey Research Methods (2022)
Vol. 16, No. 3, pp. 301-328
doi:10.18148/srm/2022.v16i3.7851
European Survey Research Association

Choosing who to follow: The long-run impact of following rules on the sample size and composition of household panel surveys

Nicole Watson
University of Melbourne, Australia, and
University of Queensland, Australia

Following rules for a household panel survey determine who is followed and interviewed when there are changes to the household structure over time. To understand the long-run impact that various following rules have on the size and composition of the sample, a household panel sample is simulated such that every household split-off is developed over a 40-year period. This long time period allows sufficient time for children to be born into the simulated sample and subsequently leave home and have their own children. Following every split-off allows for the application of the widest following rules along with a range of narrower following rules. Holding everything else constant (such as non-response rates, propensities for household formation and dissolution, fertility rates, death rates, etc.), the widest possible following rules result in a sample that is much larger and has quite a different composition to the samples with narrower following rules. Even within the narrower following rules, some small differences in sample size and composition appear by wave 16 and these become more apparent by wave 41.

Keywords: Following rules; household panel surveys; sample characteristics; sample size; simulation

1 Introduction

It is relatively common in analyses of household panel survey data for the unbalanced panel to be used, which includes all observations on all individuals regardless of how those individuals came to be part of the sample. Depending on the variables of interest, further restrictions may be made to keep only adults or specific sub-groups. In household panel surveys this unbalanced panel includes not just the original sample members recruited in wave 1, but also other people that these original sample members share a household with over time. A set of following rules determine who is followed over time if someone leaves the household. These following rules expand the sample followed to include new births but may also include the other parent of these births, new immigrants and other household members not already part of the followed sample. Further, it is also normal practice in most household panel surveys to interview all adults living with a followed sample member each wave. So even if the very narrowest of following rules are adopted, there are temporary additions to, and removals from, the sample simply because people move into and out of sampled households. These household composition changes are linked to various life cycle processes, including birth, leaving home, beginning a de-facto or marital relationship, separating from a relationship, having children, moving in with adult children, and death. Understanding how people become part of the sample and the role of the following rules is important in understanding the data available from household panel surveys and in designing new household panel surveys.

The specifications for the following rules are determined early in the life of the household panel survey (albeit with some studies amending them later) yet they have long-term consequences to the overall size of the sample and fieldwork costs. There are other factors that also affect the size of the sample, including the response rates in the initial wave, attrition rates, new sample member response rates, birth rates, death rates, marriage and cohabitation rates, divorce rates, the age of leaving the family home and the propensity for other types of household joiners and leavers. Nevertheless, the following rules and (at least to a certain degree) the response rates are within the control of the survey manager so can be designed with the impact on sample size and the budget in mind. Panel maintenance strategies have been discussed in the literature at length elsewhere (for example, see Laurie, Smith, & Scott, 1999; Smith, Lynn, & Elliot, 2009; Watson, Leissou, Guyer, & Wooden, 2018) yet decisions around the following rules have received scant attention. Further, the following rules and response rates work in opposite directions in affecting the overall sample size: a study anticipating very high response rates may decide to choose

Contact information: Nicole Watson, Level 5, FBE Building, 111 Barry Street, University of Melbourne, Victoria 3010, Australia (E-Mail: n.watson@unimelb.edu.au)

quite narrow following rules so as not to increase the sample size too quickly whereas a study with lower response rates may decide to adopt wide following rules to maintain (or at least not reduce too quickly) the overall sample size. The US Panel Study of Income Dynamics (PSID) began in 1968 with two samples—a general sample and an oversample of poor families (largely African American)—with 4800 families responding and a total of 18,200 individuals included (Institute for Social Research, 2019, Table 1). The PSID follows the original sample members and their descendants born after the study began. By 1983 (15 years later), the general sample of individuals had grown by 9% whereas the oversampled group had grown by 14%. By 2009 (41 years after the study began), the general sample of individuals had grown by 86% from the 1968 sample (the subsample of the poor was substantially reduced in 1997 due to budget constraints so is not reported here). The re-interview rates from one wave to the next for the PSID are very high, typically between 96 and 98% (Institute for Social Research, 2019, Table 8). Further, their interview rates of split-offs (e.g. adult children leaving their family home) are also reasonably high, around 85 to 90%. As the PSID only interviews one person in the household, mostly the household head, the first time an adult child is interviewed is often when they move out of home, so the response rate of new split-offs is very important for the longevity of the study. Since 2009 the PSID has lower re-interview rates of 93 to 96% so the sample size has started to decline recently.

Some household panel surveys have modified their following rules later in the life of their study in response to an emerging issue. The German Socio-economic Panel (SOEP) and the Swiss Household Panel (SHP) both began with relatively narrow following rules as per the PSID, but expanded them in later waves (wave 7 for the SOEP and wave 9 for the SHP) to include all people joining a household with a sample member. For the SOEP, this change was intended to minimise interviewer error in deciding who to follow up and for the SHP the change was to attempt to counteract attrition (Schonlau, Watson, & Kroh, 2011). The Household, Income and Labour Dynamics in Australia (HILDA) Survey changed their following rules in wave 9 to include people who recently arrived in Australia since the study began in 2001 (Summerfield et al., 2019). This was only a partial solution for missing recent immigrants until a more general solution could be actioned for this part of the population in wave 11 with a top-up sample which included recent immigrants (Watson, 2006). The problem with changing the following rules after the start of the panel is that they cannot be completely retrospectively applied. For example, in the case of the HILDA Survey, recent immigrants who had appeared in early waves but had subsequently left the sample households before the rules were changed could not be followed up.

Despite the impact that following rules can have on the size and composition of the sample, this is an area of research that has received little attention. One earlier study by Schonlau et al. (2011) examined how following rules affect sample size. They used the German SOEP, which follows everyone, to examine the differences in sample size if narrower following rules had been adopted. They find that even after 25 years the narrowest following rules of only following wave 1 sample members and associated household members accounts for 85% of the sample obtained by implementing the widest following rules of following everybody. This finding is, however, specific to the SOEP setting which has lower re-interview rates than several other major household panel surveys (Watson et al., 2018). There does not appear to be any studies of how the following rules affect the composition of the sample. Further, these following rules can change the composition of the sample in ways that may not be anticipated by researchers using the data. For example, the sample may grow in household types that frequently change (for example, larger households) and reduce in single person and couple-only households. Relatedly, the sample may on average become younger over time if the household joiners are much younger than the household leavers.

Of course, the survey statisticians who construct the weights for the household panel data are certainly aware of the implications of the impact of household joiners and the possibility that the household could have been selected through multiple pathways and therefore the need to down weight households with new entrants that are not births in the cross-sectional household weights (Heeringa, Berglund, Khan, Lee, & Gouskova, 2011; Lavallée, 1995; Lynn, 2006; Schonlau, Kroh, & Watson, 2013; Watson, 2012). The following rules also have implications for the construction of longitudinal weights. Decisions need to be made as to who to include in the longitudinal weights (original sample members, or other permanent sample members such as descendants of original sample members) (Smith et al., 2009). And if weights are provided for longitudinal populations that do not include the initial wave, decisions need to be made as to whether short- or long-term temporary sample members are included in the longitudinal weights (Watson, 2012). As a result, if data users apply the weights, issues resulting from the following rules will be addressed. But what if the users do not apply weights? Or do not take account of the changes in the composition of the unbalanced sample simply due to the following rules? What if the users are comparing the results from two different household panel surveys which have different following rules—could some of the differences found be simply due to the different following rules? Very few users appear to take this into consideration.

This paper examines what impact the following rules have on the sample composition via a simulation of a randomly selected sample followed through time. The household formation and dissolution information from the HILDA Survey is used to simulate a sample with wide following rules and no non-response (where everyone who joins a household is subsequently followed and participates fully). The simulated sample is then modified by the application of more restrictive following rules and by attrition. Differences in the distribution of a range of variables including age, sex, household size, household type and employment are examined. Through this simulation, the impact of the following rules can be distinguished independently from the influences of attrition. The findings from this simulation will assist users of household panel datasets in understanding the sample evolution over time in the unbalanced panel and how this may vary across surveys with different following rules. Further, it is also valuable for survey managers designing new household panel surveys to understand how the sample may grow and evolve over time based on decisions made about following rules very early on in the life of the survey.

2 Methods

2.1 Following rules

The term continuing sample member (CSM) is used to denote the sample members that are followed and temporary sample member (TSM) refers to those that are interviewed for as long as they live with a CSM but are not followed when they leave the CSM's household. Some TSMs may be part of the sample for a very long time if they continue to live with a CSM and others may only appear for one or two waves. Interviewing the TSM helps to provide the context for the household in which the CSM lives.

When deciding on the specific set of following rules a study will adopt, there are six different sample groups to consider:

1. Wave 1 sample members. Almost all household panel surveys include all wave 1 sample members (also known as original sample members, or OSMs) as CSMs. An exception may occur for boost samples of certain sub-groups of the population where only members of those sub-groups are considered CSMs and other household members in wave 1 of the boost sample are considered TSMs. This approach was taken for the Ethnic Minority Boost Sample in the initial wave of UK Household Longitudinal Survey, also known as Understanding Society (Institute for Social and Economic Research, 2019). In the boost sample, the sample of primary interest is the CSM belonging the key sub-populations and not the other people they live with other than to provide the context for the household in which they live.

2. Births to CSMs. For studies that (subject to funding) intend to run indefinitely, children of CSMs are converted to CSMs. In this way, the sample replaces itself over time. For studies of short duration, such as the Canadian Survey of Labour and Income Dynamics, which was designed as a 6-year rotating panel (Lavallée, 1995), there is no reason to add new births (or any other household joiner) as CSMs. For studies that do convert births of CSMs to CSM, decisions need to be made as to whether all such births should be added or if a subset of births should be added. A birth is only observed through the selection of one or both of their parents, so for replication purposes the births only need to be included in proportion to the number of CSM parents they have. This avoids oversampling children which may lead to potentially unsustainable sample growth when these children start to leave home, assuming minimal attrition. The UKHLS have operationalised this as only including births to female CSMs (Institute for Social and Economic Research, 2019). Alternatively, the births that are followed could be selected randomly in proportion to the number of parents that are CSMs. This option would limit the sibling research opportunities but would avoid the problem in the UKHLS following rules of limiting intergenerational analysis within the maternal line only. Most indefinite life studies choose to follow all births to CSMs. This allows for sibling research alongside intergenerational research within both the maternal and paternal lines. Adoptions are usually treated in the same way as births. The following rules for the PSID and the HILDA Survey explicitly include adoptions (Institute for Social and Economic Research, 2019; Summerfield et al., 2019), but they are not explicitly mentioned in the description of the following rules for the British Household Panel Study (BHPS) (Taylor, 2017).

3. Other parent of birth to CSM. The other parent of a newborn may be converted to a CSM if they are not already a CSM. This is so that both parents can be followed should they later separate after the birth of a child. This permits the study of the environment in which the child grows up in and the care arrangements agreed between the parents from both perspectives.

4. Other births to other parent. It may also be interesting to follow the other children the other parent may have (with another partner) after they have a child with a CSM. This allows greater opportunities for insights into the subsequent family dynamics of split families.

5. Recent immigrant joiners. People who arrive in the country after the initial sample (or any subsequent refreshment sample) was selected do not have a natural mechanism for becoming part of the sample. However, those that do live with a CSM can be converted to a CSM and then will be followed if they leave the CSMs household. This aims to reduce some of the consequences of undercoverage of recent immigrants as the study ages (Watson, 2006).

6. Other household joiners. Some studies choose to follow everyone who joins the household of a CSM. This has the potential to incorporate many new sample members on a continuing basis, depending on retention rates. The drawback is that these new sample members are not a cross-section of the population.

Table 1
Following rules adopted by various household panel surveys

Sample type	PSID[d] (1968)	SOEP[e] (1984)	BHPS[f] (1991)	SLID[g] (1994)	SHP[h] (1999)	HILDA[i] (2001)	UKHLS[j] (2009)
1. Wave 1 sample members	Y	Y	Y	Y	Y	Y	Y
2. Births to CSM[a]	Y[b]	Y	Y	N	Y	Y[b]	Y[c]
3. Other parent of birth to CSM	N	Y	Y	N	Y	Y[b]	Y
4. Other births to other parent	N	Y	N	N	Y	Y	N
5. Recent immigrant joiners	N	Y	N	N	Y	Y	N
6. Other household joiners	N	Y	N	N	Y	N	N

[a] Continuing sample member (i.e., individual who is followed) [b] Adoptions to CSMs are included
[c] Only of female CSMs [d] US Panel Study of Income Dynamics [e] German Socio-economic Panel
[f] British Household Panel Study [g] Canadian Survey of Labour and Income Dynamics [h] Swiss Household Panel
[i] Household, Income and Labour Dynamics in Australia Survey
[i] UK Household Longitudinal Survey (aka Understanding Society) General Population Sample.

There is a surprising amount of variability in the following rules adopted by various household panel surveys around the world. Table 1 shows the following rules adopted by seven studies. The PSID follows all wave 1 sample members and those subsequently born to or adopted by those sample members—and now after 50 years, there are as many as seven generations represented in the PSID sample (Institute for Social Research, 2019). The German SOEP and the Swiss Household Panel both began with following rules modelled on the PSID but subsequently changed these rules in wave 7 and 9 respectively, as mentioned earlier, to follow everyone who joins a sampled household. The BHPS built on the PSID rules and includes the other parent of the newborn as a CSM if they are not already a CSM (Taylor, 2017). They do not, however, follow any subsequent children this other parent may have. To distinguish them from the regular CSM, they refer to these other parents as permanent sample members (PSMs). The HILDA Survey adopted the BHPS following rules but makes three additions: i) adoptions to CSMs are treated the same as births, ii) the other parent of a birth to a CSM is treated in the same way a wave 1 sample member would be treated which means that any subsequent children this other parent may have are also followed; and iii) people who arrive in Australia for the first time after the study began and join sampled households are followed when they leave (Summerfield et al., 2019). This rule was subsequently revised after the addition of a general population top-up sample in 2011. The UKHLS adopted the BHPS following rules but walked back one of the rules so as to only follow births of female CSMs (Institute for Social and Economic Research, 2019). And finally, the SLID only followed people in wave 1 as it is a short longitudinal panel of 6 waves.

For the simulation presented in this paper, the six sets of following rules examined are those from the following studies: i) PSID; ii) SOEP; iii) BHPS; iv) two variants of the HILDA Survey rules; and v) UKHLS. The first version of the HILDA-style following rules do not include recent immigrants as CSMs but the second version of the HILDA-style following rules do include them. In terms of assessing what impact these following rules have, it was decided to simulate 40 waves of household change to allow time for new births to occur in the sample, grow up and leave home.

2.2 Data

This paper uses unit record data for the first 18 waves from the HILDA Survey, which is an Australian-wide household panel survey that began in 2001 (Department of Social Services & Melbourne Institute of Applied Economic and Social Research, 2019; Summerfield et al., 2019). Interviews are conducted annually and the core content asked every year includes questions on employment, income and families. Rotating modular content includes questions on wealth, retirement, fertility, education and health. One person in the household answers questions about the household, including changes in household composition, housing and childcare. Subsequently, interviews are conducted with all household members aged 15 and older. The HILDA Survey sample has a multi-stage, stratified and clustered design with 7682 responding households in the initial wave from a total of 11,693 in-scope households, resulting in a household response rate of 66%. The sample is restricted to usual residents in private dwellings and excludes people living in very remote parts of Australia. The wave 1 responding households contain 19,914 individuals, which encompass 13,969 respondents, 1158 non-responding adults and 4787 children aged under 15. Individuals from these responding households are followed over time and, if aged 15 years and older, interviewed. The sample is extended in subsequent waves to include other members of their household that were not part of the initial household as elaborated earlier. Over 90% of the interviews are conducted face-to-face with the remainder conducted by telephone. No proxy interviews are permitted. In wave 10 (2010) the proportion of wave 1 respondents that were re-interviewed, excluding people who had died or moved abroad, is 71% and by wave 18 (2018), it is 62%. By international standards, the HILDA Survey has one of the highest re-interview rates (Watson et al., 2018) and has reasonably wide following rules (without following everyone). In wave 18 (2018), the responding sample includes 18,324 individuals in 7616 households (not counting the top-up sample that was added in 2011).

2.3 Simulation Method

The HILDA Survey data is used to construct a simulated sample of complete cases with wide following rules to which a series of narrower following rules are applied. The impact of attrition is also modelled. This simulation is then used to estimate differences between samples (both with and without attrition) to demonstrate the impact of particular following rules.

Preparing the data for simulation. A long file of 323,225 person-wave observations is initially constructed from 18 waves of HILDA Survey data and this file has one observation per individual per wave for each wave they were part of a responding household. Information from the next wave is added to each record, this being the household identifier, whether the household underwent a transition in the next wave by gaining or losing members, and what type of household each household member transitions into (single person household, couple without children, couple with children, lone parent family, multifamily household, other type of household).

Missing information regarding household structure changes between one wave and the next is imputed (as described below) to provide a set of complete cases from which the simulation is constructed. Figure 1 provides two examples of the type of information that may be missing. Solid lines in this figure show what is known and dashed lines show what is unknown. In the first example in Figure 1, persons A and B form a household in wave t, but they do not respond in wave $t + 1$, so it is unknown if a transition occurred in their household structure and if it did then to what type of family they moved into. They could have stayed intact, or they could have split. For this type of situation, a transition indicator is predicted (i.e., deciding if they stay intact or split), along with the type of household they transition into if a transition is imputed. Note that a specific household is not imputed for A and B if a transition is imputed at this stage so to allow for multiple possibilities in the simulation should A and B's household be included. The second example shows a household with persons C, D and E in wave t, say a mother, father and adult son, and at wave $t + 1$ the son moves out of the household. It is known that the son moves out because the parents respond in wave $t + 1$, however it is not known to what type of household the son moves into because the son did not respond in wave $t+1$. Is he living alone, or has he moved in with a partner, or with housemates? To complete the household information for person C's household in wave $t + 1$, the type of household he moves into is predicted and, for simplicity in the simulation that follows, a specific household is imputed for him in wave $t + 1$ (the methods used to complete these part split households are similar to those used in the simulation below and are described in the Appendix).

On average over the first 17 waves (comprising 304,901 person-wave observations), there were 4% of person-wave observations where it is unknown if a transition occurred or not in the next wave because of attrition (like the first example in Figure 1). There were a further 4% of person-wave observations where it was known a transition occurred but it is unknown what type of household the sample members transitioned into because at least one other split-off from the household responded in the next wave (like the second example in Figure 1 or they were TSMs and were not followed).

Imputing transitions. A transition indicator is imputed for person-wave observations where it is unknown whether a transition occurred in the next wave or not. This is done by using the predicted value from multilevel logistic regression model of household transition from wave t to $t + 1$ based on the characteristics of household and the household head. Only the household head for each wave is included in the model. The model includes random effects at the wave 1 household level to allow for clustering within repeated observations of the household head over time and household splits emerging from the same initial household. New sample members are assigned the same wave 1 household group as the people in the household they join. The person defined to be the head of the household is in order of priority part of a couple, a lone parent, or other household member. Where more than one person fulfils the requirement for household head, the older person is chosen. The characteristics of the household included in the model of household transition are whether the period of time the household has been intact is left truncated, the number of intact years (6 groups), the number of adults, the age group of the oldest child (4 groups, including whether there are no children in the household), the age group of the youngest child (4 groups, including whether there is no youngest child), whether the household has a dependent student, a non-dependent child, other relatives, unrelated persons, or is a multi-family household. The characteristics of the household head include 5-year age dummies, employment status, and health status. Table 2 provides the coefficients for the variables included in the transition model. The predicted fixed and random components are both used to calculate the predicted probability for a household transition (of any type) occurring in the next wave. All household heads with a missing transition indicator are randomly assigned a transition indicator based on the predicted proba-

Table 2
Coefficients for model predicting household transition between wave t and t + 1

Variable	Coeff	S.E.
Whether years intact is left truncated	−0.05	0.04
Years intact (*Ref. categ.:* 1 year)		
2 years	-0.06^{**}	0.02
3 years	-0.16^{***}	0.03
4 years	-0.30^{***}	0.03
5 to 9 years	-0.51^{***}	0.03
10 or more years	-0.70^{***}	0.04
Number of adults in HH (*Ref. categ.:* 1 adult)		
Two adults	0.11^{***}	0.02
Three adults	0.59^{***}	0.04
Four or more adults	0.74^{***}	0.06
Age oldest child (*Ref. categ.:* no children)		
Aged 0 to 4	0.01	0.04
Aged 5 to 9	0.15^{***}	0.04
Aged 10 to 14	−0.05	0.04
Age youngest child (*Ref. categ.:* no younger child)		
Aged 0 to 4	0.45^{***}	0.07
Aged 5 to 9	-0.18^{**}	0.07
Aged 10 to 14	−0.11	0.07
Has dependent student	0.23^{***}	0.04
Has non-dependent child	1.09^{***}	0.03
Has other relative	0.99^{***}	0.04
Has other non-relative	1.56^{***}	0.05
Multifamily HH	1.58^{***}	0.08
Age of HH head (*Ref. categ.:* 15–19)		
Age 20–24	−0.00	0.07
Age 25–29	−0.09	0.07
Age 30–34	-0.16^{**}	0.07
Age 35–39	-0.44^{***}	0.07
Age 40–44	-0.85^{***}	0.07
Age 45–49	-1.09^{***}	0.07
Age 50–54	-1.18^{***}	0.07
Age 55–59	-1.30^{***}	0.07
Age 60–64	-1.53^{***}	0.07
Age 65–69	-1.83^{***}	0.08
Age 70–74	-2.02^{***}	0.09
Age 75–79	-1.76^{***}	0.09
Age 80–84	-1.41^{***}	0.09
Age 85 plus	-0.69^{***}	0.09
Employment status of HH head (*Ref. categ.:* Full-time)		
Employed part-time	−0.00	0.03
Unemployed	0.12^{***}	0.04
Not in labour force	0.05^{*}	0.03
HH head has long term health condition	0.09^{***}	0.02
Constant	-1.43^{***}	0.09

* $p < 0.1$ ** $p < 0.05$ **** $p < 0.01$

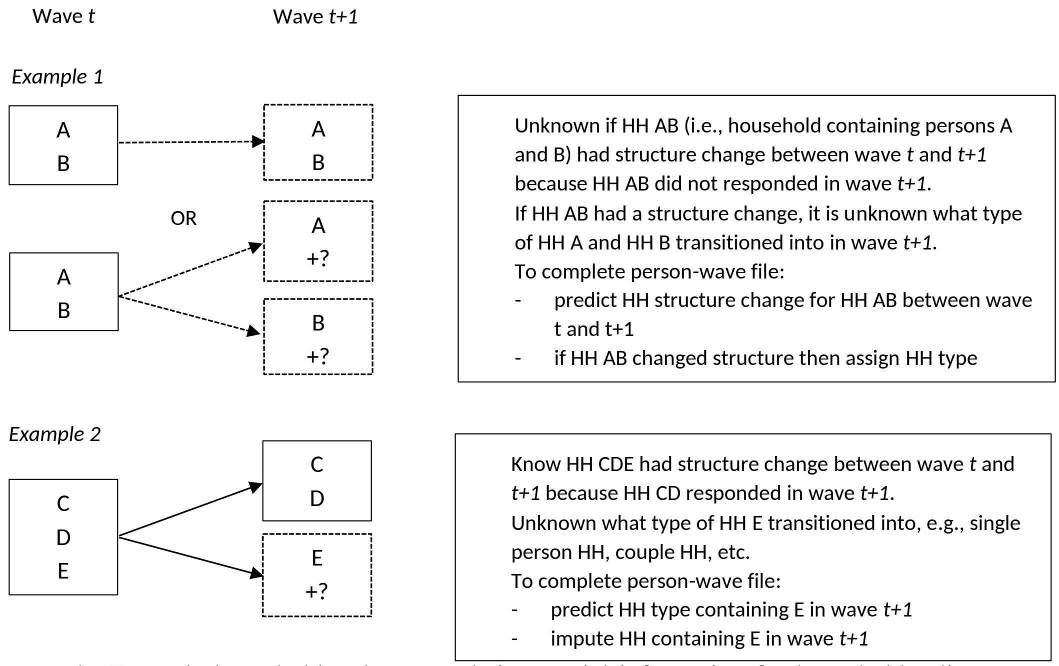

Figure 1. Example households when completing partial information for household splits. *Note:* Dashed lines indicate the status is unknown. A dashed arrow indicates that is unknown if a transition in household (HH) structure occurred between wave t and $t + 1$. A dashed box indicates that it is unknown what the household structure is in wave $t + 1$.

bility of a transition. This imputes a household transition for 26% of cases with a missing transition indicator (compared to 21% of households with transitions among those where it is known the household has a transition or not). The higher rate of transitions in the imputed group is reflective of the types of households that need to be imputed (i.e. those more likely to not respond or have TSM leavers).

Imputing broad family type. The individual-level transitions of broad family type observed in the HILDA sample from one wave (wave t) to the next (wave $t + 1$) is shown in Table 3. For the simulation, the broad family type into which the household head transitions, conditional on a transition having occurred, is imputed if it is not already known. This broad family type of the household head is later used, together with other characteristics of the household, to select a suitable household donor for the transition from wave t to $t + 1$. A multinomial logit model for each family type in wave t is estimated based on observed transitions where broad family type is known in wave $t + 1$ (see Table A2 in the Appendix). Only the head of the household (as defined earlier) is included in the model. The characteristics of the household included in the model are whether the period the household has been intact is left truncated, the number of intact years, whether the household contains other relatives or unrelated persons. The characteristics of the household head include 5-year age dummy variables, employment status, and health status. Similar variables are used by other researchers to study household transitions (for example Mutchier & Burr, 1991; Qu & De Vaus, 2011; Richards, White, & Tsui, 1987). All household heads missing the family type in wave $t+1$ are randomly assigned to one of the six family types based on the predicted probabilities estimated for each broad family type they could transition into.

Wave 1 sample. The wave 1 sample of the HILDA Survey is taken as the initial wave of the simulation from which waves 2 to 41 are simulated. This sample includes 7682 responding households comprising of 19,914 individuals (15,127 adults and 4787 children under the age of 15).

Simulating wave 2 to 41. For each subsequent wave, the simulated sample from wave t is followed into wave $t + 1$ using the completed person-wave file for waves 1 to 17 described earlier (wave 18 is used only to identify which households make a transition between wave 17 and 18 and into what sort of family the household moves into). Figure 2 shows some examples of how households may be simulated. Where a household has an observed household or (completed) set of household splits in the next wave, the details for wave $t + 1$ are taken directly from the completed person-wave file (as in Example 1 in Figure 2). Where a household is not observed in the next wave (as in Example 2 at wave 5), hot deck imputation is used. The hot deck imputation method replaces missing values of a recipient (i.e., household with missing information that need to be imputed) with observed values of a donor (i.e., household with com-

Table 3
Individual-level transitions of broad family type from wave t to t+1, conditional on a change in household structure

	Broad family type wave $t+1$						
Broad family type wave t	Single person	Couple w/o children	Couple w. children	Lone parent	Multi-family	Other	N
Single person	-	44.7	20.7	18.0	2.0	14.7	3,739
Couple without children	20.9	16.2	52.2	2.2	6.4	2.0	6,802
Couple with children	7.2	11.2	64.6	11.4	4.2	1.3	26,883
Lone parent	15.1	4.8	23.1	48.4	5.8	2.8	8,466
Multifamily	2.6	18.7	28.4	16.8	32.7	0.9	3,348
Other	38.5	16.7	7.7	7.6	1.4	28.1	1,962

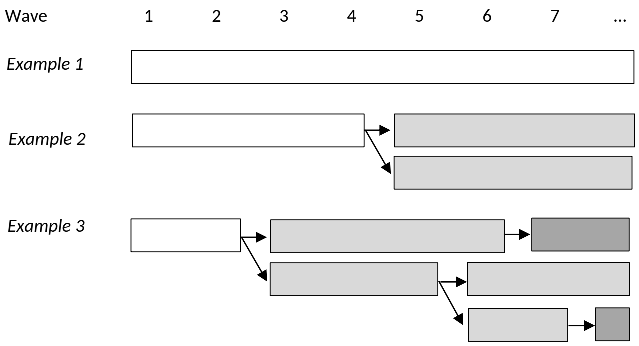

Figure 2. Simulating waves. *Note:* Shading represents a new imputed line of households drawn from the completed person-wave dataset.

plete information for the two waves) that is similar to the recipient in terms of observed characteristics (Andridge & Little, 2010). Imputation classes are formed based on the observed characteristics, providing a pool of donors from which one donor is selected randomly for each recipient in the same imputation class. The household in wave t (in this example, wave 4) is matched to a similar household in the person-wave file on a series of household characteristics defined at wave t along with an indicator of whether it makes a household composition change (imputed or observed) between wave t and $t+1$ and what type of household the household head at wave t belongs to at wave $t+1$. Example 3 in Figure 2 shows a household with an imputed transition between wave 2 and 3, then observed information is used for each of those imputed households until a new impute is needed. In the case of the first split a new imputation is needed in wave 7, and for the second split, another transition occurs between waves 5 and 6 but a new imputation is not needed until after wave 7.

The imputation classes used in the hot deck method are defined by a range of variables characterising the household at wave t, whether there is a change of household structure (a transition) or not in wave $t+1$ and if so, what type of family they transition into. The variables which define the imputation classes are in the following order: whether the household makes a transition between wave t and $t+1$, number of adults, number of children, whether a child is turning 15 (and therefore becoming eligible to be interviewed in wave $t+1$), 5-year age group of the household head, 5-year age group of the oldest person in the household, imputed family type in wave $t+1$ of the household head, relationship in household (couple, lone parent, child, other) of the household head, family type in wave t, whether the observation of the household is left truncated, the number of years the household was intact (i.e., with no leavers or joiners), the age of the youngest child, and the age of the oldest child. Where a donor could not be found, the last of these variables defining the imputation classes is dropped and an attempt to find a new donor is made. This is repeated until donors were found for all missing households. Table 4 provides the proportion of cases, averaged across waves 2 to 41, that are matched in each imputation pass. Once a household has been imputed, the subsequent observed household changes of the imputed household are followed through future waves until a point is reached where the next wave household structure is unknown and it is imputed again. The imputed household information is updated to ensure consistency with age and sex over time.

Deaths and overseas leavers are also identified as part of the simulation. Of the 19,914 wave 1 sample members, 35% had died by wave 41 and 12% had moved abroad. Moving abroad in this simulation is considered an absorbing state and 39% of these moves occurred by wave 11.

Applying the following rules. Once the simulated sample is constructed, the following rules are applied by tracking in each wave the original sample members, any babies born to CSMs (including adoptions), the other parent of those babies, and recent immigrants. Six sets of following rules are applied to simulate the rules of various studies:

SOEP CSMs are OSMs plus everybody who subsequently lives with a CSM.

HILDA1 CSMs are OSMs, babies born to CSMs, and other

Table 4
Hot deck match level to simulate households in wave 2–41

Variable	Imputation pass 1	2	3	4	5	6	7	8	9	10	11
Whether HH transition between wave t and $t+1$	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Number of adults	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Number of children	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Whether has child turning 15	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
HH head age group	Y[a]	Y[a]	Y[b]	Y[b]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]	
HH max age group	Y[a]	Y[a]	Y[b]	Y[b]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]		
HH head imputed family type in wave $t+1$	Y	Y	Y	Y	Y	Y	Y	Y			
HH head relationship in household	Y	Y	Y	Y	Y	Y	Y				
Family type in wave t	Y	Y	Y	Y	Y	Y					
Whether HH was left truncated	Y	Y	Y	Y	Y						
Years the HH was intact	Y	Y	Y	Y	Y						
Age of the youngest child	Y	Y	Y[d]	Y[d]							
Age of the oldest child	Y	Y[d]	Y[d]								
Percent of households matched at imputation pass	58	6	9	3	6	5	2	1	7	2	2

[a] 5-year age groups [b] 15-year age groups [c] age groups 0–14, 15–29, 30–59, 60 and over
[d] age groups 0–4, 5–9, 10–14

parent of CSM babies.

HILDA2 CSMs are OSMs, babies born to CSMs, other parent of CSM babies, and recent immigrants.

BHPS CSMs are OSMs and babies born to CSMs. Other parent of CSM babies are PSMs.

PSID CSMs are OSMs and babies born to CSMs.

UKHLS CSMs are OSMs and babies born to female CSMs. Other parent of CSM babies are PSMs.

Recent immigrants in the HILDA2-styled following rules are only added to the sample if they have arrived after the study began or each subsequent 10-year anniversary (i.e. wave 11, 21, 31). This specification of the rule assumes there would be an immigrant top-up selected from the population at regular 10-year intervals and that any converted immigrant CSM would continue to be followed even after an immigrant top-up has occurred. Including the recent immigrants that join sampled households is only a temporary measure until a more complete immigrant sample can be selected.

Simulating non-response. Sample growth will be limited by the nonresponse that occurs over time. To simulate the effect of this, a response model is developed for each of these seven groups of sample members:

1. Response at wave t given part of responding household at $t-1$ and:
   (a) response at $t-1$;
   (b) non-response at $t-1$;
   (c) child is turning 15 at wave t.
2. Response at wave t given part of responding household at $t-2$ (but not part of responding household at $t-1$) and:
   (a) response at $t-2$;
   (b) non-response at $t-2$;
   (c) child is turning 15 or 16 at wave t.
3. Response at wave t given adult is a new entrant at wave t.

For the scenarios 1a and 1b, a multilevel logistic regression model is used based on the HILDA Survey data. The model includes a random effect at the wave 1 household level, allowing for the multiple observations on individuals associated with the same wave 1 household over time. This allows for clustering within households and across individuals over time. A similar model is fitted for scenarios 2a and 2b together. For the remaining three scenarios, separate multilevel logistic regression models are fitted. These models include a random effect at the wave 1 household level, allowing for the clustering of multiple individuals associated with the same original household over time. These models are then used to predict the probability of response in the simulated dataset. The models include the following covariates: 5-year age groups (except for models 1c and 2c), female, health status, employment status (except for models 1c and

Table 5
Coefficients for response models at wave t

	Responding HH at $t-1$						Responding HH at $t-2$ (not at $t-1$)							
	Resp at $t-1$		Nonresp at $t-1$		Turn 15 at t		Resp at $t-2$		Nonresp at $t-2$		Turn 15, 16 at t		New entrant at t	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Age group (*Ref. categ.*: age 16–19)														
Age 20-24	−0.28***	0.05	−0.24**	0.11	-	-	−0.03	0.14	−0.49*	0.29	-	-	0.00	0.10
Age 25-29	−0.06	0.05	−0.61***	0.13	-	-	−0.12	0.15	−1.03***	0.37	-	-	−0.11	0.11
Age 30-34	0.17***	0.06	−0.56***	0.14	-	-	−0.05	0.16	−1.18***	0.42	-	-	−0.48***	0.12
Age 35-39	0.16***	0.05	−0.77***	0.15	-	-	−0.19	0.16	−0.96**	0.42	-	-	−0.19	0.14
Age 40-44	0.35***	0.05	−0.79***	0.14	-	-	−0.29*	0.15	−0.76*	0.39	-	-	−0.38***	0.14
Age 45-49	0.36***	0.05	−0.64***	0.13	-	-	−0.20	0.16	−1.31***	0.42	-	-	−0.53***	0.14
Age 50-54	0.46***	0.06	−0.91***	0.14	-	-	−0.47***	0.17	−1.55***	0.45	-	-	−0.50***	0.14
Age 55-59	0.62***	0.06	−0.77***	0.15	-	-	−0.51***	0.19	−1.47***	0.51	-	-	−0.58***	0.15
Age 60-64	0.82***	0.07	−1.06***	0.18	-	-	−0.15	0.21	−1.14**	0.57	-	-	−0.63***	0.18
Age 65-69	1.05***	0.08	−0.93***	0.22	-	-	−0.51*	0.26	−0.42	0.58	-	-	−0.44*	0.24
Age 70-74	0.69***	0.08	−1.15***	0.27	-	-	−0.29	0.25	−1.50*	0.86	-	-	−0.61**	0.27
Age 75-79	0.34***	0.09	−1.61***	0.30	-	-	−0.43*	0.25	−1.78*	0.93	-	-	−1.58***	0.31
Age 80 plus	−0.43***	0.09	−1.63***	0.25	-	-	−1.77***	0.26	−2.71**	1.20	-	-	−1.88***	0.24
Female	0.13***	0.02	0.24***	0.07	0.17*	0.10	0.15**	0.07	0.05	0.19	−0.17	0.25	0.32***	0.06
Has long term health condition	−0.05	0.03	−0.26***	0.10	−0.27	0.17	0.12	0.09	0.01	0.30	0.02	0.55	0.61***	0.09
Missing long term health condition	0.11	0.38	0.64**	0.30	-	-	1.81	1.38	0.11	1.21	-	-	−0.35	0.28
Employment status (*Ref. categ.*: employed full-time)														
Employed part-time	0.24***	0.03	0.12	0.09	-	-	−0.03	0.10	−0.29	0.29	-	-	0.37***	0.08
Unemployed	−0.12**	0.05	0.42***	0.16	-	-	0.02	0.15	0.50	0.40	-	-	0.76***	0.13
Not in the labour force	0.18***	0.04	0.12	0.09	-	-	0.02	0.10	−0.05	0.26	-	-	0.18**	0.08
Missing employment status	0.33	0.39	−1.05***	0.29	-	-	−0.57	1.15	−1.60*	0.87	-	-	−3.17***	0.23
Number of children (*Ref. categ.*: no children)														
One child	0.1***	0.04	0.22**	0.09	−0.16	0.20	0.14	0.11	−0.09	0.27	−0.74*	0.39	0.22***	0.09
Two children	0.24***	0.04	0.42***	0.11	−0.04	0.20	0.19	0.12	0.12	0.32	−0.93**	0.40	0.23**	0.11
Three or more children	0.35***	0.07	0.50***	0.17	−0.03	0.22	0.16	0.19	0.07	0.49	−0.60	0.44	0.04	0.18
Number of adults (*Ref. categ.*: two adults)														
One adult	0.04	0.06	-	0.06	−0.62***	0.16	0.37***	0.10	-	0.10	0.31	0.35	-	-
Three	−0.18***	0.06	0.27*	0.15	0.19	0.13	0.05	0.11	0.14	0.25	0.19	0.31	−0.43***	0.07
Four or more adults	−0.32***	0.07	0.13	0.16	−0.10	0.17	−0.42***	0.13	0.04	0.25	−0.41	0.42	−0.72***	0.08
HH moved	−0.15***	0.03	−0.25***	0.09	0.05	0.18	0.25***	0.10	−0.34	0.29	0.12	0.38	−0.27***	0.06
HH splits subsequently	0.07*	0.04	0.59***	0.11	−0.26	0.17	0.54***	0.11	0.87***	0.31	0.58*	0.32	0.08	0.06
HH has 2-wave non-resp.	−0.55***	0.05	−1.35***	0.08	−2.71***	0.15	−0.21*	0.13	−0.53**	0.25	−0.80*	0.45	−1.83***	0.23
HH has 2-wave non-resp. & splits subseq.	1.03***	0.12	0.27*	0.15	0.61*	0.37	−0.28	0.40	−0.58	0.41	0.58	0.98	0.65**	0.31
Wave *t* (*Ref. categ.*: wave 13 or higher)														
Wave 2	−0.47***	0.06	0.50***	0.19	−1.24***	0.18	-	-	-	-	-	-	0.35***	0.12
Wave 3	−0.38***	0.06	1.22***	0.19	−0.92***	0.20	0.00	0.13	−0.66*	0.37	−0.05	0.43	−0.17	0.11
Wave 4	−0.45***	0.06	0.86***	0.22	−0.97***	0.19	0.33**	0.13	0.61*	0.36	0.12	0.45	−0.50***	0.11
Waves 5-8	−0.13***	0.05	0.84***	0.13	−0.08	0.15	0.40***	0.10	0.81***	0.25	0.65*	0.36	0.09	0.08
Waves 9-12	0.02	0.05	0.67***	0.13	0.23	0.16	0.22**	0.11	0.57**	0.27	0.19	0.42	0.07	0.08

Continues on next page

Continues from previous page

	Responding HH at $t-1$						Responding HH at $t-2$ (not at $t-1$)							
	Resp at $t-1$		Nonresp at $t-1$		Turn 15 at t		Resp at $t-2$		Nonresp at $t-2$		Turn 15, 16 at t		New entrant at t	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Number of adults by wave (*Ref. categ.:* two adults, wave 13 or higher)														
One adult, wave 2	0.39***	0.12	0.39***	0.12	-	-	-	-	-	-	-	-	-	-
One adult, wave 3	0.11	0.12	0.11	0.12	-	-	-	-	-	-	-	-	-	-
One adult, wave 4	0.37***	0.13	0.37***	0.13	-	-	-	-	-	-	-	-	-	-
One adult, wave 5-8	0.24***	0.09	0.24***	0.09	-	-	-	-	-	-	-	-	-	-
One adult, wave 9-12	0.14	0.09	0.14	0.09	-	-	-	-	-	-	-	-	-	-
Three adults, wave 2	−0.35***	0.11	−0.87***	0.29	-	-	-	-	-	-	-	-	-	-
Three adults, wave 3	−0.54***	0.12	−0.32	0.29	-	-	-	-	-	-	-	-	-	-
Three adults, wave 4	−0.42***	0.12	−0.24	0.32	-	-	-	-	-	-	-	-	-	-
Three adults, wave 5-8	−0.09	0.09	−0.09	0.20	-	-	-	-	-	-	-	-	-	-
Three adults, wave 9-12	0.11	0.09	0.06	0.21	-	-	-	-	-	-	-	-	-	-
Four or more adults, wave 2	−0.32***	0.12	−0.33	0.27	-	-	-	-	-	-	-	-	-	-
Four or more adults, wave 3	−0.52***	0.12	−0.60**	0.29	-	-	-	-	-	-	-	-	-	-
Four or more adults, wave 4	−0.52***	0.13	−0.65*	0.33	-	-	-	-	-	-	-	-	-	-
Four or more adults, wave 5-8	−0.34***	0.09	−0.31	0.21	-	-	-	-	-	-	-	-	-	-
Four or more adults, wave 9-12	0.05	0.10	0.24	0.21	-	-	-	-	-	-	-	-	-	-
Constant	2.94***	0.06	−0.97***	0.15	2.76***	0.22	−1.80***	0.17	−2.69***	0.38	−0.79*	0.48	1.69***	0.12

* $p < 0.1$ ** $p < 0.05$ **** $p < 0.01$

2c), number of children, number of adults, household moved (at time of last response), household split after last response, whether the household contains a persistent non-respondent (of at least two waves), indicator variables for wave 2, 3, 4, 5–8 and 9–12, and (for the model for scenario 1a and 1b) the interaction of wave and number of adults. These covariates are limited to the information available about all members of the household in this simulation (it does not include detailed information from a previous interview as is done, for example, by Lepkowski & Couper, 2002; Uhrig, 2008; Watson & Wooden, 2009) as this information is not available for all individuals). The coefficients for the four models are presented in Table 5. Note that for presentational purposes, the coefficients for respondents presented here are the linear combination of the coefficient of the particular variable for non-respondents (the base category) and the interaction of response (at wave $t-1$ or $t-2$) with the particular variable. These models are used to predict a propensity to respond at wave t for individuals in the simulated sample, depending on which scenario they fall into. The predicted fixed and random components are both used to calculate the response propensity. Households are assigned a random number each wave from the uniform distribution from 0 to 1 and all individuals receive this random number, including new entrants. Individuals are imputed to be a respondent (i.e. provide an individual interview) at wave t if the random number is below the predicted probability of response and not otherwise. This is done wave by wave, so that past simulated response behaviour can be incorporated into the prediction of response at wave t.

If a household has at least one person classified as a respondent, then all members of the household are considered to be enumerated (i.e. part of a responding household). If a household is non-responding for two sequential waves then, by the design of the response models presented earlier, they are deemed to be non-responding in all future waves along with any new entrants the household may have. Further, children turning 15 or new entrants in a household where no other sample members respond are also set as non-respondents the first two waves they are eligible to be interviewed.

The actual HILDA Survey sample at wave 16 (in terms of sample size, sample characteristics and response rates) matches reasonably closely to the simulated wave 16 sample after applying the HILDA1-styled following rules and non-response. While it is not a perfect replica, the simulation sample provides a consistent basis on which to compare the impact of alternative following rules.

3 Results

3.1 Size of the sample

The sample begins in wave 1 with 19,914 sample members and by wave 41, the simulated sample contains 3.9 million sample members. This excludes 83,200 sample members who have died and 131,700 sample members who moved overseas by that time. This is the size of the sample with SOEP-styled following rules where everyone is followed and makes the unrealistic assumption that everyone who is eligible to be interviewed responds. Narrower following rules result in sample sizes between 36,100 (under the UKHLS-styled following rules) to 54,000 (under the HILDA2-styled following rules), again under the unrealistic assumption of full response (see Figure A1 in the appendix).

When the HILDA-styled non-response models are applied to the sample, the sample sizes are much reduced. Figure 3 shows how the sample size changes across the 41 waves when different following rules and non-response apply. Compared to the sample with full response, the sample with SOEP-styled following rules is reduced by 85% at wave 41 and the samples with narrower following rules are reduced by 60 to 62%. The greater reduction with the SOEP-styled following rules is due to the larger amount of growth occurring in later waves. Nevertheless, even with HILDA-styled non-response applied, the simulated sample with SOEP-styled following rules and HILDA-styled non-response is still exceptionally large.

With non-response applied equally to all scenarios, differences in overall sample size due to the following rules over the long-term are very apparent. Compared to following births of all CSMs (as in the BHPS-styled following rules) with a wave 41 sample size of 17,800, only adding the births of female CSMs (in the UKHLS-styled following rules) saves 19% in the size of the sample. Following only the children of CSMs (as per the PSID-styled following rules) and not adding the other parent of a birth as a PSM (as is done in the BHPS-styled following rules) saves 8% in the size of the sample. Adding the other parent of a birth and following them like any other CSM (as per the HILDA1-styled sample), rather than following only that person (i.e. as a PSM) as per the BHPS-style following rules, adds 10% of the size of the sample at wave 41. Adding recent immigrants as CSMs (as per the HILDA2-styled sample) adds 4% to the size of the HILDA1-styled sample. Similar sample size differentials are evident in the sample with full response.

The growth in the sample from the BHPS-styled following rules matches the loss from attrition resulting in a relatively steady sample size in the long-run. The PSID- and UKHLS-styled following rules do not add enough permanent sample members to keep pace with the sample loss from the HILDA-styled non-response, but the sample loss with the PSID-styled following rules is less than what occurs with the

UKHLS-styled following rules. The two HILDA-styled following rules result in sample additions outpacing losses from attrition so the overall active sample grows.

Of course, different levels of non-response or characteristics of non-response will result in different sized samples over the long-run but this simulation illustrates the two opposing forces (i.e., following rules and attrition) on the overall sample size.

3.2 Composition of the sample

The following analysis of the composition of the sample focuses on two particular waves: wave 16 and 41. Wave 16 is chosen as it is within the current window of the HILDA Survey data. It is also the first wave in which all of the children observed in the initial sample are of an age to be interviewed and all of the children in wave 16 have been born into the sample after the initial wave. Wave 41 is chosen as the last wave within the simulated sample so as to allow time for the children born into the sample to subsequently leave home. Key socio-demographic variables often used by researchers are compared across the different sets of following rules. Also compared is the distribution of the number of times sample members respond in each simulated sample.

There are substantial differences in sample composition between the sample with SOEP-styled following rules and the samples with narrower following rules. Figure 4 shows the age distribution of each of the samples at wave 16 (on the left) and wave 41 (on the right) with non-response being applied. The age distribution for wave 1 is added to each of these graphs as a comparison. Note that none of the samples in wave 16 or 41 resemble the wave 1 age distribution. This is because of the (temporary or permanent) addition to the sample of people who subsequently live with the OSMs. By wave 16, the age profile of the sample with the SOEP-styled following rules, compared to the other samples, includes proportionally many more people in their 20s and 30s, an associated increase in the proportion of children aged 0 to 9 and a smaller proportion of people aged 40 and older. Some small differences in the age distribution are beginning to appear for the narrower following rules in the 0 to 9 year old children but there are few differences in the other parts of the age distribution. These differences between the sample with the SOEP-style following rules and the other samples are due to the higher rate of household structure changes that occur for people in their 20s and 30s (e.g. moving out of home, partnering, sharing with housemates, etc). By wave 41, the differences in the age profile of the sample with the SOEP-styled following rules and the narrower following rules become somewhat more pronounced. More interestingly, the UKHLS-style following rules that limits the births that are followed display a different age profile in the sample to the other following rules where all births of CSMs are followed. The differences primarily occur in the 20–34 age group, which is when the births that occurred in the first 15 years of the panel would be leaving home. Similar differences are evident in the full sample (see Figure A2 in the appendix) though without the simulated non-response there is a slightly higher proportion of children and people in their 20s and 30s and a slightly lower proportion of people aged 40 and over.

The left graph of Figure 5 shows more people in the responding SOEP-style sample have moved in the last 12 months in wave 16 than under the other following rules. This is because the SOEP-style following rules bring into the sample people who are more likely to change household structure (which is also often accompanied with a household move). By wave 41 there are some differences emerging between the UKHLS-style following rules and the other narrow following rules. A very similar pattern is seen in the entire simulated sample (including people in responding and non-responding households), though the mobility rates are slightly higher for all samples (see Figure A3 in the appendix).

Turning now to other characteristics of the responding sample under the different following rules, Figure 6 presents various characteristics of the sample in wave 16. The first two graphs show the average number of children and adults per household. All samples have a similar number of adults per household. The SOEP-style sample has, on average, more children per household, reflecting the younger nature of the sample. The BHPS, HILDA, and PSID following rules produce samples that are all very similar with respect to the number of children, however the UKHLS sample has less children per household than the HILDA samples. The next two graphs in Figure 6 show the proportion in full-time employment and the proportion of people with a long-term health condition. The SOEP-style sample has more people in full-time employment and fewer people with a long-term health condition than the other samples, owing to the greater proportion of people in their 20s and 30s in the sample. The next three graphs in Figure 6 examine the differences in household type by showing the proportion of sample members who are part of a couple, a lone parent, or live alone. The SOEP-style sample has fewer couples and single person households and more lone parents than the other samples. The last graph in Figure 6 shows the number of years the households have been intact (i.e., without any structural changes). The SOEP-style samples average 3.5 intact years, whereas the other following rules average 6 intact years.

The differences seen in the characteristics of the samples are more pronounced at wave 41, as shown in Figure 7. The SOEP-style sample continues to be quite different from the other samples, having on average a greater number children per household and fewer adults per household and years intact. The SOEP-style sample also has a greater proportion of people working full-time but a smaller proportion of people with a long-term health condition, couples, and lone per-

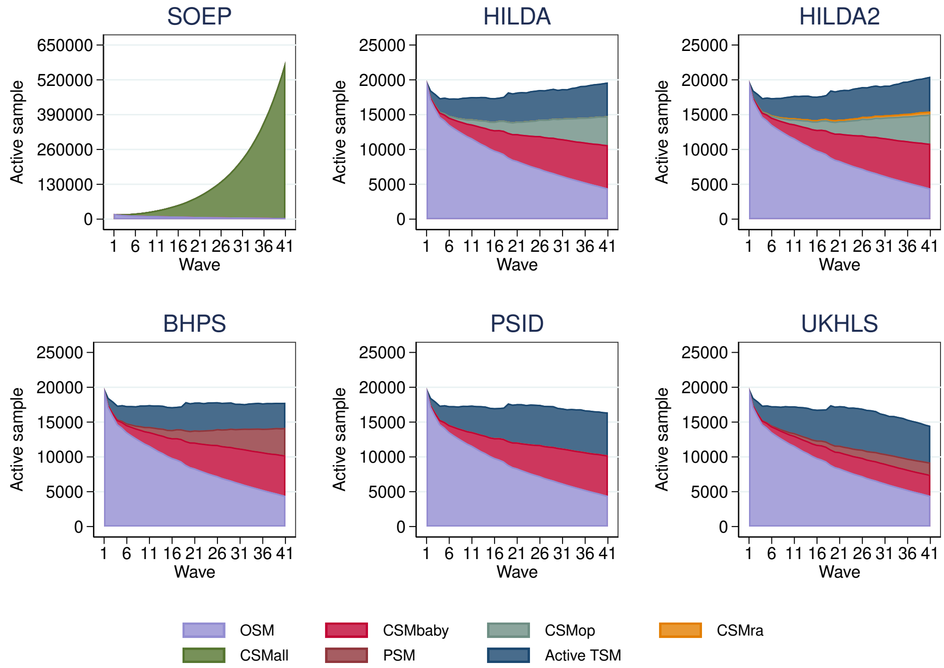

Figure 3. Sample size for waves 1 to 41 for different following rules with non-response applied. *Notes:* OSM: original sample member (a CSM); CSM baby: baby born to a CSM; CSMop: other parent of a CSM; CSMra: recent arrival meeting CSM rules; CSMall: any co-resident of CSM; PSM: other parent of a CSM baby; active TSM: temporary sample member co-resident with a CSM.

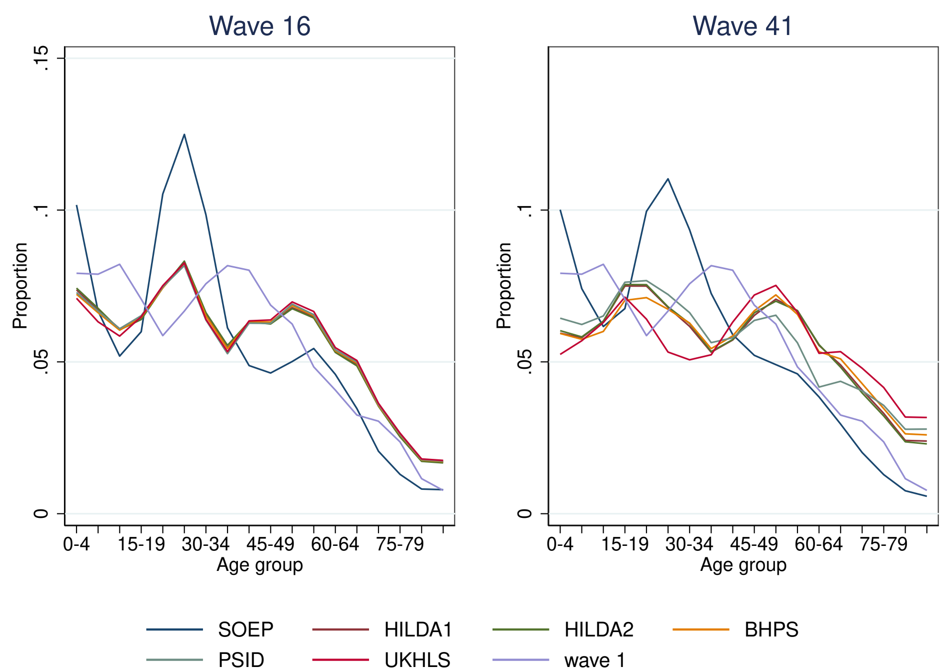

Figure 4. Age distribution of the simulated samples at wave 16 and 41 for the various following rules, non-responding households excluded

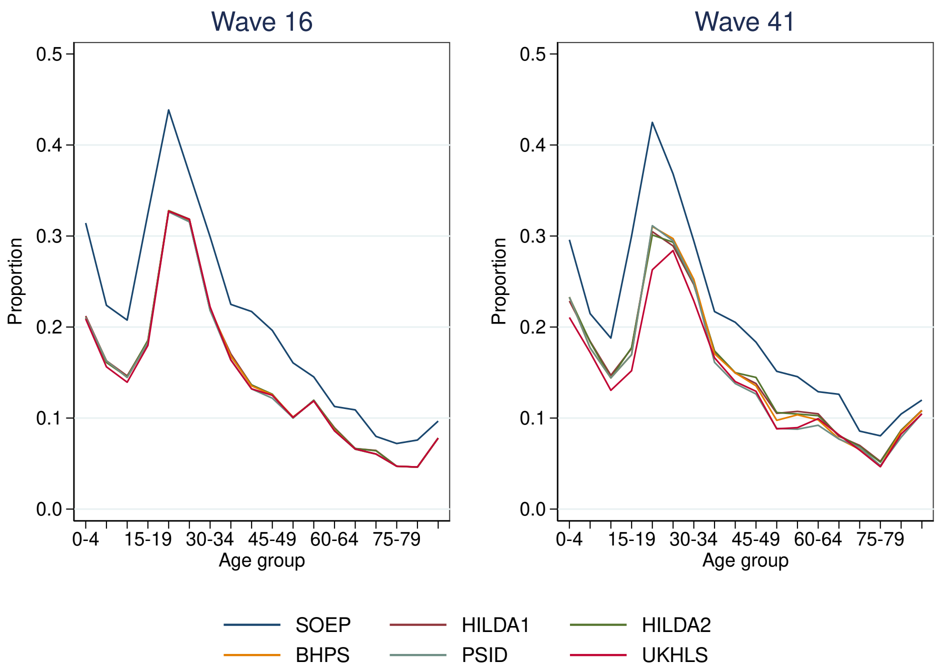

Figure 5. Proportion of people moving in the last 12 months in wave 16 and 41, by age group (non-responding households excluded)

sons. It also has a greater proportion of lone parents than other samples (with the exception of the sample with BHPS-style following rules). The UKHLS-style sample has generally a greater proportion of people with a long-term health condition and a larger number of years intact than the other remaining samples, this being due to the sample being older on average. The PSID-style sample is also a little different from the other samples with narrow following rules, with more children per household on average. This is likely due to not following the other parent of births if they leave the household of a PSM.

Finally, Figure 8 shows the distribution of the number of times sample members respond within the first 16 waves and then across all 41 waves (after the non-response models have been applied). This shows the composition of cases available for longitudinal analysis under the different following rules. The SOEP-styled following rules have a greater proportion of people observed for a shorter amount of time than the other following rules. There is very little difference between the other following rules in terms of the number of times people respond.

4 Conclusion

This simulation of a household panel sample which follows everyone who joins a CSM household provides a platform from which it is possible to study the effect of alternate following rules on the size and composition of the sample. The use of the HILDA Survey data to underpin the simulation which has reasonably wide following rules (without following everybody) and high response rates means the amount of imputed data required due to nonresponse and TSM leavers is relatively modest. Further, being able to assess the difference in the size and composition of the sample with and without non-response applied in a consistent fashion (albeit using a limited range of variables) across the samples allows some further insights. A total of 41 waves are simulated, allowing for the effect of the following rules to be studied over a 40-year window. This allows time for children to be born into the sample and subsequently leave home.

The SOEP-style following rules produces a sample that is quite different from the samples with narrower following rules. The simulated SOEP-styled sample is much larger (the responding sample is more than 30 times larger by wave 41), has a greater proportion of people in their 20s and 30s, has on average more children in each household, a lower proportion of people with a long term health condition, a lower proportion of couples and single person households and a higher proportion of lone parent households, and the households are on average intact for a smaller number of years. The BHPS- and HILDA-style following rules (all of which follow all births to CSMs) produce samples that have relatively similar sample characteristics. The UKHLS-style following rules (which follow a smaller sample of births to CSMs) dif-

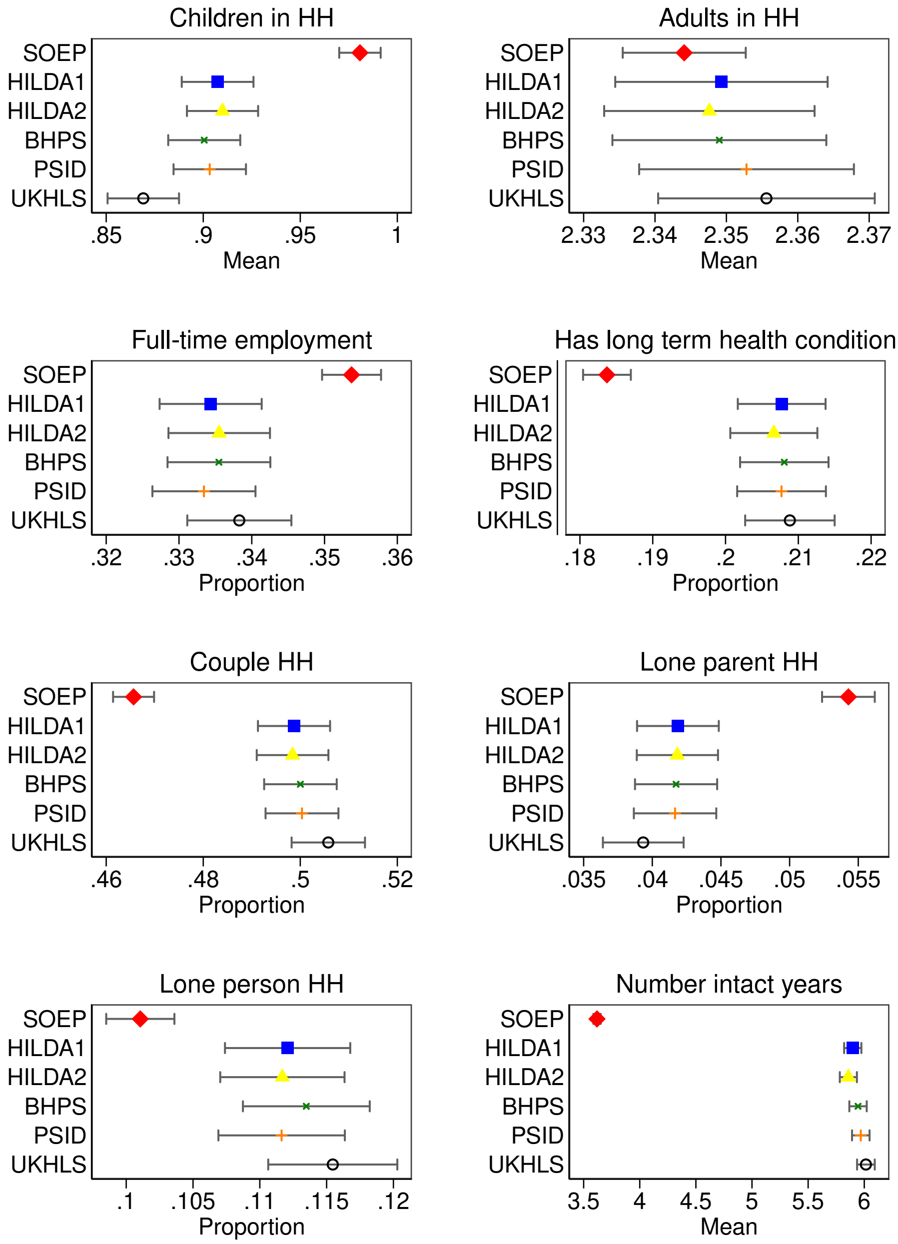

Figure 6. Characteristics of the sample in wave 16 for various following rules (non-responding households excluded)

fer somewhat in sample characteristics to the samples following all births (but not everybody): this sample tends to move less frequently, have a higher proportion of people with a long-term health condition, and stay intact for longer than the PSID-, BHPS- and HILDA-style samples. The PSID-styled following rules result in a sample that tends to have more children on average per household and fewer lone person households.

An obvious difference between the simulated results here and the actual sample size numbers achieved in SOEP and UKHLS in particular is due to the much higher nonresponse rates in these studies compared to the HILDA Survey that has been used to generate the nonresponse models (Watson et al., 2018). Conversely, the PSID has higher response rates than the HILDA Survey, so rather than see the sample diminish over time as simulated here, their sample has grown. Higher non-response rates, as seen more recently in the PSID (Institute for Social Research, 2019) will slow or eliminate this

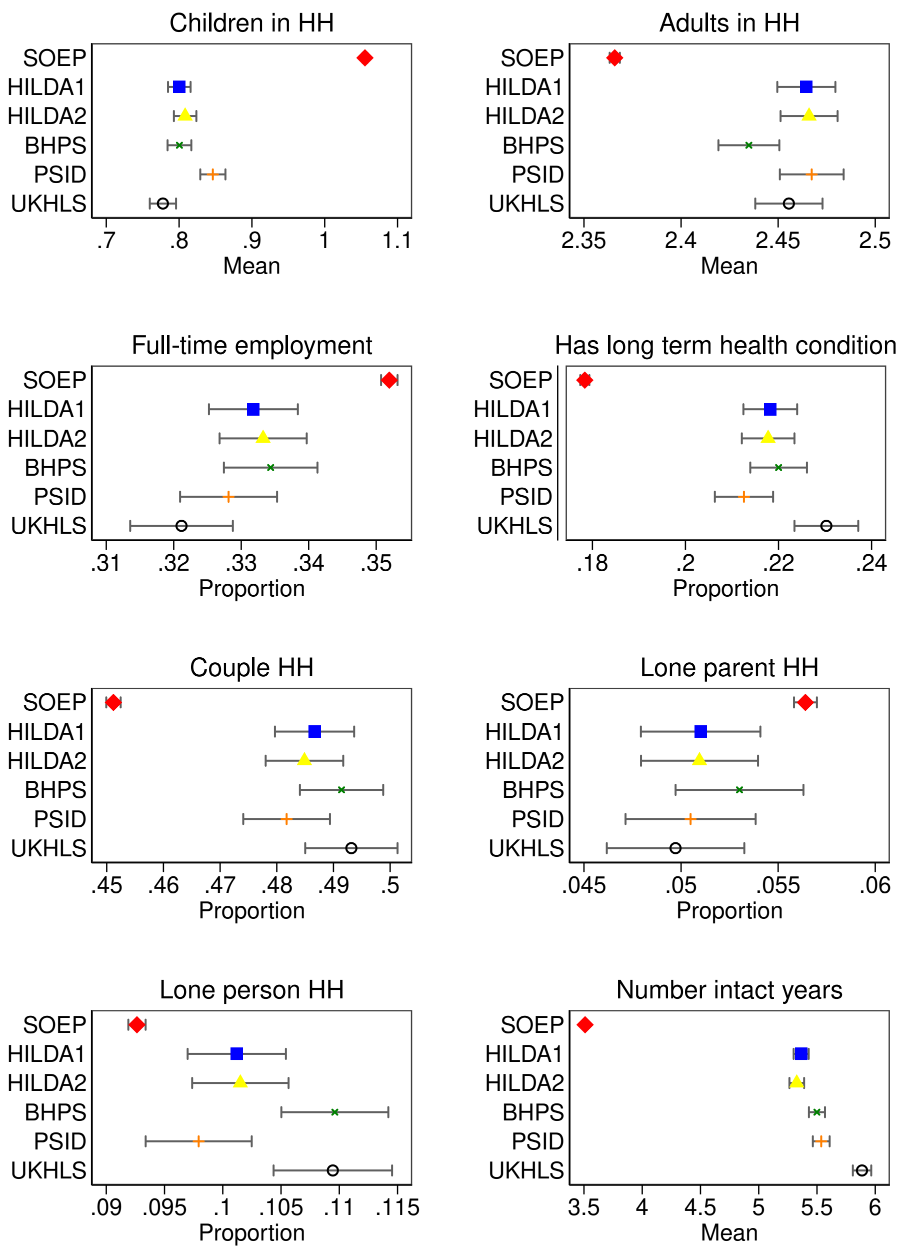

Figure 7. Characteristics of the sample in wave 41 for various following rules (non-responding households excluded)

sample growth.

These results indicate that the following rules chosen at the outset of conducting a household panel survey have long term consequences on the size and composition of the sample. This is important for survey managers in understanding the likely cost differentials of certain decisions around the following rules. Although changes can be made to these following rules in later waves it will not be possible to re-engage with sample members that have left the sample (i.e., become inactive TSMs) if the following rules are widened at a later point in time. And if the following rules are narrowed later then it means that following and interviewing those sample members that are later thought excessive has added unnecessarily to the cost of the study.

These results are also relevant to the users of household panel survey data. The unbalanced panel is not equivalent to a series of cross-sections or cohorts added together. There are additions to the sample due to the practice of interview-

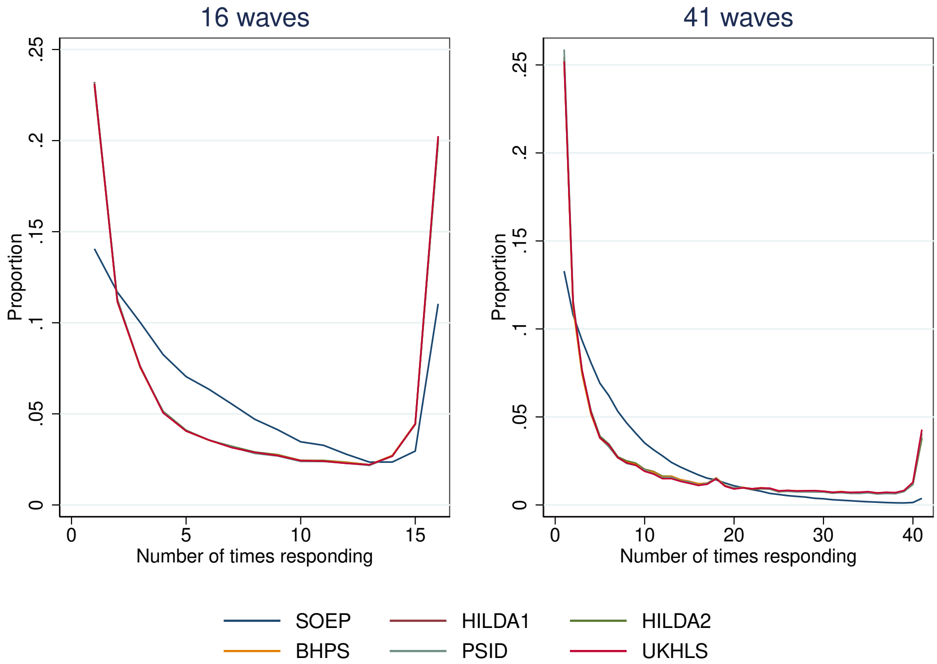

Figure 8. Distribution of the number of times sample members respond after 16 and 41 waves

ing all people the CSM shares a house with as well as who is converted to a CSM and followed should they leave the household of a CSM. These additional people alter the size and composition of the sample and this needs to be taken into account in some way when analysing the data. Or at least undertaken as part of the sensitivity analysis. One way to take these differences into account is through the restriction of the sample to the OSMs and (some or all of) their descendants. Ideally, the choice to include children in the analysis should be proportional to the number of OSM parents that they have. An alternative is to use weights in the analysis which already take into account these differences in probabilities of inclusion in the sample. Understandably, the use of weights may be a concern to the analyst as it will result in restrictions to the sample. One potential option that needs some further research could be to use the design weight if this is available post wave 1. The author is aware of only one study, the UKHLS, which provides such a weight.

It is also worth noting that studies such as the SOEP have supplemented their original sample with a number of additional samples. The impact of such additions could be explored by combining the simulated sample from different waves together. For example, if a sample addition of the size of the original sample was added in wave 11 of the survey, then at wave 16 this would be equivalent to adding the wave 6 sample for the top-up to the original sample at wave 16. Alternatively, if the top-up was focused on households with particular characteristics, then a subset of such households could be selected from the wave 1 sample and subsequent households in later waves linked to those original households could be added to the appropriate wave of the original sample. Such explorations of the size and composition of multiple combined samples is beyond the scope of this paper.

This simulation has a number of limitations. The first limitation is that re-joiners to a household (i.e. someone who left a sampled household and returns at a later wave) are treated as the same as a new entrant. Sample members who leave a household are treated as separate lines followed through time (as illustrated in Figure 2) and it would be very difficult to recross those lines when a re-joiner to a household occurs, certainly when the two lines have used different donors in the intervening waves and could be at different stages of development. Re-joiners make up 13% of the household joiners on average across waves 3 to 18 of the HILDA Survey. Nevertheless, treating re-joiners as new entrants will increase the size of the simulated sample somewhat. The second limitation is that there is a small amount of leakage in the simulation system, resulting in 0.05% of the sample each wave not matching to a suitable donor. It is not expected that this would materially affect the results of the simulation. The third limitation is that the household transition, compositional change and response models are limited to a certain set of variables and as a result the simulated processes are based only on these variables. For non-response in particular, this means that nonresponse is simulated via a missing at random process specific to this set of variables and that other variables—such as household structural changes (known only for responding households)—which may be associated with non-response are not able to be included. The fourth limitation is that the fertility rates, mortality rates, emigration and immigration rates applied in this simulation study are those from 2002 to 2018 in Australia based on the HILDA Survey data. Further, it is assumed that these rates are relatively steady over the 40-year window of the simulation. Equivalent simulations using data from other countries may produce different results simply because of these differences in demographic statistics, but the broad findings are expected to be similar.

Acknowledgement

This paper uses data from the Household, Income and Labour Dynamics in Australia Survey which was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the author and should not be attributed to the Australian Government, DSS or the Melbourne Institute. The data are available through Dataverse at the Australian Data Archive (dataverse.ada.edu.au). The author also appreciates the helpful feedback provided by John Henstridge, Michele Haynes, Janeen Baxter, and the anonymous referees on earlier versions of this paper.

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1). doi:10.1111/j.1751-5823.2010.00103.x

Department of Social Services, & Melbourne Institute of Applied Economic and Social Research. (2019). The Household, Income and Labour Dynamics in Australia (HILDA) survey, general release 18 (waves 1–18). ADA Dataverse, V5. doi:10.26193/IYBXHM

Heeringa, S. G., Berglund, P. A., Khan, A., Lee, S., & Gouskova, E. (2011). PSID cross-sectional individual weights, 1997–2009. Retrieved from https://psidonline.isr.umich.edu/data/weights/cross_sec_weights_09.pdf

Institute for Social and Economic Research. (2019). Understanding society: The UK Household Longitudinal Study Waves 1–9 user guide. Retrieved from https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/mainstage/user-guides/mainstage-user-guide.pdf

Institute for Social Research. (2019). PSID main interview user manual: Release 2019. Retrieved from https://psidonline.isr.umich.edu/data/Documentation/UserGuide2017.pdf

Laurie, H., Smith, R. A., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, *15*(2).

Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, *21*(1).

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey nonresponse*. New York: Wiley.

Lynn, P. (2006). Quality profile: British Household Panel Survey: Version 2.0, waves 1–13. Retrieved from https://www.iser.essex.ac.uk/files/bhps/quality-profiles/BHPS-QP-01-03-06-v2.pdf

Mutchier, J. E., & Burr, J. A. (1991). A longitudinal analysis of household and nonhousehold living arrangements in later life. *Demography*, *28*(3).

Qu, L., & De Vaus, D. A. (2011). Starting and ending one-person households: A longitudinal analysis. *Journal of Family Studies*, *17*(2).

Richards, T., White, M. J., & Tsui, A. O. (1987). Changing living arrangements: A hazard model of transitions among household types. *Demography*, *24*(1).

Schonlau, M., Kroh, M., & Watson, N. (2013). The implementation of cross-sectional weights in household panel surveys. *Statistics Surveys*, *7*.

Schonlau, M., Watson, N., & Kroh, M. (2011). Household survey panels: How much do following rules affect sample size? *Survey Research Methods*, *5*(2).

Smith, P., Lynn, P., & Elliot, D. (2009). Sample design for longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys*. Chichester: Wiley.

Summerfield, M., Bright, S., Hahn, M., La, N., Macalalad, N., Watson, N., … Wooden, M. (2019). HILDA user manual—release 18. Retrieved from https://melbourneinstitute.unimelb.edu.au/__data/assets/pdf_file/0008/3247289/HILDA-User-Manual-Release-18.0.pdf

Taylor, M. F. (2017). British Household Panel Survey user manual volume a: Introduction, technical report and appendices. Retrieved from http://doc.ukdataservice.ac.uk/doc/6614/mrdoc/pdf/6614_bhps_user_manual_volume_a.pdf

Uhrig, S. C. N. (2008). The nature and causes of attrition in the British Household Panel Study. Retrieved from https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2008-05.pdf

Watson, N. (2006). Options for a top-up sample to the HILDA survey. Paper presented at the ACSPRI Social Science Methodology Conference, Sydney. Retrieved from https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-discussion-papers/Watson_Topup.pdf

Watson, N. (2012). Longitudinal and cross-sectional weighting methodology for the HILDA survey. Retrieved from https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec212.pdf

Watson, N., Leissou, E., Guyer, H., & Wooden, M. (2018). Best practices for panel maintenance and retention. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3mc)*. Ney York: Wiley.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys*. Chichester: Wiley.

Appendix

Details of the simulation

Imputing broad family type for partial household splits

The broad family type into which each household split transitions in wave $t + 1$, conditional on a transition having occurred, is imputed if it is not already known (as in example 1 in Figure 1). The process to impute this is similar to that described in Section 2.3 for imputing the broad family type of the household head. A multinomial logit model for each family type in wave t is estimated based on observed transitions where broad family type is known in wave $t + 1$ (models not shown). The head of each part household split is included in the model. The person defined as the head of the part split is in order of priority part of a couple, a lone parent, or other household member. Where more than one person fulfils the requirement for part split head, the older person is chosen. The characteristics of the household included in the model are whether the period the household has been intact is left truncated, the number of intact years, whether the household contains other relatives or unrelated persons. The characteristics of the part household head include 5-year age dummy variables, employment status, and health status. Six predicted probabilities are estimated for each head of a household split, one for each broad family type they could transition into in wave $t + 1$. All heads of a split household missing the family type in wave $t + 1$ are randomly assigned to one of the six family types based on the predicted probabilities.

Imputing household structure for partial household splits

For the 4% of person-wave observations where it is known a transition occurred between t and $t+1$ but the household was not observed in wave $t + 1$ (as in Example 2 in Figure 1), hot deck imputation is used. Donors are randomly matched to recipients within imputation classes. The imputation classes are defined by a range of variables based on the information specific to the portion of the household at wave t that splits out at wave $t+1$ along with the imputed family type for wave $t + 1$. The variables used to define the imputation classes are in the following order: number of adults, number of children, whether a child was turning 15 (and therefore becoming eligible to be interviewed in wave $t + 1$), 5-year age group of the head of the part split, 5-year age group of the oldest person in the part split, imputed family type in wave $t + 1$, relationship in household (couple, lone parent, child, other) of the head of the part split, family type in wave t, whether the observation of the household was left truncated, the number of years the household was intact, the age of the youngest child, and the age of the oldest child. Where a donor could not be found, the last of these variables defining the imputation classes was dropped and an attempt to find a new donor is made. This is repeated until donors were found for all missing household splits. Table A1 shows the number of donors found at each level of matching. The donor could come from any of the first 17 waves of the sample. New household identifiers are assigned to these imputed part household splits, the longitudinal individual identifiers of the recipients are carried forward into the new household and new household members are identified (if any). All household information (such as age, sex, household relationships, employment status, health status) is taken from the donor and updated to ensure age and sex is consistent with the recipient's previous wave record.

Imputing family type household head transitioned into

Where a transition in household structure (i.e., there was a joiner or leaver or both) occurs between wave t and $t + 1$, broad family type of the household head at wave $t + 1$ is imputed where it is unknown. Table A2 shows the multinomial model of broad family type (in 6 categories) in wave $t + 1$ fitted for each of the broad family types at wave t.

Selected results with full response

Figure A1 shows the growth in the sample before non-response has been applied. Figures A2 and A3 show the age distribution and the proportion who move house in these samples at waves 16 and 41.

Table A1
Hot deck match level to complete part household splits

Variable	Imputation pass 1	2	3	4	5	6	7	8	9	10	11	12
Number of adults	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Number of children	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Whether has child turning 15	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Part split head age group	Y[a]	Y[a]	Y[b]	Y[b]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]		
Part split max age group	Y[a]	Y[a]	Y[b]	Y[b]	Y[c]	Y[c]	Y[c]	Y[c]	Y[c]			
Part split head imputed family type in wave $t + 1$	Y	Y	Y	Y	Y	Y	Y	Y				
Part split head relationship in household	Y	Y	Y	Y	Y	Y	Y					
Family type in wave t	Y	Y	Y	Y	Y	Y						
Whether HH was left truncated	Y	Y	Y	Y	Y							
Years the HH was intact	Y	Y	Y	Y	Y							
Age of the youngest child	Y	Y	Y[d]	Y[d]								
Age of the oldest child	Y	Y[d]	Y[d]									
Percent of part household splits matched at imputation pass	62	2	7	1	4	6	4	2	11	0	0	0

[a] 5-year age groups [b] 15-year age group [c] age groups 0–14, 15–29, 30–59, 60 and over [d] age groups 0–4, 5–9, 10–14

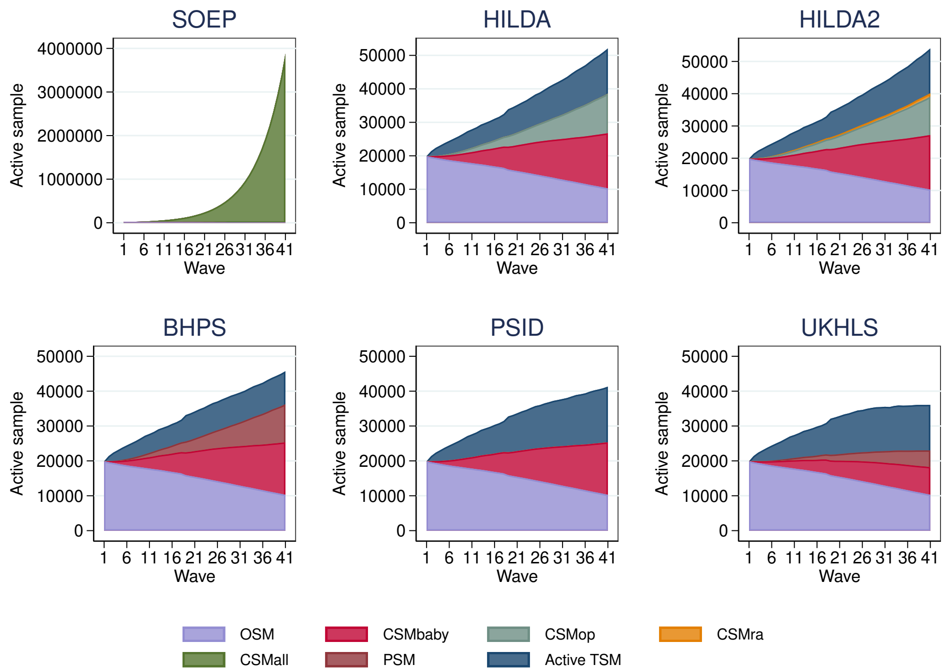

Figure A1. Sample size for waves 1 to 41 for different following rules with full response. *Notes:* OSM=original sample member (a CSM); CSM baby=baby born to a CSM; CSMop=other parent of a CSM; CSMra=recent arrival meeting CSM rules; CSMall=any co-resident of CSM; PSM=other parent of a CSM baby; active TSM=temporary sample member co-resident with a CSM.

Table A2

Multinomial model of broad family type at wave $t + 1$ given transition between t and $t + 1$

	Single person		Couple w/o child.		Couple w. child.		Lone parent		Multifamily HH		Other	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Transition to couple without children at wave t + 1 (base outcome)												
Transition to single person HH at wave t + 1												
Years intact is left truncated	-	-	0.35	0.42	0.13	0.33	0.65	0.62	1.53	1.48	−0.07	0.61
Years HH intact (ref. categ.: one year)												
2 years	-	-	−0.41**	0.20	−0.58***	0.18	0.15	0.31	−1.69	1.07	−0.06	0.27
3 years	-	-	−0.35	0.25	−0.52**	0.21	−0.47	0.34	0.19	1.12	−0.29	0.42
4 years	-	-	−0.24	0.31	0.28	0.23	0.47	0.55	1.03	1.21	−1.22**	0.57
5 to 9 years	-	-	−0.05	0.25	0.34*	0.20	−0.20	0.32	−17.66	-[a]	−0.27	0.70
10 years or more	-	-	0.20	0.39	0.62**	0.31	0.68	0.64	−0.70	-[a]	−4.98***	1.92
Has other related	-	-	−4.05***	0.34	0.47	0.42	−1.52***	0.45	−17.99	-[a]	−0.65	0.48
Has other not related	-	-	−3.40***	0.35	0.14	0.73	−1.88***	0.56	1.97*	1.17	−0.21	0.51
Age HH head (ref. categ.: 25-29 years)												
Age 15-19	-	-	−0.16	0.52	-	-	−16.06	-[a]	-	-	0.80	0.59
Age 20-24	-	-	0.01	0.23	-	-	−18.10	-[a]	-	-	−0.13	0.26
Age 30-34	-	-	0.29	0.26	−1.93*	1.07	−0.05	-[a]	−0.28	-[a]	0.21	0.34
Age 35-39	-	-	0.40	0.36	−2.09*	1.07	−15.51	-[a]	−18.55	-[a]	0.68	0.49
Age 40-44	-	-	0.65*	0.37	−3.81***	1.03	−16.41	-[a]	1.05	1.17	0.60	0.53
Age 45-49	-	-	1.45***	0.46	−4.94***	1.02	−15.65	-[a]	0.04	0.98	0.45	0.63
Age 50-54	-	-	0.69*	0.37	−6.32***	1.03	−16.18	-[a]	−0.27	0.97	0.86	0.60
Age 55-59	-	-	−0.32	0.32	−7.38***	1.05	−15.55	-[a]	−2.42*	1.33	1.20	0.93
Age 60-64	-	-	−0.31	0.32	−7.45***	1.07	−16.18	-[a]	−2.62*	1.35	2.48**	1.25
Age 65-69	-	-	−0.09	0.37	-	-	−16.69	-[a]	-	-	16.54	-[a]
Age 70-74	-	-	0.68	0.46	-	-	−15.39	-[a]	-	-	17.00	-[a]
Age 75-79	-	-	1.43**	0.64	-	-	−15.78	-[a]	-	-	17.67	-[a]
Age 80-84	-	-	0.89	0.69	-	-	4.20	-[a]	-	-	16.36	-[a]
Age 85 plus	-	-	17.44	-[a]	-	-	−16.04	-[a]	-	-	18.35	-[a]
Age 15-24	-	-	-	-	−2.42**	1.15	-	-	−1.82	1.33	-	
Age 65 plus	-	-	-	-	−8.37***	1.24	-	-	−3.11**	1.40	-	

Continues on next page

Continues from previous page

	Single person		Couple w/o child.		Couple w. child.		Lone parent		Multifamily HH		Other	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Employment status of HH head (ref. categ.: employed full-time)												
Employed part-time	-	-	0.00	0.22	−0.46**	0.18	0.21	0.28	−0.11	1.12	0.11	0.25
Unemployed	-	-	−0.14	0.38	0.47	0.34	2.31**	1.03	−17.42	-[a]	2.66**	1.03
Not in labour force	-	-	0.17	0.25	−0.48**	0.20	0.40	0.31	1.49**	0.59	0.64	0.41
HH head has long term health condition	-	-	−0.35*	0.19	0.53***	0.16	−0.07	0.25	−0.34	0.61	0.24	0.35
Constant	-	-	1.05***	0.19	4.24***	1.02	17.54	-[a]	−1.38	0.85	1.07*	0.56
Transition to couple with children at wave t + 1												
Years intact is left truncated	0.11	0.17	0.40	0.39	0.19	0.19	0.44	0.63	1.25	0.93	−0.05	0.90
Years HH intact (ref. categ.: one year)												
2 years	−0.29**	0.12	−0.30*	0.17	−0.19*	0.10	0.39	0.31	−0.65**	0.33	−0.30	0.41
3 years	−0.31**	0.15	0.00	0.22	−0.33***	0.12	−0.08	0.34	−0.04	0.51	−0.70	0.70
4 years	−0.12	0.17	0.26	0.27	−0.20	0.15	1.09*	0.57	−0.23	0.89	−1.69	1.12
5 to 9 years	−0.47***	0.15	0.10	0.23	−0.03	0.12	0.35	0.33	0.89	0.82	−0.37	1.00
10 years or more	−0.37	0.29	−0.53	0.39	0.58***	0.16	0.41	0.73	20.48	-[a]	−20.97	-[a]
Has other related	-	-	−3.68***	0.20	1.32***	0.26	−1.33***	0.41	0.66	0.49	−0.73	0.82
Has other not related	-	-	−3.64***	0.28	1.26***	0.44	−1.24**	0.51	0.01	0.96	−0.69	0.85
Age HH head (ref. categ.: 25-29 years)												
Age 15-19	0.82***	0.19	−0.61	0.47	-	-	0.13	-[a]	-	-	0.69	0.95
Age 20-24	0.30**	0.14	−0.51**	0.21	-	-	−16.98	-[a]	-	-	0.87**	0.43
Age 30-34	0.17	0.17	1.12***	0.22	−1.62	1.06	−0.04	-[a]	20.29	-[a]	0.58	0.57
Age 35-39	0.87***	0.18	1.22***	0.31	−1.67	1.06	−16.60	-[a]	0.56	0.83	1.23*	0.70
Age 40-44	0.98***	0.19	0.55	0.34	−3.44***	1.01	−18.17	-[a]	1.48*	0.78	0.79	0.82
Age 45-49	0.49**	0.21	1.29***	0.44	−4.54***	1.01	−18.57	-[a]	0.31	0.56	1.36	0.87
Age 50-54	0.05	0.24	0.53	0.34	−5.26***	1.00	−20.00	-[a]	0.07	0.54	1.23	0.87
Age 55-59	−0.51	0.31	0.01	0.27	−5.93***	1.00	−20.20	-[a]	−0.80	0.56	1.27	1.38
Age 60-64	−0.59	0.41	−0.61**	0.29	−6.39***	1.01	−20.65	-[a]	−1.23**	0.58	3.30**	1.41
Age 65-69	−0.03	0.44	−0.60*	0.34	-	-	−21.97	-[a]	-	-	0.62	-[a]
Age 70-74	−0.97	0.78	−0.70	0.47	-	-	−20.18	-[a]	-	-	0.77	-[a]
Age 75-79	0.25	0.54	−0.55	0.70	-	-	−20.22	-[a]	-	-	0.82	-[a]
Age 80-84	0.69	0.70	−0.61	0.72	-	-	−18.54	-[a]	-	-	0.62	-[a]
Age 85 plus	−14.44	-[a]	14.81	-[a]	-	-	−37.24	-[a]	-	-	19.54	-[a]
Age 15-24	-	-	-	-	−2.53**	1.12	-	-	−0.97*	0.58	-	-
Age 65 plus	-	-	-	-	−6.85***	1.01	-	-	−2.30***	0.69	-	-

Continues on next page

Continues from previous page

	Single person		Couple w/o child.		Couple w. child.		Lone parent		Multifamily HH		Other	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Employment status of HH head (ref. categ.: employed full-time)												
Employed part-time	0.26**	0.12	−0.23	0.20	−0.33***	0.11	0.41	0.29	0.55	0.39	−0.18	0.38
Unemployed	0.78***	0.17	0.00	0.33	−0.12	0.27	1.31	1.04	1.16**	0.53	1.55	1.25
Not in labour force	0.48***	0.17	0.00	0.22	−0.04	0.11	0.29	0.32	0.44	0.32	0.10	0.59
HH head has long term health condition	−0.17	0.13	−0.43**	0.18	0.06	0.10	−0.15	0.26	−0.03	0.28	0.21	0.50
Constant	−1.05***	0.12	2.05***	0.17	6.22***	1.00	19.86	-^{a}	0.22	0.48	−0.60	0.94
Transition to lone parent family at wave t+1												
Years intact is left truncated	0.12	0.19	0.99	0.65	0.56**	0.28	0.51	0.61	1.62	1.03	0.08	0.93
Years HH intact (ref. categ.: one year)												
2 years	−0.12	0.13	0.06	0.38	−0.61***	0.17	0.23	0.30	−0.52	0.45	−0.48	0.45
3 years	−0.46***	0.17	−0.07	0.54	−0.71***	0.20	−0.11	0.32	0.36	0.66	−1.09	0.75
4 years	−0.33*	0.19	−0.03	0.67	−0.41	0.25	0.64	0.54	0.80	0.94	−1.06	0.82
5 to 9 years	−0.66***	0.16	−0.40	0.60	0.36*	0.19	−0.01	0.31	0.78	1.21	−1.21	1.05
10 years or more	−0.78***	0.28	−0.04	0.88	0.82***	0.28	0.65	0.63	−1.41	-^{a}	−3.18*	1.66
Has other related	-	-	−1.54***	0.40	0.87**	0.36	0.45	0.37	0.77	0.60	−0.52	0.78
Has other not related	-	-	−2.00***	0.64	0.40	0.63	−0.52	0.44	0.43	1.08	−0.90	0.80
Age HH head (ref. categ.: 25-29 years)												
Age 15-19	0.87***	0.24	0.38	0.77	-	-	−0.07	-^{a}	-	-	2.40***	0.85
Age 20-24	0.37*	0.19	0.16	0.45	-	-	−16.88	-^{a}	-	-	1.14**	0.55
Age 30-34	0.19	0.24	−0.36	0.68	−1.94*	1.07	0.13	-^{a}	20.00	-^{a}	0.74	0.74
Age 35-39	1.12***	0.24	1.33**	0.56	−1.59	1.07	−15.82	-^{a}	0.61	1.17	1.21	0.87
Age 40-44	1.60***	0.23	0.53	0.72	−3.20***	1.03	−16.97	-^{a}	2.45**	0.96	1.26	0.88
Age 45-49	1.77***	0.23	0.60	0.87	−4.54***	1.02	−16.88	-^{a}	0.67	0.81	2.12**	0.90
Age 50-54	1.81***	0.23	−0.29	0.83	−6.00***	1.03	−17.85	-^{a}	−0.25	0.84	1.29	1.03
Age 55-59	1.77***	0.24	−0.74	0.70	−6.96***	1.04	−17.78	-^{a}	−1.26	0.94	3.90***	1.06
Age 60-64	1.61***	0.29	−1.60*	0.82	−8.18***	1.13	−18.71	-^{a}	−0.95	0.85	3.97***	1.40
Age 65-69	1.61***	0.35	−16.19	-^{a}	-	-	−18.98	-^{a}	-	-	17.59	-^{a}
Age 70-74	0.81*	0.49	−1.60	1.15	-	-	−17.57	-^{a}	-	-	18.51	-^{a}
Age 75-79	1.52***	0.43	−16.64	-^{a}	-	-	−20.27	-^{a}	-	-	20.12	-^{a}
Age 80-84	2.18***	0.56	0.39	0.95	-	-	0.65	-^{a}	-	-	16.97	-^{a}
Age 85 plus	2.65***	0.62	−2.00	-^{a}	-	-	−18.78	-^{a}	-	-	19.10	-^{a}
Age 15-24	-	-	-	-	−2.37**	1.15	-	-	−0.49	0.84	-	-
Age 65 plus	-	-	-	-	−7.01***	1.06	-	-	−1.33	0.89	-	-

Continues on next page

Continues from previous page

	Single person		Couple w/o child.		Couple w. child.		Lone parent		Multifamily HH		Other	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Employment status of HH head (ref. categ.: employed full-time)												
Employed part-time	0.63***	0.14	1.11***	0.42	0.60***	0.16	0.49*	0.27	1.41***	0.50	−0.12	0.44
Unemployed	1.29***	0.18	1.27**	0.57	1.11***	0.33	1.69	1.03	2.06***	0.63	3.10***	1.12
Not in labour force	1.21***	0.16	1.67***	0.44	1.06***	0.17	1.04***	0.30	1.25***	0.43	0.23	0.56
HH head has long term health condition	0.31**	0.12	0.30	0.36	0.08	0.16	−0.26	0.24	0.07	0.37	1.04**	0.44
Constant	−2.17***	0.17	−2.24***	0.43	3.62***	1.02	19.25	-^{a}	−1.72**	0.74	−1.44	0.94
Transition to multifamily at wave t + 1												
Years intact is left truncated	0.37	0.42	0.32	0.51	−0.01	0.36	0.94	0.67	0.80	0.92	−15.96	-^{a}
Years HH intact (ref. categ.: one year)												
2 years	−0.27	0.31	−0.16	0.25	−0.42**	0.19	0.27	0.35	−0.35	0.29	−0.97	1.11
3 years	0.31	0.34	0.46	0.29	−0.66***	0.24	−0.22	0.41	−0.12	0.49	−16.07	-^{a}
4 years	−0.39	0.55	0.29	0.37	−0.28	0.27	−0.43	0.80	0.12	0.72	−16.87	-^{a}
5 to 9 years	−2.21**	1.02	0.13	0.32	−0.30	0.23	−0.33	0.44	1.01	0.74	−15.62	-^{a}
10 years or more	−13.26	-^{a}	−0.19	0.52	0.30	0.28	0.65	0.80	−1.11	-^{a}	−5.38	-^{a}
Has other related	-	-	−2.11***	0.26	1.08***	0.38	0.23	0.43	1.64***	0.43	−16.34	-^{a}
Has other not related	-	-	−2.25***	0.40	1.16*	0.59	−0.76	0.60	1.35	0.83	−15.66	-^{a}
Age HH head (ref. categ.: 25-29 years)												
Age 15-19	1.27***	0.44	1.32**	0.54	-	-	1.25	-^{a}	-	-	0.65	1.27
Age 20-24	0.64*	0.38	0.42	0.30	-	-	−16.58	-^{a}	-	-	−0.51	0.72
Age 30-34	−1.06	0.78	0.24	0.36	−1.90*	1.14	−0.36	-^{a}	19.12	-^{a}	−16.29	-^{a}
Age 35-39	−0.41	0.79	0.06	0.53	−1.53	1.13	−15.77	-^{a}	1.63*	0.88	−16.10	-^{a}
Age 40-44	−0.24	0.79	0.03	0.57	−2.76**	1.08	−17.35	-^{a}	1.82**	0.86	−16.08	-^{a}
Age 45-49	0.26	0.68	1.94***	0.52	−3.17***	1.06	−16.92	-^{a}	1.49**	0.64	−15.36	-^{a}
Age 50-54	0.72	0.58	1.26***	0.44	−3.96***	1.06	−18.63	-^{a}	1.20*	0.62	−15.39	-^{a}
Age 55-59	0.23	0.69	1.13***	0.35	−4.30***	1.06	−17.38	-^{a}	0.56	0.63	−16.28	-^{a}
Age 60-64	−0.52	1.08	0.44	0.39	−4.72***	1.08	−18.36	-^{a}	−0.18	0.65	3.95**	1.71
Age 65-69	−13.39	-^{a}	0.48	0.46	-	-	−36.94	-^{a}	-	-	1.40	-^{a}
Age 70-74	−13.99	-^{a}	0.74	0.58	-	-	−17.29	-^{a}	-	-	15.48	-^{a}
Age 75-79	−14.12	-^{a}	0.44	0.93	-	-	−37.14	-^{a}	-	-	16.20	-^{a}
Age 80-84	−13.55	-^{a}	−0.32	1.20	-	-	−18.08	-^{a}	-	-	3.94	-^{a}
Age 85 plus	−14.13	-^{a}	16.06	-^{a}	-	-	−36.66	-^{a}	-	-	17.66	-^{a}
Age 15-24	-	-	-	-	−1.41	1.20	-	-	0.04	0.66	-	-
Age 65 plus	-	-	-	-	−4.64***	1.09	-	-	−0.14	0.67	-	-

Continues on next page

Continues from previous page

	Single person		Couple w/o child.		Couple w. child.		Lone parent		Multifamily HH		Other	
	b	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.	*b*	S.E.
Employment status of HH head (ref. categ.: employed full-time)												
Employed part-time	1.13***	0.31	−0.34	0.28	−0.29	0.21	−0.19	0.36	0.65*	0.37	0.05	0.77
Unemployed	1.56***	0.39	−0.05	0.45	0.21	0.42	1.73	1.08	1.14**	0.53	2.81*	1.51
Not in labour force	1.62***	0.36	−0.08	0.29	0.41**	0.19	0.70*	0.36	0.54*	0.30	−14.93	-[a]
HH head has long term health condition	0.18	0.32	−0.06	0.23	−0.10	0.18	0.15	0.30	−0.21	0.27	0.16	1.14
Constant	−3.93***	0.37	−0.65**	0.26	2.48**	1.06	17.37	-[a]	−0.83	0.58	14.00	-[a]
Transition to other household type at wave t+1												
Years intact is left truncated	−0.02	0.20	1.04	0.72	−13.76	-[a]	0.97	0.95	−12.70	-[a]	−0.65	0.74
Years HH intact (ref. categ.: one year)												
2 years	0.00	0.13	−0.32	0.43	−0.01	0.51	0.74	0.52	1.41	1.42	−0.09	0.30
3 years	−0.19	0.17	−0.07	0.55	−1.33	1.07	−0.57	0.82	−13.50	-[a]	−0.78	0.51
4 years	−0.03	0.19	−1.15	1.07	−0.49	1.07	0.21	1.17	−12.32	-[a]	−0.93	0.61
5 to 9 years	−0.15	0.16	−1.58	1.06	0.04	0.81	−1.28	1.10	−0.16	-[a]	−0.22	0.79
10 years or more	−0.76**	0.37	−13.81	-[a]	1.10	0.84	0.64	1.21	29.01	-[a]	−4.13**	2.05
Has other related	-	-	−2.25***	0.55	−12.45	-[a]	0.91	0.58	2.18	1.40	0.36	0.45
Has other not related	-	-	−2.52***	0.76	1.55	1.14	1.41**	0.64	−13.92	-[a]	0.70	0.48
Age HH head (ref. categ.: 25-29 years)												
Age 15-19	0.55***	0.20	0.76	0.89	-	-	2.22	-[a]	-	-	0.88	0.61
Age 20-24	0.03	0.15	0.55	0.47	-	-	−16.63	-[a]	-	-	−0.02	0.28
Age 30-34	−0.30	0.19	0.19	0.64	−17.92	-[a]	−14.62	-[a]	2.93	-[a]	0.30	0.37
Age 35-39	0.06	0.22	−0.32	1.09	−2.10	1.34	−16.04	-[a]	−17.13	-[a]	0.09	0.56
Age 40-44	−0.35	0.27	0.78	0.74	−4.11***	1.37	−15.59	-[a]	−15.72	-[a]	−0.15	0.62
Age 45-49	−0.06	0.25	1.47*	0.79	−5.30***	1.37	−15.97	-[a]	−16.45	-[a]	−0.11	0.76
Age 50-54	−0.63**	0.30	0.20	0.85	−5.33***	1.28	−16.88	-[a]	−17.42	-[a]	−1.69	1.15
Age 55-59	−0.32	0.29	−1.17	1.09	−6.26***	1.37	−15.87	-[a]	−17.34	-[a]	0.51	1.06
Age 60-64	−0.22	0.33	−15.17	-[a]	−6.03***	1.38	−16.38	-[a]	−16.75	-[a]	1.87	1.33
Age 65-69	0.32	0.36	−15.17	-[a]	-	-	−34.11	-[a]	-	-	16.62	-[a]
Age 70-74	0.21	0.45	−15.04	-[a]	-	-	−33.53	-[a]	-	-	16.35	-[a]
Age 75-79	−0.79	0.61	−14.95	-[a]	-	-	−15.61	-[a]	-	-	16.61	-[a]
Age 80-84	0.45	0.64	−15.85	-[a]	-	-	−16.62	-[a]	-	-	16.98	-[a]
Age 85 plus	0.25	0.81	−0.30	-[a]	-	-	−34.17	-[a]	-	-	18.63	-[a]
Age 15-24	-	-	-	-	−1.19	1.34	-	-	−0.31	1.35	-	-
Age 65 plus	-	-	-	-	−23.61	-[a]	-	-	−14.81	-[a]	-	-
Employment status of HH head (ref. categ.: employed full-time)												
Employed part-time	0.49***	0.14	0.20	0.47	0.05	0.67	0.18	0.57	−13.58	-[a]	0.34	0.27
Unemployed	0.20	0.24	−1.14	1.09	0.97	0.85	1.49	1.30	1.48	1.44	2.80***	1.05
Not in labour force	1.15***	0.16	0.14	0.62	1.13**	0.52	0.55	0.56	−13.26	-[a]	0.18	0.47
HH head has long term health condition	0.04	0.14	0.51	0.45	−0.20	0.58	0.63	0.45	1.19	1.40	0.26	0.38
Constant	−1.30***	0.13	−1.70***	0.44	0.96	1.22	14.08	-[a]	−3.03**	1.53	−0.37	0.54

[a] Large standard error (> 100)   * $p < 0.10$ ** $p < 0.05$ **** $p < 0.01$

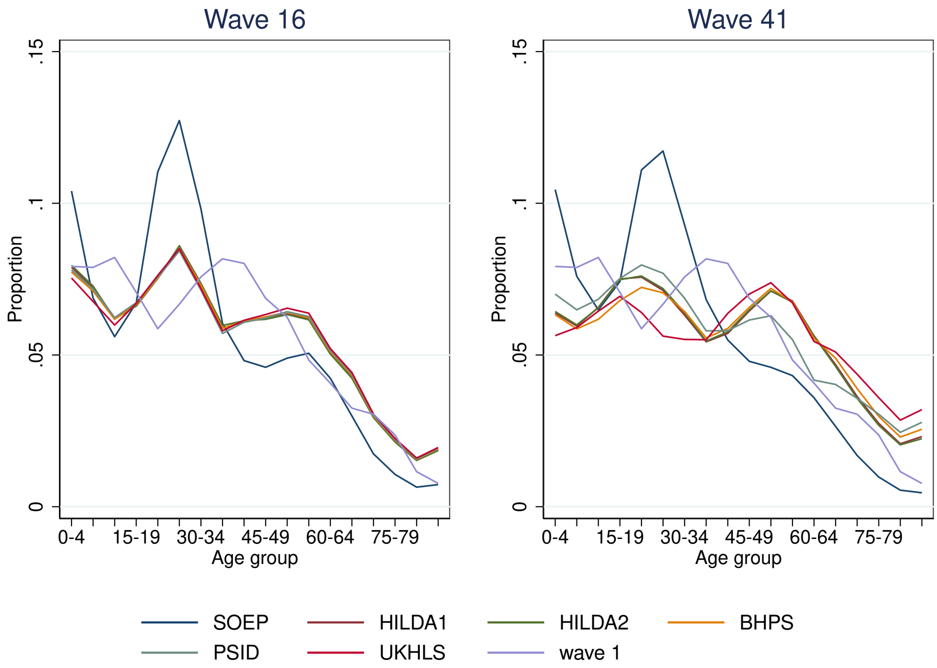

Figure A2. Age distribution of the simulated samples at wave 16 and 41 for the various following rules, assuming full response

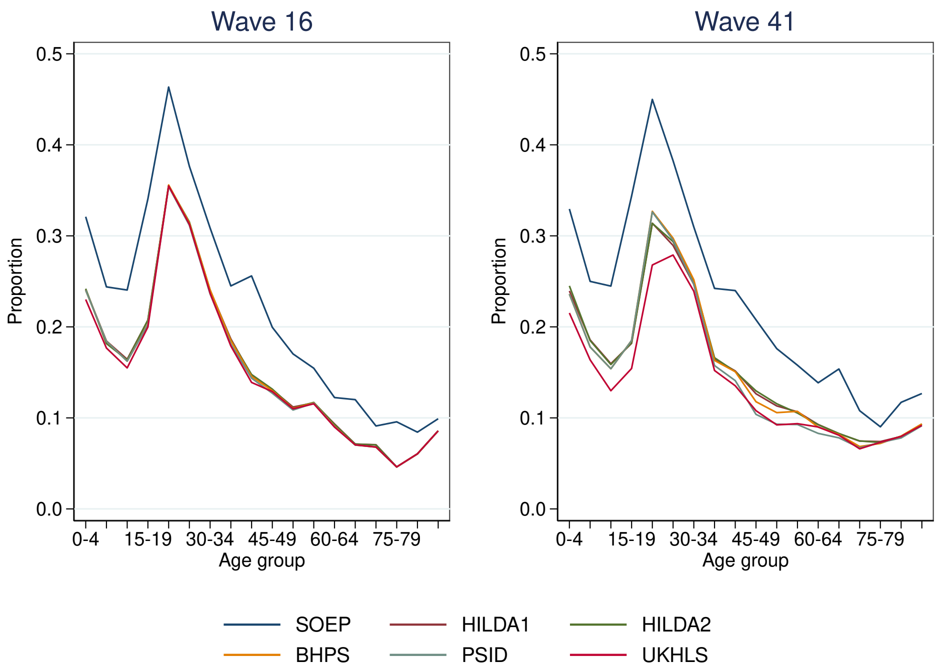

Figure A3. Proportion of people moving in the last 12 months in wave 16 and 41, by age group, assuming full response