# Common themes in design and analysis of longitudinal surveys

Paul A. Smith

S3RI and Dept of Social Statistics & Demography, University of Southampton

July 2019

This work was funded by the Economic & Social Research Council under grant no. ES/T001038/1.

# Common themes in design and analysis of longitudinal surveys

Paul A. Smith
S3RI/Dept of Social Statistics & Demography, University of Southampton

**Executive Summary**

Some common themes emerged from the scoping studies, and are discussed in more detail.

The UK has world-leading longitudinal data resources. They have accumulated as a result of a series of design decisions as new studies have been added in ways that made the best use of the resources available at the time. The system as a whole is not optimised with clear objectives; such a system should include elements for coverage of smaller subpopulations. Research into design and optimisation of a system of longitudinal survey resources should be commissioned, and is potentially valuable in providing evidence to support long-term funding.

Several longitudinal survey options rely on the combination of information from different types of surveys. Microdata linkage is a well-known topic, but how to account for linkage errors is an area of active research, which should be supported. Additionally, data integration is needed where two datasets collected with different designs or procedures need to be put together. How to integrate data in these circumstances, and methods to analyse the resulting data in an objective way, also need exploration and case studies.

**Introduction**

The ESRC commissioned a series of scoping studies on different aspects of longitudinal surveys as a follow-up to the recommendations of the Longitudinal Studies Review (Davis-Kean et al. 2017). Six such topics were considered by the University of Southampton (refs?). While each aspect presented its own challenges, it quickly became clear that there were several overarching themes, which influenced the design and methods in multiple places. It seemed sensible to add some additional commentary about these issues.

This document covers the challenges in designing and maintaining a system of longitudinal survey resources, the interplay between cohort studies and focussed studies of subpopulations (noting that this also affects the system of surveys), combining different sources for analysis, and what population we are able to make inferences about.

**Design and optimisation of a system of longitudinal data resources**

Several of the questions considered in these short studies relate to how best to gather information in a way that provides the required range of variables, with appropriate quality and frequency, from a range of longitudinal survey and administrative sources. Each individual question is a reasonable inquiry in its own right, but an ideal approach would be to group many of these together with the aim of optimising a *system* of resources. Optimising across surveys is a topic which has not been widely researched. Occasional government reviews have considered whether the range of surveys is appropriate, and whether the division of sample resources between them is appropriate – for example Penneck *et al*. (1993) considered whether sample sizes in official business surveys in the UK were appropriately allocated for the production of the national accounts. The national accounts provided a unifying framework in this case which is missing in the social survey space, but nonetheless the

expected accuracy for a range of variables, associations and analyses could be compared to investigate whether the arrangement of surveys meets the range of needs.

Indeed, the current suite of surveys has clearly not been designed in isolation, and the links between BHPS and Understanding Society are obvious. The intentions of setting up national cohort studies approximately every ten years means that considering studies together implicitly generates an accelerated longitudinal design (ALD) over specific periods, without explicitly being designed in this way. Analysing two successive cohorts in this fashion is possible. But by designing future cohort studies with accelerated designs, it may be possible to link better between cohorts (which will not be so widely spaced) and have a wider palette of analyses. Further, a *regular* spacing of cohorts backed up by design considerations indicating the benefits of new data collection and consistency might make the task of finding funding more straightforward. Regular spacing is also likely to simplify analyses, and this simplicity is likely to increase the effective use of the infrastructure.

Such a design need not restrict itself to the current arrangement. There are competing calls for many different kinds of data to be included in the longitudinal studies, and this can lead to a heavy burden on respondents. Countries are considering their systems of social surveys and how they can be reorganised to maintain response and collect the range of data needed (see Karlberg et al. (2015), Ioannidis et al. (2016)). The same kinds of thinking can be applied to a system of longitudinal surveys, to define what questions should be included, and with what frequency. This approach is already applied in Understanding Society, which follows a detailed topic plan (Davis-Kean *et al*. 2017, section 2.1). But there has not been the same sort of approach to defining the priority for topics to be included in different surveys in the longitudinal portfolio, at what times or stages, and with how much consistency. The design process needs to account for the types of analysis that are likely to be required – Chipperfield (2016) and Ioannidis *et al*. (2016) in discussion point out that certain higher-order interactions might not be estimable in some designs, but that it might be extremely challenging to estimate these statistics in a usable way. It would nonetheless be valuable to consider to what extent different designs would restrict the analysis that could be done. This could be extended to examining what constraints on analysis are imposed by the current portfolio of surveys and topics, to provide a comparator. (The challenges of the effect of data integration on associations resurface in the sections on *Linkage* and *Combining different data sources* below).

There is a series of methodological challenges with this sort of integration. There is a clear need for harmonisation, and for a process for agreeing across a range of surveys a suitable data collection procedure. Then the datasets must be merged (not linked at the microdata level, since there should be (almost?[1]) no overlap) to map the different variables and timings to form an analysis dataset. Then, not least of the challenges, the methods for analysing such a dataset and dealing with the designed presence or absence of data, and the different periodicities, need to be developed and made available to researchers in as easily implementable a form as possible. Some training in the use of these approaches is likely to be needed, particularly to avoid drawing false conclusions from inappropriately simple analysis.

Dolson (2016) in discussing Ioannidis *et al*.'s paper points out that the benefits of survey integration are only gradually realized, and the long-term nature of longitudinal surveys makes the return on the investment potentially even longer term (though some of the changes, for example through harmonised question sets and procedures, can be implemented much more quickly and the benefits realised on a shorter timescale). Arguably, changes in longitudinal studies can be managed more effectively because there is already good information on the respondents from previous waves which can be used to design an effective change process and adjust for any differences afterwards.

[1] A possibly interesting, though potentially sensitive, analysis would be to examine whether anyone has been selected for multiple longitudinal surveys in the UK.

**Recommendation**: Commission research into the optimisation of resources with respect to outputs across a range of longitudinal surveys. Evaluate the current portfolio against this to identify areas most in need of further development. This project could include the following recommendation as one element.

**Recommendation**: Commission research into extending the idea of modular design to a system of longitudinal surveys. Explicitly cover both the design and analysis perspectives.

A designed system of longitudinal surveys could be less vulnerable to shifts in funding. At the least, a designed system would provide the information on the effects of funding changes on key objectives and outputs, and therefore the evidence to support continued funding.

**Cohorts, minority groups and domains**

There is an interrelationship between the discussion of the competing benefits of cohort and panel studies in WP4 (Lugtig & Smith 2019) and the arguments about sample boosting for minority groups or small populations in WP2 (Smith 2019). Panel studies cover the general population, and may be oversampled for particular groups of interest, particularly where higher rates of attrition are expected. Cohort studies relax one element of the panel design, namely the need for coverage of the whole population, in exchange for the ability to target the topic coverage to particular characteristics and still provide sufficient sample sizes for subgroup analyses. In this sense a cohort study looks like a separate focussed study of a minority population, which in this case is a specific cohort. Separate samples for minority groups also relax the requirement of complete population coverage, but in a different fashion where elements of the population are excluded on characteristics other than age; we assume that they are also undertaken longitudinally to provide the same sorts of information.

These survey choices also contribute a dimension to the system of longitudinal surveys. It becomes challenging to manage a system consisting of a large number of separate studies. The challenges of integrating the different studies for particular analyses would need to be solved afresh for each study (in combination with other resources). So there is a simplicity argument for making national panel and cohort studies sufficiently large (combined with a manageable level of oversampling) to give adequate sample sizes for a defined set of subpopulations of particular interest. It is difficult to make general recommendations on the sample size requirements without specific details of the objectives, however. Large studies offer the flexibility to analyse many subpopulations because they have sufficient size, but it is challenging to recruit enough participants in a large sample, as was shown by the Life Study.

Both cohort studies and separate focussed panel studies present interesting analysis challenges if they are to be related to a general population survey. Coordination of core topics is required, so that they can be used to compare the characteristics of participants in the different types of studies, and be used as additional variables to include in models to help to compensate for differences during analysis. Data integration is described in more detail below.

**Combining different data sources for analysis**

There are a number of situations where data arise from two different but related sources, and could be combined for analysis:

- in the Life Study (Dezateux et al. 2016) there were two sample components, one a national probability sample, and one from location sampling - recruitment at specified locations;
- there is a range of options for non-probability designs, and some offer relatively inexpensive ways to gather data; in principle these extra cases could be added to probability based samples, although it is not clear how much users of the data would benefit (WP5)

- separate studies for particular subpopulations may be better undertaken with their own designs and objectives, but where there are common variables, it would be advantageous to combine these with a national probability sample, possibly to add extra cases for the minority group, but particularly to make inferences about differences between the subpopulation and main populations in a rigorous way (WP2)

These three approaches all suggest that models for combining data from probability and non-probability sources are needed, ideally 'borrowing strength' and taking the best properties of each. One strategy would be to focus on producing a minimum mean squared error estimator, but it is not clear how this would be constructed in a general case, and whether guidance and/or software to assist analysts with these approaches could be developed and made available. Some case studies illustrating how such models could be constructed would be valuable for researchers.

A second strategy would be to link all the data through an administrative data spine, which might give access to common variables measured in a consistent manner and facilitate the building of models where relationships can be analysed through such an intermediary variable.

**Recommendation:** more research into models for combining and analysing data of different types (including through an administrative data spine) is needed, together with some case studies.

At a microdata level, linkage is a critical procedure for creating datasets which cover a range of topics, particularly in the case where associations between variables are the target of inference. A high quality linkage process is important to ensure that a linked dataset contains all the appropriate records with complete observations, and no erroneous linkages. A failure to link records belonging to the same entity may act like differential sampling to reduce the numbers of certain types of records in the datasets, and lead to biased analyses. Linking records that do not belong to the same entity will create records where variable values will be associated when they should not be, and similarly lead to bias. Where both biases are present they are likely to be cumulative rather than offsetting. Microdata linkage is particularly important when administrative data are being used to create a spine, and when survey data are being linked to administrative data (through a spine or not).

Various methods are available for analysing linked datasets accounting appropriately for the linkage error (see Work Package 3), and in order to facilitate their application it is important that metadata and paradata about the linkage process are recorded and made available to researchers alongside the linked data (Gilbert *et al*. 2017).

The use of administrative data in design (WP5 & 6) can produce savings by reducing variability for variables and models of interest. At the design stage there is less requirement for the matching to be of very high quality – as long as the matched administrative variable is a good predictor of the variable(s) of interest, its use will improve the design. But if the same data are then used in analysis, rather than validated or collected at interview, then the methods from WP3 would be needed.

**What population can we make inferences about?**

This is a challenging definitional problem in several contexts. A longitudinal study by definition covers people (or other sample units) at more than one time, and the population to which they belong evolves. This leads to several ways of defining the 'longitudinal population' (Smith *et al*. 2009). The cohort approach is often the simplest, because we define the population of interest as a specific cohort, and do not allow it to increase (WP4). Panels however can increase by application of the following rules, and then we can analyse in terms of the initial, final, ever-present or always-present population. Which choice to make will often depend on the question of interest, but may also be driven by data availability together with suitable assumptions.

This situation becomes further complicated with administrative data, which is updated dynamically, and may exist in a system which has a range of different vintages (WP1). It may be possible to use some methods designed to adjust for measurement error to account for these differences, for example latent class models (WP3).

There are also challenges with network sampling approaches where the population is unknown and there are no calibration totals. IN these cases the stability of the population is difficult to assess, and may be important in order to make an objective analysis of the question of interest. Some follow-up of study members who leave the key population of interest may be needed in order to get a good estimate of the rate of change of the population, but this may be more difficult to achieve if the study becomes less salient. And in a dynamic population it will be necessary to add people who join the population, necessitating regular top-up or boosts.

**References**


Chipperfield, J. (2016) Discussion [of Ioannidis *et al*. (2016)]. *Journal of Official Statistics* **32** 287-289.

Davis-Kean, P., Chambers, R.L., Davidson, L.L., Kleinert, C., Ren, Q. & Tang, S. (2017) *Longitudinal Studies Strategic Review*. Report to the Economic and Social Research Council.

Dezateux, C., Knowles, R., Brocklehurst, P., Elias, P., Burgess, S., Colson, D., Elliot, P., Emond, A., Goldstein, H., Graham, H., Kelly, F., Kiernan, K., Leon, D., Reay, D., Sera, F., Vignoles, A. & Walton, S. (2016) Life Study Scientific Protocol. (Life Study Working Papers). Life Course Epidemiology and Biostatistics/ UCL Institute of Child Health: London, UK. http://discovery.ucl.ac.uk/1485668/.

Dolson, D. (2016) Discussion [of Ioannidis *et al*. (2016)]. *Journal of Official Statistics* **32** 291-294.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.-C., Smith, P., Dibben, C. & Goldstein, H. (2017) GUILD: GUidance for Information about Linking Data sets. *Journal of Public Health* **40** 191-198. doi:10.1093/pubmed/fdx037.

Ioannidis, E., Merkouris, T., Zhang, L.-C., Karlberg, M., Petrakos, M., Reis, F. & Stavropoulos, P. (2016) On a modular approach to the design of integrated social surveys. *Journal of Official Statistics* **32** 259-286 and 301-305.

Karlberg, M., Reis, F., Calizzani, C. & Gras, F. (2015) A toolbox for a modular design and pooled analysis of sample survey programmes. *Statistical Journal of the IAOS* **31** 447-462.

Lugtig, P. & Smith, P. (2019) Work package 4: the choice between a panel and cohort study design. Working paper.

Penneck, S., Walker, C. & Wood, N. (1993) *CSO Business Inquiries: Scrutiny of Structure and Conduct*. Newport: CSO.

Smith, P., Lynn, P. & Elliot, D. (2009) Sample design for longitudinal surveys. Pp 21-33 in P. Lynn (ed) *Methodology of longitudinal surveys*. Wiley: Chichester.

Smith, P.A. (2019) Booster samples of marginal groups vs separate focussed studies. Working paper.