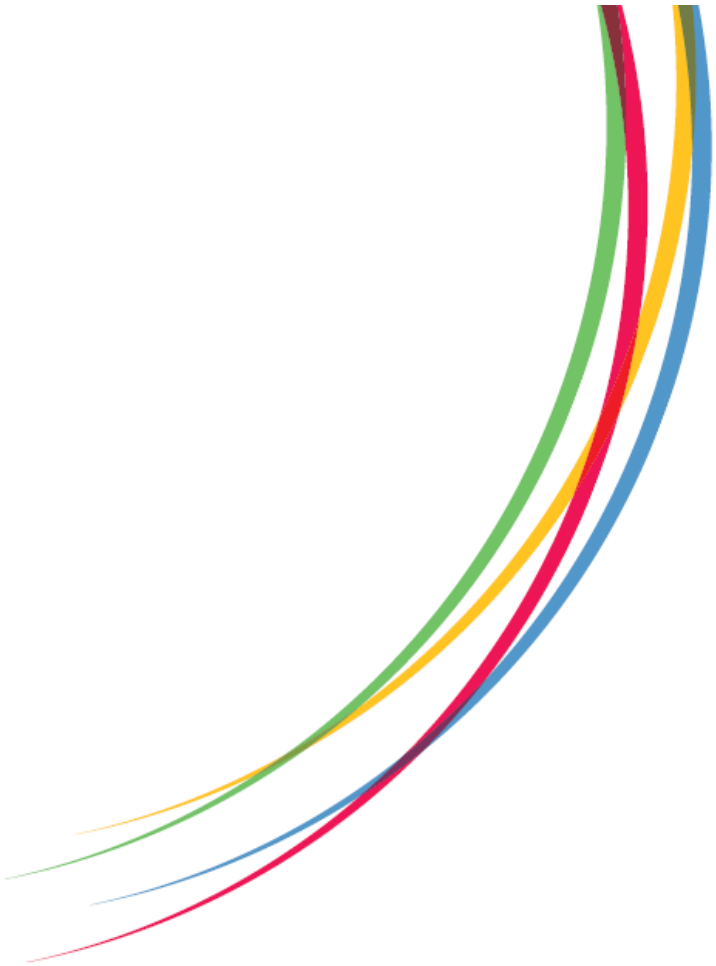

Understanding Society

Working Paper Series

No. 2017 – 05

May 2017

Mounting Multiple Experiments on Longitudinal Social Surveys: Design and Implementation Considerations

Peter Lynn & Annette Jäckle

Institute for Social and Economic Research, University of Essex

Mounting Multiple Experiments on Longitudinal Social Surveys: Design and Implementation Considerations

Peter Lynn & Annette Jäckle

Non-Technical Summary

Sample-based surveys can be designed to represent an entire population and they can therefore be used to quantify the characteristics of the population. If the survey involves collecting data repeatedly over a period of time from the same set of people, this is known as a longitudinal survey. Longitudinal surveys can be used to study how people's behaviour, circumstances or attitudes change over time.

Experiments, on the other hand, are used to study the effect of a treatment of some kind on a specific outcome of interest. A sample is randomly divided in groups and each group is administered a different treatment. The outcome is subsequently observed. Typically, the sample is not representative of an entire population.

Mounting experiments on a longitudinal survey brings together the advantage of being able to identify a causal effect (experiment), the advantage of population representation (survey) and the advantage of longitudinal observation (measuring change). This is a very powerful design. However, it is also a somewhat complex design, particularly when many experiments are mounted on the same survey. This article outlines some of the challenges in designing and carrying out experiments on longitudinal surveys and discusses some ways in which these challenges can be met.

The issues that we discuss include how to avoid one experiment influencing the results of another one, how to allocate sample persons to treatment groups, and how best to take advantage of the longitudinal context.

Mounting Multiple Experiments on Longitudinal Social Surveys: Design and Implementation Considerations

Peter Lynn & Annette Jäckle

Abstract

Mounting experiments on longitudinal surveys adds a further dimension to the value of randomised experiments (designed to maximise internal validity) mounted on probability surveys (to maximise external validity): for example, repeated measurement over time can be used to test effects on inherently longitudinal outcomes, or to test inherently longitudinal treatments. The unique value of experimentation in longitudinal surveys is, however, matched by unique challenges in design and implementation. We summarise key methodological features and challenges based on experiences with the *Understanding Society* Innovation Panel, a probability-based household panel with annual interviews that exists solely for experimentation and methodological development.

Key words: experimental design; longitudinal surveys

JEL classifications: C81, C83

Author contact details: plynn@essex.ac.uk

Acknowledgements: A revised version of this article will be published as a chapter in the book *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment* (editors Paul J. Lavrakas, Edith D. de Leeuw, Allyson Holbrook, Courtney Kennedy, Michael W. Traugott, and Brady T. West), to be published by Wiley in 2018. Readers wishing to cite this article should cite the book version, not this Working Paper. Both authors are part-funded by awards from the UK Economic and Social Research Council to the University of Essex for *Understanding Society*, the UK Household Longitudinal Study.

Mounting Multiple Experiments on Longitudinal Social Surveys: Design and Implementation Considerations

Peter Lynn & Annette Jäckle

1. Introduction and Overview

There are now various longitudinal surveys that are used for experimentation. Several of these studies specifically invite proposals from external researchers. The longitudinal aspect adds a further dimension to the value of randomised experiments (designed to maximise internal validity) mounted in probability surveys (designed to maximise external validity): the repeated measurement of sample members over time can be used to test effects on inherently longitudinal outcomes, to take the histories of sample members into account in analysing experimental outcomes, to analyse the long-term effects of experimental treatments on outcomes measured in later waves, or to exploit within as well as between respondent allocations to treatments by repeating experiments across waves. The unique value of experimentation in longitudinal surveys is matched by unique challenges in successfully designing and implementing experiments in a longitudinal context.

This chapter summarises key methodological features and challenges based on experiences with the *Understanding Society* Innovation Panel, a probability-based household panel with annual interviews that exists solely for the purposes of experimentation and methodological development. The aim is to raise awareness of unique issues that arise when mounting multiple independent experiments on the same survey vehicle and, particularly, when the survey and the experiments are longitudinal.

Section 2 provides an overview of the types of experiments that can be carried in longitudinal surveys, section 3 discusses the distinction between longitudinal experiments and experiments in longitudinal studies, and section 4 provides an

overview of international longitudinal studies that are used as platforms for experimentation. Section 5 provides further information on the design of the *Understanding Society* Innovation Panel and the types of experiments that have been implemented on it, while sections 6 and 7 discuss how to avoid confounding and how to allocate units to treatments, respectively. Section 8 discusses the introduction of refreshment samples, which is a particular feature of longitudinal surveys that can strengthen the experimental setting. The final section provides a discussion of key lessons learned and possible future methodological developments.

2. Types of Experiments that can be Mounted in a Longitudinal Survey

Different types of experiments can be embedded in a longitudinal survey. Table 1 presents an overview, where the horizontal axis indicates a hypothetical sequence of interviews at different points in time. Each design can be used to address different types of (longitudinal) research questions.

The classic experimental design, as described in standard handbooks such as Campbell and Stanley (1963) or De Vaus (2001), is a *pre-test post-test* control group design (1): the outcome of interest is measured before and after the exposure to treatment and a randomised control group is measured at the same points in time but not exposed to the treatment. To estimate the treatment effect the changes in the outcomes of the treatment and control groups are compared. Other information collected about participants in the pre-test interview can be used substantively to study whether treatment effects depend on pre-existing characteristics of study participants, or methodologically to select sub-samples with certain characteristics for the experiment, or to test for and estimate the impact of differential attrition in the experimental conditions (see, Farrington, Loeber and Welsh 2010). If the topic of the experiment does not require prior measures of either the outcome or covariates, the pre-test measurement may be omitted: in a *post-test* design (2) the treatment effect is estimated by comparing the outcomes of the treatment and the control groups.

In many survey experiments the exposure to treatment and measurement of the outcome occur simultaneously: for example in experiments testing the effects of

reversing the order of response options in survey questions, participants are exposed to the treatment when they answer the question that measures the outcome. In this case the classic experimental design becomes a *pre-test test* design (3) and the post-test design becomes a *cross-sectional test* (4). Cross-sectional tests can equally well be carried out in a single cross-sectional survey, as they do not make use of the repeated measures nature of longitudinal data.

Further standard experimental designs include *multiple post-tests* (5) or *multiple pre-tests* (6) where participants are interviewed multiple times before or after the treatment exposure. *Multiple pre-test* studies can be used to identify existing change, or time trends, in the outcome before the treatment exposure. *Multiple post-test* studies (sometimes referred to as growth designs) can be used to estimate long-term effects of treatments, for example to study how outcomes evolve with age, to identify delayed effects or to compare short and long-term effects (i.e. the growth trajectory of the treatment effect). Compared to studies with a single post-test interview, multiple post-test studies can identify and distinguish immediate lasting effects of a treatment, immediate but short-lived effects, delayed lasting effects, delayed undesirable effects on an existing trend, no effects of a treatment because of a pre-existing trend and haphazard oscillation (see Figure 24.1 in Farrington, Loeber and Welsh 2010). For example, in criminology multiple post-tests are used to study the effect of interventions such as counselling or training of pre-school or school age children. Follow-up interviews may take place at irregular intervals over several decades, to study outcomes such as criminal activity, drug use, educational attainment and labour market outcomes (Farrington, Loeber and Welsh 2010). Similarly, in developmental epidemiology preventive trials with long-term follow-ups are for example used to test ways of reducing the risk of mental health problems in children of divorce, with behavioural or learning problems or whose parents are being treated for depression (Brown and Liao 1999).

Repeated tests (7) can be used to study how long-term treatments should best be applied to maximise effectiveness. For example clinical dynamic treatment trials are sequences of randomized trials, where at each stage the randomizations may depend on outcomes of previous stages, with the aim of developing optimum sequences of treatments by exploiting carry-over effects (Chakraborty and Murphy 2014). Repeated experiments can also be used to study how reliable and

reproducible estimated treatment effects are if the conditions of the experiment are varied (see, John and Quenouille 1977), or to study longitudinal outcomes.

In repeated designs the treatment conditions can either be held the same across implementations (between subject designs), or crossed (within subject designs). With cross-over designs each subject is observed in multiple treatment conditions and the same outcome is measured in each condition, such that each subject contributes multiple scores. Compared to between-subject designs this offers two advantages (Maxwell and Delaney 2004). Firstly the repeated measures mean that a smaller number of subjects is needed to reach a certain level of statistical power. This is a clear advantage when the costs of recruiting subjects is high, in terms of money, time, or effort. Secondly, as each subject serves as his or her own control, variability in individual differences between subjects is removed from the error term, which increases statistical power.

Table 1: Typology of experiments in longitudinal surveys

Type of experiment	R	Interview 1		Interview 2	Interview 3		Interview 4
1 Pre-test post-test	T	O	X	O			
	C	O		O			
2 Post-test	T		X	O			
	C			O			
3 Pre-test test	T	O		XO			
	C	O		O			
4 Cross-sectional test	T			XO			
	C			O			
5 Multiple post-tests	T	(O)	X	O	O		O
	C	(O)		O	O		O
6 Multiple pre-tests	T	O		O	O	X	O
	C	O		O	O		O
7 Repeated tests	T	O		XO	XO		XO
	C	O		O	O		O

Notes: R = randomisation, T = treatment group, C = control group, O = observation or measurement of outcome, X = exposure to treatment.

3. Longitudinal Experiments and Experiments in Longitudinal Surveys

Not all longitudinal experiments, where data are collected about respondents at two or more points in time, are mounted on longitudinal surveys. There are many examples of free-standing field experiments that include follow-up surveys designed to test the long-term effects of treatments or interventions in economics (e.g. Aguila, Kapteyn and Smith 2015; Dupas and Robinson 2013), education (e.g. Hu et al. 2007), psychology (e.g. Acredolo 1978; Yeager et al. 2013), health (e.g. Marcus 1982; Olds et al. 2004), criminology (e.g. Belfield et al. 2006; Boisjoli et al. 2007; Ellickson and Bell 1990; McCord 2003), computer science (e.g. Lee et al. 2012; Wiedenbeck et al. 2005), market research (e.g. Aaker, Fournier and Brasel 2004; Bolton and Drew 1991), and management research (Dvir et al. 2002; Workman and Bommer 2004). Such experimental studies correspond to designs (2) or (5) in Table 1.

Where experiments are embedded in a pre-existing longitudinal survey, two scenarios can be distinguished. In the first scenario, the researchers responsible for a longitudinal survey may carry out an experiment on their survey to inform design decisions. For example, before switching from annual to biennial interviewing, the National Longitudinal Survey of Youth 1979 implemented an experiment to see how less frequent interviews would affect the quality of recall data (Pierret 2001). A sub-sample of respondents were asked to report on events in the past two years, rather than the year since the previous interview, simulating the two-year interviewing schedule and allowing comparison with data collected in an annual interview schedule. Similarly, in the 2013 wave of the UK National Child Development Study an experiment was conducted to test the effects of introducing web as a main mode of interviewing on attrition and measurement (Brown and Hancock 2015). A random control group were assigned to telephone only interviewing. The rest of the sample were invited to complete the survey online and nonrespondents were followed up by telephone interviewers.

In the second scenario the longitudinal survey is a multi-purpose vehicle forming part of the infrastructure for academic research. This chapter focuses on the latter: multi-purpose longitudinal surveys that are used for multiple experiments.

4. Longitudinal Surveys that Serve as Platforms for Experimentation

Table 2 provides a summary of international longitudinal surveys of probability samples that are used as multi-purpose platforms for social scientists to collect experimental data. There also exist a small number of commercial longitudinal panels based on probability samples that are occasionally used for social science experiments, but we focus here on the panels that exist primarily for experimental research purposes:

- The Innovation Panel: a household panel study in Great Britain that is part of the UK Household Longitudinal Study: *Understanding Society*
- The SOEP Innovation Sample (SOEP-IS): a household panel study in Germany that is part of the German Socio Economic Panel study
- The LISS panel: a probability-based online panel in the Netherlands
- The GESIS Panel: a probability-based mixed mode (online-mail) panel in Germany
- The ELIPSS panel: a probability-based online panel in metropolitan France
- The American Life Panel: an online panel in the U.S. that grew out of studies exploring the opportunities for internet interviewing in the Health and Retirement Study
- The Understanding America Study (UAS): a probability-based online panel in the U.S.

All of the panel surveys are scientific infrastructure projects enabling academic researchers in the social sciences to collect data. The funding for four of the surveys (the Innovation Panel, SOEP-IS, GESIS Panel and ELIPSS) is such that standard data collection is free for proposers; while the other panels' proposers have to pay for the data collection. The Innovation Panel and the SOEP-IS are modelled on their 'parent' household panel surveys. The American Life Panel originally derived from the Survey of Consumer Attitudes, but later added new samples based on face to

face interviewing and address based sampling. The other panels are free-standing surveys.

Sampling methods differ between the studies, but all are probability based. The Innovation Panel, SOEP-IS, the LISS panel, and UAS interview all members of households, while the other panels are samples of individuals. Sample sizes vary, but all studies regularly add refreshment samples. The frequency of interviewing varies: interviews are annual in the Innovation Panel and SOEP-IS, every two months in the GESIS Panel, monthly in the LISS and ELIPSS panels and twice a month in the American Life Panel and UAS. Both the Innovation Panel and the SOEP-IS are primarily CAPI surveys, with some experimental testing of other modes. All other surveys are primarily online surveys, where sample members without internet are either sent a paper questionnaire (GESIS Panel), or loaned a tablet (ELIPSS) or computer (LISS, American Life Panel, UAS) with broadband access. Design and fieldwork procedures are summarized in Blom et al (2015) for LISS and ELIPSS, Bosnjak et al (2017) and the GESIS Panel, Al Baghal and Jäckle (2016) for the Innovation Panel and Richter and Schupp (2015) for the SOEP-IS.

All panels have carried experiments testing aspects of questionnaire design and question wording. Several of the panels have also experimentally tested survey procedures such as different modes of data collection, respondent incentives, or audio-recording versus writing in responses to open ended questions. Other types of experiments include information treatments, experiments to measure risk attitudes, financial decision making or decision making under uncertainty, or factorial surveys where the content of vignettes and allocation to respondents rely on randomisation.

Researchers wishing to implement experiments on any of these panels submit proposals and draft questionnaires which are peer reviewed and assessed for scientific merit. Successful proposals are implemented by the survey teams, except for the GESIS Panel where proposers have to programme the questionnaire themselves using the guidelines provided by GESIS. All studies provide scientific open access to the data collected on their panel. Table 2 provides links to further documentation for each of the studies.

5. The *Understanding Society* Innovation Panel

The *Understanding Society* Innovation Panel is a platform for longitudinal methods research and social science experiments. It is an integral part of the design of the *Understanding Society* survey funded by the UK Economic and Social Research Council. Its purpose is to develop key innovations in survey methods and content that will ensure the future success of the *Understanding Society* survey, and more broadly to advance knowledge in the social sciences and in the methodology of designing longitudinal surveys. In addition to experimentation in the annual interviews, the focus of this chapter, the Innovation Panel is used as a base for Associated Studies collecting data using new and innovative mixed method approaches (see https://www.understandingsociety.ac.uk/research/get-involved/associated-studies).

The design of the Innovation Panel is based on the main *Understanding Society* survey. It consists of an original sample of around 2,500 persons, clustered within households, first fielded in 2008, plus refreshment samples of around 700 persons added in 2011 and again in 2014. Attrition rates are documented in the User Guide at https://www.understandingsociety.ac.uk/documentation/innovation-panel. The sample is a stratified, clustered sample of all persons resident in Great Britain, excluding northernmost Scotland. An equal probability sample of addresses was drawn from the UK Postcode Address File and all residents at selected addresses at the time of wave 1 became sample members (see Lynn 2009). Refreshment samples are added by selecting additional addresses from the existing primary sampling units.

All sample members are eligible for annual interviews and followed if they move within Great Britain. Wave 1 was fielded in 2008 and data from each wave are deposited the following year with the UK Data Service (https://discover.ukdataservice.ac.uk/catalogue/?sn=6849), from where they are available to researchers. New household members are eligible for interviews as long as they live with a sample member, but not followed if they move out. To maintain contact with participants and update addresses, a between-wave-mailing is sent out. The mailing includes a report of research findings, an address confirmation slip that

respondents are asked to return, and materials to encourage registration with the participant website (https://www.understandingsociety.ac.uk/participants).

The modes of data collection in wave 1 were CAPI with a paper self-completion module for adults (aged 16+), and a paper self-completion questionnaire for youth aged 10-15. Wave 2 included an experiment where for a random two-thirds of households interviews were first attempted by telephone and non-respondents were followed up by CAPI interviewers. The control group were interviewed by CAPI, as in wave 1 (see Lynn 2013). Waves 5 to 9 included an experiment where two-thirds of households were first invited to complete the survey online and non-respondents were followed up by CAPI interviewers. The control group were again interviewed in CAPI (Bianchi, Biffignandi and Lynn in press; Jäckle, Lynn and Burton 2015). The mode of the self-completion module was experimentally varied in waves 4 to 6, with a random half of respondents interviewed in CAPI allocated to a CASI version and the control maintaining the paper self-completion.

The initial waves of the Innovation Panel were used for development and testing of the main *Understanding Society* survey. Since wave 4 experiments are selected through an open competition: proposers submit a case for support including specification of the study design and draft questionnaires; the survey team assess feasibility and costs; a review panel including external reviewers assesses the scientific merit and value for money and suggests a ranking of proposals; the *Understanding Society* Executive Team decide which proposals to accept. The criteria for the competition are described at https://www.understandingsociety.ac.uk/research/get-involved/innovation-panel-competition.

Over the first nine waves, a total of 42 unique experiments have been fielded in the Innovation Panel, some of which are replicated over multiple waves. The experiments have included experiments with survey procedures (such as the mode experiments described above, experiments with the value of respondent incentives, or with the format and content of between wave mailings), experiments with generic questionnaire design issues (such as the use of showcards, with the labelling or direction of scales, or with the wording of dependent interviewing questions), and experiments with questionnaire design to measure specific concepts (such as testing

ways of measuring consumption or wealth, life satisfaction, identity, or self-assessed disability). In addition there have been vignette studies, studies with randomised information treatments, and some non-experimental methodological studies, for example measuring finger lengths as indicators of prenatal testosterone exposure, or testing time use diaries. All experiments implemented to date are described in the User Guide, including references to resulting publications (https://www.understandingsociety.ac.uk/documentation/innovation-panel). For each wave of the Innovation Panel an *Understanding Society* Working Paper is published which documents the rationale, design and early findings from each experiment (Al Baghal 2015; Al Baghal 2016; Al Baghal 2014; Burton 2012 ; Burton 2013; Burton et al. 2011; Burton, Laurie and Uhrig 2008; Burton, Laurie and Uhrig 2010).

6. Avoiding Confounding of Experiments

Experimentation relies on randomised allocation of observational units (sample members) to treatments in order to ensure that the effect of treatment is not confounded with any other factor that could influence the observed outcomes. Pure random allocation (effectively simple random sub-sampling) ensures the absence of confounding *on average* (i.e., invoking the expected value under the sub-sampling distribution), but the sample may not be well balanced between treatments, due to random sampling (allocation) variance. This risk is particularly great when there are many potential confounding factors, as when many experiments are carried out on a longitudinal survey. The larger the number of potential confounding factors (other experiments) the greater the chance of observing severe imbalance with respect to at least one of those factors, under simple random allocation.

With respect to any particular potential confounding factor, balance can be ensured by using stratified random allocation rather than simple random allocation, where the potential confounding factor acts as the stratification variable[1]. For example, if the outcome of interest is expected to be strongly influenced by the participant's age,

[1] The strata are referred to as 'blocks' in the classical experimental design literature (Addelman 1969), though in the survey context it is not necessary to assume that the stratification is explicit rather than implicit.

participants could be listed in age order before allocating alternately to treatment and control groups, thus ensuring a similar age distribution in each of the two groups. In the longitudinal survey context it is in principle therefore possible to achieve balance with respect to any other experiments administered at the current wave or at any previous wave, or with respect to any survey data collected previously. However, in practice there is a limit to the number of factors for which this can be done. Suppose we wish to allocate sample members to one of two treatment groups, A1 and A2, with an equal sample size to be allocated to each group. Table 3 illustrates the extent of imbalance that can arise with simple random allocation (upper panel) and how this imbalance can be removed with stratified random allocation (lower panel). The imbalance between the two treatment groups for experiment A is shown with respect to the treatments for three other experiments B, C and D, which have four, three and two treatment groups respectively. These may be experiments that were carried at previous waves of the survey, or they may be planned for the same wave as experiment A. In either case, in order to use stratified random allocation the allocation to treatments for B, C and D must have been made before the allocation to treatments for A. The distribution in the upper panel was obtained by allocating 504 sample units to each of A1 and A2 randomly, without regard to the distribution of the other three experimental indicators.

The potential for imbalance to affect observed outcomes can be seen in the upper panel of Table 3. For example, treatment group A1 contains a higher proportion of sample units allocated to D2 than treatment group A2. A simple comparison of the outcome between groups A1 and A2 will confound the effect of A2-A1 with a small proportion of the effect of D1-D2. This can be overcome either by controlling for the effect of D1-D2 in the analysis (for example, by carrying out a weighted analysis or by including the experiment D allocation as a covariate) or by controlling through design, as in the lower panel of the table. The distribution in the lower panel was obtained by sorting the sample units by the cross-classification of the other three experimental indicators before allocating alternately. The appendix provides `Stata` syntax for implementing each of these two allocation methods.

Table 3: Sample distributions generated by two alternative allocation methods

Sample distribution: 4 experiments, simple random allocation

	B1	B2	B3	B4		C1	C2	C3		D1	D2
A1	133	130	119	122		161	181	162		238	266
A2	119	122	133	130		175	155	174		266	238

Sample distribution: 4 experiments, stratified random allocation

	B1	B2	B3	B4		C1	C2	C3		D1	D2
A1	126	126	126	126		168	168	168		252	252
A2	126	126	126	126		168	168	168		252	252

With modest sample sizes, allocation cannot be fully controlled by design for more than a few factors. In the *Understanding Society* Innovation Panel, with more than forty experiments, any one experiment is likely to be unbalanced with respect to the majority of other experiments. In this situation it becomes important to identify the experiments that are most likely to affect the outcome of interest and to at least stratify the allocation with respect to those experiments. For example, at wave 4 eleven new experimental manipulations were to be introduced. Fully-crossing all eleven would have been impossible as this would lead to 45,056 experimental groups with a sample size of 2,445 to allocate. Instead, full orthogonality was restricted to subsets of the eleven experiments that were likely to influence the same outcomes. Four of the experiments were explicitly designed to influence unit nonresponse rates (Burton 2012). A fully-crossed experimental design was used to allocate sample units to these experiments, to ensure that the impact of each on unit nonresponse could be separately identified. As the experiments had eleven, four, two and two treatments respectively, this involved randomly assigning the 2,445 units to 176 groups. While the other seven experiments could conceivably have had some impact on unit nonresponse, this was felt to be unlikely as they mainly involved manipulations to question wording or question placement, designed to influence measurement. Similarly, for each of the seven measurement experiments, the allocation was crossed with two or three other experiments that could conceivably

have affected the measurement outcomes of interest. Complete confounding was avoided for every combination of experiments out of the eleven (and, indeed, with experiments carried at previous waves) by always assigning randomly within the groups defined by the crossing of experiments and making each assignment independently. However, for some of these combinations (the ones that were not expected to influence common outcomes) the sample distribution can be somewhat unbalanced due to the play of random chance.

In some cases the experiments that are amongst the most likely to influence a particular outcome of interest may be ones that were carried at an earlier wave. An example is a set of five experiments concerned with measurement of change between waves. These particular experiments were introduced at wave 3 and outcome measures were comparisons of responses given at wave 3 with those given at wave 2. However, wave 2 had involved an experimental allocation to mode treatments and the mode could have affected the response given at wave 2 to some of the relevant questions. The allocation to the wave 3 measurement of change experiments was therefore fully crossed with the allocation to the wave 2 modes experiment.

7. Allocation Procedures

In this section we discuss two other important considerations regarding the allocation of sample units to treatments. Given that survey samples are often hierarchically structured in some way, the first consideration concerns the choice of the level at which to assign units to treatments. The second consideration comes into play when an experiment involves treatment at multiple waves. The researcher must decide how to allocate sample units to multi-wave combinations of treatments (design (7) in table 1).

Assignment within or between households and interviewers?

In the *Understanding Society* Innovation Panel, sample units are individual persons, but these are clustered within households, and households are clustered in turn within both primary sampling units (PSUs) and interviewer assignments (which are strongly correlated with each other, but not identical). It is possible to allocate

experimental treatments at the level of PSU, interviewer, household or individual. Each may have advantages and disadvantages, depending on the nature of the experiment. In principle, statistical power is greatest when allocation is made at the lowest level (individuals, in our case) for reasons that are analogous to those set out in the previous section regarding confounding and balance. Allocating individuals to treatments, with stratification by household and PSU, will ensure that all PSUs (and as many households as possible) are represented in each treatment group, so that the power to observe a treatment effect is not reduced by systematic differences between PSUs. However, there are a number of reasons why allocating at a higher level will sometimes be preferable.

The effects of some treatments may be contaminated if respondents are aware that some people received different treatments to them. In the case of the *Understanding Society* Innovation Panel (as with several of the other surveys outlined in section 4 above) the clustering of individuals within households makes it quite likely that many respondents will be aware of the treatment received by other household members – at least for some types of experiments – so contamination effects are a serious concern. For that reason, most experiments have been allocated at the household level, so that all respondents in the same household receive the same treatment. For example, in the online/CAPI mode experiment introduced at wave 5 and described in section 5 above, the reaction of an individual to the single-mode face-to-face protocol might be different if they knew that someone else in their household was offered an opportunity that they themselves were not offered, to complete the survey online. Furthermore, in this example the treatment of interest to the experiment proposers (the scientific leadership team of *Understanding Society)* was inherently a household-level treatment. They wished to compare the existing single-mode design with one designed to reduce survey costs by getting a proportion of households to participate entirely online. This specific objective of getting everyone in the household to participate online can only be achieved if everyone in the household is invited to participate online, so a design in which only some household members receive such an invite was simply not of interest. There are other examples of treatments that are inherently household-level. One of these is part of a series of experiments on respondent incentives: one treatment involves offering each individual in the household an additional incentive payment conditional on every

individual in the household participating. The motivation for assessing this treatment is that it might increase the proportion of sample households for which data is successfully obtained from every individual.

There are at least two situations in which allocation to treatment may be best done at the level of interviewer assignment, rather than allocating households within each assignment. First, some experimental manipulations must be administered by the interviewer, such as when the interviewer has to present respondents with alternative versions of survey materials. In this situation, interviewers are less likely to make mistakes and the administration is more likely to be smooth and efficient if each interviewer only has one version to administer to all his or her respondents. The *Understanding Society* Innovation Panel learned this lesson the hard way at wave 1 with an experiment in which show cards were to be shown to half the sample, but not the other half. Assignment to treatments was crossed with interviewers, to minimise any interviewer effect on the results, but it turned out that many interviewers, once equipped with a set of show cards, found it hard to remember that they should not always hand the cards to the respondent at the start of the interview (the usual procedure on the survey and on most other surveys).

The second situation in which allocation to treatment may be best done at the level of interviewer assignment is when one or more of the treatments is specifically designed to change interviewer behaviour in some way. An example would be any treatment that should affect calling patterns in a face-to-face survey. Calls are not made independently for each unit in an assignment. Rather, an interviewer will often make additional call attempts while they are in the area visiting other sample units. Thus, for the treatment to be a realistic replication of how it would work if applied to a whole survey, all units in an interviewer assignment should receive the same treatment. At wave 4 of the *Understanding Society* Innovation Panel an experiment was run in which a proportion of sample households were offered, via the advance letter, the opportunity to telephone their interviewer to make an appointment at a convenient time rather than waiting for the interviewer to visit them. Brown and Calderwood (2014) found only a very small reduction in the number of interviewer calls required to complete the interviews - a finding which could have been affected by the fact that each interviewer assignment included some treated cases and some control cases.

Switching treatments between waves

The longitudinal survey context gives researchers the possibility of mounting experiments over two or more waves. There are a number of situations in which this can be desirable, and a number of possible multi-wave designs. For example, each sample unit could continue to receive the same treatment at each wave; each sample unit could switch from one treatment to another at the next wave; or treatments could be assigned randomly at each wave, without regard to the treatment assigned previously. The most appropriate choice should depend on the objectives of the experiment and the extent to which errors in the outcome variable(s) are likely to be correlated between waves.

For example, suppose there are two alternative treatments, A and B, to be compared. Using a within-subject design controls the between-subject component of variance and hence improves the accuracy of estimates of the effect of B rather than A. If the treatments are of a kind that cannot be both administered in the course of the same interview, a within-subject option in a longitudinal survey context is to administer one treatment at one wave and the other at the next wave. However, if all sample members are administered treatment A at wave t and treatment B at wave t+1, this risks confounding the relative effects of treatment B with a) real change in the outcome between t and t+1, and b) a 'priming' effect caused by having previously been administered treatment A. To avoid such confounding, a crossover design can be used. In a crossover design, one group of respondents would receive treatment A at wave t and treatment B at wave t+1, while another group would receive treatment B at wave t and treatment A at wave t+1. If the errors are uncorrelated between waves (or, more realistically, have very low correlation) this design should maximise the precision of estimates of the effect of B rather than A.

For some research questions, the relevant treatments are themselves inherently longitudinal. Consider, for example, the choice of question wording or response options for a question that is to be repeated at each wave of a longitudinal survey for the purpose of measuring change. Suppose there are two candidate versions of a question, labelled A and B. The researcher wishes to know whether it is better to repeat version A at each wave or to repeat version B at each wave. If the purpose is to inform development of a new survey, with no *a priori* reason for preferring either

question version, a simple repeated test design (type 7 in table 1) could be implemented, in which one group is asked version A at each wave and another group is asked version B. If the experiment shows that repeated use of version B is superior, the survey will adopt that version. But what if the experiment is to inform an existing survey, in which version A is currently used? The researcher might also need to assess the effect of transitioning from version A to version B. Thus, a third treatment group could be introduced in which version A is administered initially, with the treatment switching to version B after a number of waves.

Sometimes, the accuracy of both cross-sectional and longitudinal measures is of importance. If the survey questions in the example of the previous paragraph are about the level of savings held by a household, the answers could be used either to construct a (cross-sectional) measure of current savings or a (longitudinal) measure of change in savings since the previous wave. Even if the survey designs under consideration are only those that involve repeating the same question at each wave, a crossover design might provide a more accurate estimate of the quality of the cross-sectional measure.

Recognising the competing design implications of different research objectives, an experiment carried at waves 3 and 4 of the *Understanding Society* Innovation Panel involved four treatment groups: one group was administered version A at both waves, one was administered version B at both waves, one was administered version A at wave 2 and version B at wave 3, and the final group was administered version B at wave 2 and version A at wave 3. The experiment concerned several measures of change. For example, one of the questions involved was designed to ascertain for how long the respondent had lived at their current address. Version A of this question asked (of people who had not lived at their current address their whole life) “In what month and year did you move to this address?” while version B asked “How long have you lived at this address?” The partial crossover in the experimental design enhances the precision of estimates of differences between the two question versions in cross-sectional measures, while the two simple repeated test treatments allow comparison of measures of change when either of the questions versions is repeated. A similar four-treatment design was used for an experiment at waves 2 and 3 in which show cards were used with half of the respondents at each wave.

8. Refreshment Samples

The *Understanding Society* Innovation Panel introduces an additional sample, known as a refreshment sample[2], each three years. To date, refreshment samples have been added at waves 4, 7 and 10. The main reason for doing this is to maintain the size of the panel, but an additional advantage is that the practice adds an extra dimension to experiments mounted on the panel.

Time in sample may affect respondents' familiarity with the survey questions, trust in the interviewer/survey, and knowledge of the topic(s) of the survey. These changes may affect the responses that are given to survey questions, producing 'panel conditioning' (Struminskaya 2015; Warren and Halpern-Manners 2012). Furthermore, sample members at later waves of a panel may tend to be easier to contact and more co-operative than at earlier waves (Uhrig 2008; Watson and Wooden 2012). Thus, the results of any experiment designed to affect either survey responses or fieldwork outcomes may depend on the survey wave at which the experiment is conducted. The strength of a design with regular refreshment samples is that the extent and nature of dependency on survey wave can be estimated by comparing outcomes between samples. For example, an experiment mounted at wave 8 will be administered to three samples consisting of respondents for whom it is their second, fifth and eighth wave of participation. This strength was exploited in the analysis of an experiment with targeted advance letters carried out at wave 6 of the *Understanding Society* Innovation Panel (Lynn 2016). In the CAPI single-mode part of the sample, the targeted letters improved response rates significantly for the refreshment sample (who had only participated in two previous waves) but not for the original sample (who had participated in five previous waves).

Another strength of a design with regular refreshment samples is that it may be possible to control for any possible exposure of respondents to relevant stimuli or experimental treatments at previous waves. For example, an experiment regarding ways to introduce a new, sensitive, topic to respondents clearly requires a context in

[2] On other surveys, similar additional samples are sometimes known as replenishment samples, refresher samples, or top-up samples.

which the respondents have not previously been asked questions on that topic. Having run such an experiment once, the findings may suggest a further line of enquiry that would require further experimentation. But further experimentation on the same sample would not provide a realistic setting. The existence of a new refreshment sample on whom the first experiment had not been administered would provide an opportunity for the second experiment to be carried out in broadly the same context as the first (same survey).

9. Discussion

Randomised experiments mounted on probability-based longitudinal surveys have considerable strengths. The randomisation provides internal validity, while probability sampling provides external validity. In this high-validity context, the longitudinal design provides opportunities to study dynamics in both the outcomes of experimental treatments and in the treatments themselves, as well as opening up the possibility of both treatments and analysis being cognisant of past experiences, prior characteristics or even past survey behaviour. The range of design types that are possible with a longitudinal survey context are outlined in section 2 of this chapter. A particular strength of the longitudinal survey context is the variety of repeated test designs that are possible, defined by whether and how treatments are varied within participants across waves (discussed in section 7). Another advantage is the potential provided by regular refreshment samples (section 8) to study the effect of time-in-sample and to control for previous exposure to similar treatments. Many examples of experimental studies that take advantage of these multiple strengths can be found amongst the experiments that have been mounted on the *Understanding Society Innovation Panel* (described in section 5).

However, these strengths come at the price of challenge and complexity in design and implementation. The complexity increases when multiple independent experiments are to be carried on the same survey, and over many waves. This chapter has outlined some of these challenges and complexities and has demonstrated some of the ways in which the challenges can be met in order to ensure the success of the experiments. A key design objective is to avoid confounding between experiments and to maximise the statistical power of

experiments. In the context of longitudinal surveys such as those described in section 4 of this chapter, this is particularly challenging because of the large number of experiments carried on the same survey and because of the evolving nature of the experimentation: experiments in later waves are not yet conceptualised at the time of the design of earlier waves. In section 6 we have described how confounding can be avoided through the use of stratified random allocation, where the treatment groups for other experiments constitute the strata. We have also discussed (section 7) issues involved in choosing the level at which randomised allocation should take place (primary sampling unit, interviewer, household or individual). There is a balance to be struck between statistical and practical considerations and we have mentioned examples that demonstrate why neglecting the latter will not necessarily benefit the former.

Truly longitudinal experimentation on probability-based longitudinal surveys is still an evolving methodology. There are relatively few longitudinal surveys designed for this purpose and there is very little literature on design issues. Research would benefit from further study of the relative advantages of different longitudinal designs for different analytical purposes. For example, there is little guidance to be found on when a simple crossover design should be preferred to a crossover design with repeated-treatment groups, or how best to determine the optimum group sizes in the latter design. Analysis methods too are under-developed: for example, standard error estimation that takes into account stratified random allocation to treatments within a survey with a complex design. While the strengths of randomised experiments mounted on probability-based longitudinal surveys are truly considerable, work remains to ensure that study designs can take full advantage of these strengths.

Appendix: Stata syntax to produce table 3 treatment allocations

Simple random allocation to two groups:

```
ge rand=runiform()
sort rand
ge treatA=1 if trunc((_n-1)/2)==trunc(_n/2)
recode treatA .=2
```

Stratified random allocation, where stratification is by the treatment groups for three other experiments:

```
ge rand=runiform()
sort treatB treatC treatD rand
ge treatA=1 if trunc((_n-1)/2)==trunc(_n/2)
recode treatA .=2
```

References

Aaker, Jennifer, Susan Fournier, and S Adam Brasel. 2004. "When Good Brands Do Bad." *Journal of Consumer research* 31(1):1-16.

Acredolo, Linda P. 1978. "Development of Spatial Orientation in Infancy." *Developmental Psychology* 14(3):224.

Addelman, Sidney. 1969. "The Generalized Randomized Block Design." *The American Statistician* 23(4):35-36.

Aguila, Emma, Arie Kapteyn, and James P Smith. 2015. "Effects of Income Supplementation on Health of the Poor Elderly: The Case of Mexico." *Proceedings of the National Academy of Sciences* 112(1):70-75.

Al Baghal, Tarek. 2015. "Understanding Society Innovation Panel Wave 7: Results from Methodological Experiments." in *Understanding Society Working Paper 2015-03*. Colchester: University of Essex.

—. 2016. "Understanding Society Innovation Panel Wave 8: Results from Methodological Experiments." Colchester: University of Essex.

Al Baghal, Tarek (ed.). 2014. "Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments." in *Understanding Society Working Paper 2014-4*. University of Essex.

Al Baghal, Tarek, and Annette Jäckle. 2016. "Understanding Society: The UK Household Longitudinal Study Innovation Panel, Waves 1-8, User Manual." Colchester: University of Essex.

Belfield, Clive R, Milagros Nores, Steve Barnett, and Lawrence Schweinhart. 2006. "The High/Scope Perry Preschool Program Cost–Benefit Analysis Using Data from the Age-40 Followup." *Journal of Human Resources* 41(1):162-90.

Bianchi, Annamaria, Silvia Biffignandi, and Peter Lynn. in press. "Web-CAPI Sequential Mixed Mode Design in a Longitudinal Survey: Effects on Participation Rates, Sample Composition and Costs." *Journal of Official Statistics*.

Blom, Annelies G, Michael Bosnjak, Anne Cornilleau, Anne-Sophie Cousteaux, Marcel Das, Salima Douhou, and Ulrich Krieger. 2015. "A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe." *Social Science Computer Review* 34(1):8-25.

Boisjoli, Rachel, Frank Vitaro, Eric Lacourse, Edward D Barker, and Richard E Tremblay. 2007. "Impact and Clinical Significance of a Preventive Intervention for Disruptive Boys." *The British Journal of Psychiatry* 191(5):415-19.

Bolton, Ruth N, and James H Drew. 1991. "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes." *The Journal of Marketing* 55(1):1-9.

Bosnjak, Michael, Tanja Dannwolf, Tobias Enderle, Ines Schaurer, Bella Struminskaya, Angela Tanner, and Kai W Weyandt. 2017. "Establishing an

Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel." *Social Science Computer Review*:preprint.

Brown, C Hendricks, and Jason Liao. 1999. "Principles for Designing Randomized Preventive Trials in Mental Health: An Emerging Developmental Epidemiology Paradigm." *American journal of community psychology* 27(5):673-710.

Brown, Matt, and Lisa Calderwood. 2014. "Can Encouraging Respondents to Contact Interviewers to Make Appointments Reduce Fieldwork Effort? Evidence from a Randomized Experiment in the UK." *Journal of Survey Statistics and Methodology* 2(4):484-97.

Brown, Matt, and Maggie Hancock. 2015. "National Child Development Survey. 2013 Follow-Up: A Guide to the Datasets." London: Institute of Education.

Burton, Jonathan. 2012 "Understanding Society Innovation Panel Wave 4: Results from Methodological Experiments." in *Understanding Society Working Paper 2012-06*. Colchester: University of Essex.

—. 2013. "Understanding Society Innovation Panel Wave 5: Results from Methodological Experiments." in *Understanding Society Working Paper 2013-06*. Colchester: University of Essex.

Burton, Jonathan, Sarah Budd, Emily Gilbert, Annette Jäckle, Stephanie McFall, and SC Noah Uhrig. 2011. "Understanding Society Innovation Panel Wave 3: Results from Methodological Experiments " in *Understanding Society Working Paper 2011-05*. Colchester: University of Essex.

Burton, Jonathan, Heather Laurie, and SC Noah Uhrig. 2008. "Understanding Society: Some Preliminary Results from the Wave 1 Innovation Panel." in *Understanding Society Working Paper 2008-03*. Colchester: University of Essex.

—. 2010. "Understanding Society Innovation Panel Wave 2: Results from Methodological Experiments." in *Understanding Society Working Paper 2010-04*. Colchester: University of Essex.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.

Chakraborty, Bibhas, and Susan A Murphy. 2014. "Dynamic Treatment Regimes." *Annual review of statistics and its application* 1:447-64.

De Vaus, David A. . 2001. *Research Design in Social Research*. London: Sage.

Dupas, Pascaline, and Jonathan Robinson. 2013. "Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya." *American Economic Journal: Applied Economics* 5(1):163-92.

Dvir, Taly, Dov Eden, Bruce J Avolio, and Boas Shamir. 2002. "Impact of Transformational Leadership on Follower Development and Performance: A Field Experiment." *Academy of management journal* 45(4):735-44.

Ellickson, Phyllis L, and Robert M Bell. 1990. "Drug Prevention in Junior High: A Multi-Site Longitudinal Test." *Science* 247(4948):1299-305.

Farrington, David P., Rolf Loeber, and Brandon C. Welsh. 2010. "Longitudinal-Experimental Studies." Pp. 503-18 in *Handbook of Quantitative Criminology*, edited by Alex R. Piquero and David Weisburd: Springer New York.

Hu, Paul Jen-Hwa, Wendy Hui, Theodore HK Clark, and Kar Yan Tam. 2007. "Technology-Assisted Learning and Learning Style: A Longitudinal Field Experiment." *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 37(6):1099-112.

Jäckle, Annette, Peter Lynn, and Jon Burton. 2015. "Going Online with a Face-to-Face Household Panel: Effects of a Mixed Mode Design on Item and Unit Non-Response." *Survey Research Methods* 9(1):57-70.

John, J. A., and M.H. Quenouille. 1977. *Experiments: Design and Analysis*. High Wycombe: Charles Griffin & Co. Ltd.

Lee, Min Kyung, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. "Personalization in HRI: A Longitudinal Field Experiment." Pp. 319-26 in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*: IEEE.

Lynn, Peter. 2009. "Sample Design for Understanding Society." in *Understanding Society Working Paper 2009-01*. Colchester: University of Essex.

—. 2013. "Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs." *Journal of Survey Statistics and Methodology* 1:183-205.

—. 2016. "Targeted Appeals for Participation in Letters to Panel Survey Members." *Public Opinion Quarterly* 80(3):771-82.

Marcus, Alfred C. 1982. "Memory Aids in Longitudinal Health Surveys: Results from a Field Experiment." *American journal of public health* 72(6):567-73.

Maxwell, Scott E., and Harld D. Delaney. 2004. *Designing Experiments and Analysing Data: A Model Comparison Perspective*. London: Lawrence Erlbaum Associates.

McCord, Joan. 2003. "Cures that Harm: Unanticipated Outcomes of Crime Prevention Programs." *The Annals of the American Academy of Political and Social Science* 587(1):16-30.

Olds, David L, Harriet Kitzman, Robert Cole, Joann Robinson, Kimberly Sidora, Dennis W Luckey, Charles R Henderson, Carole Hanks, Jessica Bondy, and John Holmberg. 2004. "Effects of Nurse Home-Visiting on Maternal Life Course and Child Development: Age 6 Follow-Up Results of a Randomized Trial." *Pediatrics* 114(6):1550-59.

Pierret, Charles R. 2001. "Event History Data and Survey Recall: An Analysis of the National Longitudinal Survey of Youth 1979 Recall Experiment." *Journal of Human Resources* 36(3):439-66.

Richter, David, and Juergen Schupp. 2015. "The SOEP Innovation Sample (SOEP IS)." *Schmollers Jahrbuch* 135:389-400.

Struminskaya, Bella. 2015. "Respondent Conditioning in Online Panel Surveys: Results of Two Field Experiments." *Social Science Computer Review* 34(1):95-115.

Uhrig, S. C. Noah. 2008. "The Nature and Causes of Attrition in the British Household Panel Survey." in *ISER Working Paper 2008-05*. Colchester: University of Essex.

Warren, John Robert, and Andrew Halpern-Manners. 2012. "Panel Conditioning in Longitudinal Social Science Surveys." *Sociological Methods & Research* 41(4):491–534.

Watson, Nicole, and Mark Wooden. 2012. "The HILDA Survey: A Case Study in the Design and Development of a Successful Household Panel Study." *Longitudinal and Life Course Studies* 3(3):369-81.

Wiedenbeck, Susan, Jim Waters, Jean-Camille Birget, Alex Brodskiy, and Nasir Memon. 2005. "PassPoints: Design and Longitudinal Evaluation of a Graphical Password System." *International Journal of Human-Computer Studies* 63(1):102-27.

Workman, Michael, and William Bommer. 2004. "Redesigning Computer Call Center Work: A Longitudinal Field Experiment." *Journal of Organizational Behavior* 25(3):317-37.

Yeager, David S, Adriana S Miu, Joseph Powers, and Carol S Dweck. 2013. "Implicit Theories of Personality and Attributions of Hostile Intent: A Meta-Analysis, an Experiment, and a Longitudinal Intervention." *Child development* 84(5):1651-67.

Table 2: Overview of longitudinal surveys that regularly field experiments

Survey	UKHLS Innovation Panel (IP)	SOEP Innovation Sample (SOEP-IS)	LISS panel	GESIS Panel	ELIPSS	American Life Panel	Understanding America Study (UAS)
Funder	Economic and Social Research Council; non-standard elements funded by proposer	Leibniz Association**;** non-standard elements funded by proposer	Costs paid by proposers of experiments	Leibniz Association	Agence Nationale de la Recherche	Costs paid by proposers of experiments	Costs paid by proposers of experiments
Geographical coverage	Great Britain	Germany	Netherlands	Germany (German speaking population)	France (metropolitan area)	U.S.A.	U.S.A.
Sample units	All members of sampled households	All members of sampled households and new members	All members of sampled households	Individuals	Individuals	Individuals, but other household members invited to participate	All members of sampled households followed over time
Sample design	Clustered, stratified sample based on postal addresses	Random route	Probability sample of households from population register	Random sample drawn from municipal population registers	Two stage random sample of individuals in households listed in 2011 census	Sample members recruited from multiple sources with different probability sample designs	Clustered, stratified sample of households based on postal addresses
Frequency	Annual since 2008	Annual since 2011	Monthly since 2007	Every two months since 2014	Monthly since 2012	Once or twice a month since 2006	Once or twice a month, depending on demand
Modes	CAPI; mixed mode experiments with CATI and Web	CAPI; experiments with smartphones and web	Web; households without internet access are loaned computer and broadband connection	Web and mail; paper questionnaire sent to those unable/unwilling to participate online	Web; participants are loaned a tablet with mobile internet connection	Web; households without internet access are loaned computer and broadband connection	Web; households without internet access are loaned computer and broadband connection
Website	https://www.understandingsociety.ac.uk/about/innovation-panel	www.diw.de/soep-is	http://www.lissdata.nl/lissdata/Home	http://www.gesis.org/en/services/data-collection/gesis-panel/	http://quanti.dime-shs.sciences-po.fr/en/	https://alpdata.rand.org	https://uasdata.usc.edu/surveys