Booster samples of marginal groups vs separate focussed studies

Paul A. Smith

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

This work was undertaken in cooperation with NatCen Social Research

July 2019

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

Booster samples of marginal groups vs separate focussed studies

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

Executive summary

Recent longitudinal study designs have had samples which are sufficiently large to support a range of subgroup analyses. But for very specialised subpopulations, the sample sizes are too small. The design can be adjusted to over-represent these cases, but this has an impact on general analysis, particularly by increasing variability in the weights. It seems methodologically preferable to use standalone studies in these cases, though these would be vulnerable to funding changes.

Small studies should be harmonised as far as possible with the main panel/cohort, to facilitate combined analysis. How to combine samples potentially sampled in quite different ways to make the best use of the data is an open question needing further research.

Boost samples could be employed regularly in the main studies to replace sample losses from attrition and to increase coverage of subpopulations of particular interest. How to manage analysis in such a dynamic system also needs to be addressed, but it is not so far removed from a rotating panel design. Coordinated sampling might help to make fieldwork procedures more efficient.

A small-scale comparison of probability and non-probability sampling approaches for marginal groups in a UK context would be valuable; it should contain at least two waves to investigate the impact on sample retention.

1. Introduction

The UK has a well-developed longitudinal survey data infrastructure based on a series of cohort and panel studies which have been running since the 1940s. Many of the studies which started earlier have relatively modest sample sizes, but the Millennium Cohort Study and Understanding Society had substantially larger initial samples, 18,000 children and 30,000 households respectively. One of the main reasons for these larger sample sizes was to provide adequate numbers of cases for analyses of smaller or marginal groups within the population. In considering the future of longitudinal data provision, the ability to analyse information for smaller groups continues to be important, and we consider the options for how data on such groups may be collected and analysed.

In this document we take *marginal group* to mean a subpopulation whose defining characteristics are rare in the general population for whatever reason; see Kish (1991) for a taxonomy of such subpopulations. They are implicitly subpopulations of research or policy interest. We specifically consider how these subpopulations are included in longitudinal studies, to reflect the aspirations of the Longitudinal Studies Strategic Review (Davis-Kean et al. 2017).

One thing that is specific to subpopulations in a longitudinal survey is that they are not necessarily stable. It is clearly easier to deal with subpopulations defined by variables that do not change with time, though Kish (1991) points out that communities can retain their characteristics “in spite of constant changes due to migrations and vital processes”. And there will be occasions when subpopulations defined by unstable characteristics are of primary interest (eg poverty status, recent mothers). If the characteristics which define the marginal group change regularly then the longitudinal change of interest is in the categories as much as in the associated variables, and this then becomes a problem of estimating changes between states, for which the whole sample is needed.

Longitudinal studies are often used to generate and investigate hypotheses concerning particular subpopulations, since they provide nationally representative samples and many variables of interest to researchers. In designing a longitudinal study, however, it is challenging to provide for many subpopulations in the sampling, and this suggests two possible strategies for obtaining the required information for studies on particular subpopulations

- ensuring that nationally representative longitudinal studies have adequate numbers of cases for defined subpopulations through the sample design; and
- undertaking separate studies of subpopulations.

These approaches both have the same challenges over how to identify and sample adequate numbers of cases in the subpopulations of interest. The gold standard approach is to use a form of probability sampling (Lynn *et al*. 2017), and this is likely to be essential in data collected as a resource for a wide variety of research such as a cohort or panel study (see also WP5). For specific marginal groups, however, a general purpose sample may not include sufficient cases for a specific analysis, and some supplementary data collection to provide sufficient cases may be needed.

In the next two sections we consider the properties of these two approaches as they relate to data collection and analysis for subpopulations. In section 4 we consider strategies for sampling sufficient cases. In section 5 we widen the scope slightly to consider periodic boosts to panel surveys, and in section 6 consider how this might fit with a responsive design framework and how it might be facilitated by coordinated sampling in section 7. Finally in section 8 we consider the analysis of the data collected, before pulling the main points together in a discussion. The development of the ideas in this work package has been supported by a literature review, which is attached as Annex A, and contains further information.

2. Subpopulation samples within a longitudinal study

If a longitudinal study is designed to have a large achieved sample size, then a subgroup can still have an achieved sample which is adequate for analysis. It is not necessary to rely on a proportionate representation of the population characteristics in the sample; one strategy is to oversample defined subpopulations as part of the design of a longitudinal survey. When there is a small number of such subpopulations of particular interest, this oversampling can be built into the original sample design of the longitudinal study, using whatever related information is available (for example small area census data, see Lynn et al. 2018) to identify areas where there is relatively high prevalence of the required minorities. This requires that (ideally) the populations are identifiable, or that some known predictor of the density of the population in particular areas is available. Then more of the sample can be allocated in these cases/areas. But this does not work for hidden subpopulations, where there is no good predictive information on which to base a sample design; in this case screening interviews may be needed to identify members of the population to include in the sample without the sample costs being taken up by cases not in the minorities of interest.

Building the subpopulation requirement directly into the design in this way has a number of advantages:

- it ensures integration of the data collection for the minority, particularly the baseline data collection which provides the foundation for the analysis of individual changes.
- the population is representative in the same way as for other parts of the longitudinal survey sample (see Goldstein *et al*. 2015)

but also has a number of costs (or possibly disadvantages):

- interviewing minorities can be more expensive, so with a fixed overall budget it may reduce the total number of cases that can be included; this may include screening costs.
- the oversampling will increase the variance of the weights, which is known to increase the variance of estimates (Kalton 2014)

- the subpopulation will be an integral part of the sample design, so at least the original questions for the survey must be included. It may be possible to add further questions tailored to the subpopulation of interest, but any increase in the survey length could potentially reduce cooperation and increase attrition, so it makes sense to keep such question to a minimum. This also makes questionnaire design simpler with less routing.
- the oversampling may need to be quite extreme if the original sample is to yield an adequate sample for the subpopulation over many waves in a longitudinal survey, particularly if it is expected that the subpopulation will be subject to higher attrition rates than in the remainder of the sample.

A second approach is to have a boost (supplementary sample) which uses the same data collection procedures as the main survey, though possibly not the same sampling procedures, and which is used at a wave after wave 1. The intention is to collect comparable data from all cases in the main survey and the boost, but to have an increased number of cases for the target subgroup(s). Some variation to the data collection may be needed to collect some retrospective data. There is then a question of whether this boost should become part of the main study, with the same following rules and follow-up waves. This is the approach used in the immigrant and ethnic minorities boost to Understanding Society (Lynn et al. 2018). In principle this approach is efficient, because a boost can be added when the sample size in subpopulations of interest becomes too small. On the other hand, the main interest is in *longitudinal* analysis, so it would be good to have longitudinal data back to the beginning of the study for modelling.

If there are several boosts for different subpopulations this process may be expensive in sample management, KIT, tracing, interview and nonresponse conversion resources, and will also complicate the analysis of the full dataset. To ameliorate this problem, boosts could be followed up less frequently than the main longitudinal survey. This would reduce the costs, and would allow different subpopulations to be interleaved across waves so that they are all covered regularly. But we might expect that attrition and KIT costs will both be higher in this scenario as respondents' contacts with interviewers are less frequent. There may be further options for reducing costs, such as using a short web-completion questionnaire in years when the boost is not followed up, to maintain contact and provide data to help identify and explain transitions.

a. Following rules

The *following rules* for a longitudinal survey are an important part of the data collection, which define which of the sample members will be followed in a subsequent wave. These rules are important, because some transitions are associated with changes in household composition (for example divorce, or children leaving home), and these events may be of particular interest (Iacovou & Lynn 2013). If these are not properly followed, there will be a shortage of data on these cases and some transitions of interest may not be analysable. In this sense these subpopulations are of particular interest, but no focused collection is needed – just the proper implementation of a defined set of following rules. A corollary is that in analysing life events that are associated with changes of household structure, it is important to reflect the impact of the following rules on the analysis.

3. Separate subpopulation samples

'Separate focussed studies' can be more tailored to the subpopulations they are intended to study. They can ask a range of questions which are very specific and would not be reasonable in a general population survey. A core set of basic questions will be needed for classification and for comparative analysis with other subpopulations, and these should use harmonised survey questions so that they are comparable across different studies.

The main advantage to separate focused studies is that the sample can be designed to provide a sufficient number of cases from the target subpopulation, using whatever approach seems to offer the best trade-off of costs and bias. Most of the sampling methods (see section 4) which produce relatively large samples in specific subpopulations are either susceptible to bias through exclusion of parts of the population, or dependent for their properties on assumptions which may be unlikely to hold. But it is usual to accept such biases because of the reduction in costs per case of sampling in these ways.

However, separate studies are likely to be frequently at risk, since they will require relatively small amounts of funding, and therefore be much easier to stop during a period of financial austerity. We assume, however, that the whole reason for wanting to study a subpopulation in this way is that there are questions which are best or can only be answered by a longitudinal survey. So once a separate study has started, it would be unfortunate if it was not continued.

These samples cannot easily be integrated with general population data from a probability sampling scheme. In the cases where the methods of sample recruitment mean that analysis is reliant on adjustments derived from models, for example, the estimation procedures are different (although elements of model-based approaches are already used in the probability sampling approach, such as for nonresponse adjustment). In a separate study the estimation can be tailored to the sampling used (although for general research purposes some training of researchers in the appropriate quantitative methods would be useful). The gains from the relatively few cases in the general population data is unlikely to be important in many cases. Where the objective of a focused study is to compare the outputs with similar analyses from the general population study, we need to develop procedures that allow these comparisons, accounting for the differences in inference. This is an instance of a general problem, of how to combine data from separate studies (many of them using model-based approaches) with a general population sample (using a design-based approach). More research on how to make these combinations, and 'borrow strength' (use the best properties of both parts of the data) to get the best or most comparable estimates, would be worthwhile.

4. Methods for gathering adequate numbers of cases for rare subpopulations

A variety of methods have been used to capture cases from hard to sample populations. These methods tend to suffer from bias, high variance or high cost relative to standard probability designs for longitudinal studies where there is a good auxiliary information to help with an efficient design.

Screening

Samples of minority groups can be taken using a direct approach where households are only included if they contain people who are part of the target population. This may implemented by screening households using a small number of questions to see whether they belong to the subpopulation(s) of interest first, and then asking the full survey only if they do. Using this process tends to be expensive, because potentially many households must be contacted for each eligible member, and in cross-sectional surveys this unused information from screened-out households is expensive. In a longitudinal survey, however, a screening approach may be more cost effective relative to the total cost (the screening procedure is an investment in recruiting a sufficient number of cases in the subpopulation of interest with a high-quality probability sampling method, which will be included over multiple waves). This does not however change the actual cost of the screening, which must be done up-front at the first wave.

A screening procedure gives a probability design, which allows unbiased estimation, but the variances can potentially be large – particularly if the size of the population needs to be estimated, rather than being available from a control source (such as the population census), since this contributes an extra component of variance. For relative measures such as proportions of the subpopulation with certain characteristics, the variance will be less affected by the screening procedure.

There are ways to make screening more efficient in some circumstances. One strategy is to use an existing survey sample as the basis for selection, since then the screener question can be added to that survey at minimal cost. Account would need to be made of undercoverage, nonresponse and attrition in the screening survey in analysing the screened sample, but this extra complication in analysis might be worth the saving in fieldwork costs. A second strategy is for two surveys to share the screening. The Health and Retirement Survey (HRS) and the Panel Survey of Income Dynamics (PSID) shared screening in the US, and in the age range where their screening requirements overlapped the HRS took the screened households, with PSID using the relatives of screened in members to identify further people in the required age range and generate boost samples of recent immigrants (Sastry 2017 in Watson & Lynn 2020).

Design – disproportionate stratification

A less targeted approach is not to screen, but to adjust the probability that a unit will be included in the sample based on the characteristics of the area. For example the Wealth and Assets Survey was designed to oversample higher earners by using higher selection probabilities in areas where the rate of self assessment income tax was higher (Smith *et al*. 2011). The level of oversampling may need to be adjusted (and was in the example given) to ensure that target sample sizes for the minorities are met. Disproportionate stratification is only effective when

- the target subpopulation has much higher prevalence in the oversampled strata
- the proportion of the subpopulation in the oversampled strata is high
- the unit cost of data collection is not substantially higher than the screening cost

(Kalton 2009). In practice the gains from disproportionate stratification are typically modest, because the oversampling increases the range of the weights, which inflates the variance.

If administrative data (for example from a spine as promoted in Davis-Kean et al. (2017)) gives good predictor variables for the subgroup of interest, then they can be used as the basis for stratification rather than area-level variables, and the conditions in the bullet points above are more likely to be met. The ethical basis for using administrative data in such a way may need to be clarified, since it will use personal administrative data before there has been a chance to ask the respondent for consent.

Network sampling and indirect sampling

Network sampling encompasses a range of methods, including multiplicity sampling, link-tracing (snowball) sampling and respondent driven sampling (RDS). A recent review is given by Heckathorn & Cameron (2017). These approaches can also be viewed as a form of indirect sampling (Lavallée 2014) with estimation following with the generalised weight share method.

Link-tracing and respondent driven sampling (RDS) both rely on the members of a subpopulation knowing each other and identifying other members to the researchers. They both require model assumptions to hold to make unbiased estimates (though Zhang & Patone 2017 present a condition under which Horvitz-Thompson estimation can be used, although it is unlikely to be met in practice), and this model-based inference does not fit well with the predominantly design-based analysis procedures of longitudinal studies. Therefore in a situation where the only way to include sufficient cases is to use one of these techniques, the analysis would be much more straightforward if the survey were treated as a separate study. This would help make it clear to users of the data that different methods should be used, and underline the different interpretations of the main study and separate focused study.

An interesting idea is to start a link-tracing sample from the existing members of the target subpopulation in a panel (but not a cohort, as people identified by link-tracing wouldn't in general belong to the cohort of interest). Although a probability sample of such *seeds* is not actually a requirement of such a procedure (Salganik & Heckathorn 2017), this does offer some reassurance that the coverage starts from the same basis as the original longitudinal design. Equally, this does not solve the generalisability problem, but the common cases between the panel and the snowball study offer some interesting possibilities for modelling.

For a longitudinal application of link-tracing or RDS there is a question of what population is being covered. A simple option is to undertake the link-tracing at a specific time to get a sample, and then to follow this sample longitudinally. This mirrors the approach in cohorts and panels where the population is defined as that at the first sampling occasion. However, we might be interested in change in the subpopulation as well as the characteristics of the people who form it, which might argue for top-up link-tracing at later waves to get a better cross-sectional representation as well as the longitudinal one. There may be some ethical challenges with this approach, in trying not to allow the link-tracing procedure to approach people who have previously refused to participate. The question of how to integrate such a series of link-tracing samples has not been addressed in the literature, and some exploratory research into how this type of data could be analysed would be worthwhile.

Adaptive cluster sampling (Thompson 1990) is another type of network (or indirect) sampling. An initial set of clusters is selected randomly, and if a surveyed cluster contains a member of the target subpopulation, then its neighbours are also added to the sample, continuing until no further neighbours belong to the subpopulation of interest. Some consideration is needed of how to implement this in practice, because most social surveys use a cluster sampling approach, where only a few households in each cluster are selected, so an algorithm for selecting cases within a cluster would be needed. In principle the technique would work with non-geographical clusters as long as a means of defining “neighbours” to sample members on the variable of interest is available (for example based on some administrative variable); whether such an approach could be implemented in the field would also need to be tested.

Focused enumeration is a network sampling variant of screening where after assessing the eligibility of the sampled unit, the respondent is asked whether they know other units in the target subpopulation (for example among neighbours). In principle this increases the number of cases at the cost of some additional clustering and the introduction of the network. But an analysis of examples in a presentation (Smith et al. 2010) suggested that these theoretical benefits might not be achieved in practice. Reichel & Morales (2017) report an example where this approach was effective as one component of a recruitment (but for a cross-sectional survey). A similar approach, but called shadow sampling and relying on observable characteristics of neighbouring dwellings of the English Housing Survey sample and which could be identified by the interviewer, was used on the English House Conditions Survey in 2005 (DCLG 2007). A review of the evidence for the effectiveness of focused enumeration in general, and its potential to generate longitudinal samples of subpopulations of interest, would be valuable.

Location, time-location sampling

A frame of locations (or *venues* or *intercept points*) or locations × times can be constructed as the basis for sampling. Under certain assumptions about whether the subpopulation of interest has a chance to visit a location/time in this frame, it is possible to use this approach to make estimates. When people are the unit of analysis some adjustment has to be made when they have multiple chances to be included in the survey at different locations or times; this can drastically increase the variability of the weights, but is necessary for any even approximately unbiased estimation. Baio *et al*. (2011) set out the methodological framework for cross-sectional samples, but it would be interesting to see how well contact information could be obtained and a follow-up wave administered in a longitudinal survey.

Multiple frame sampling

A lot of research attention has been focussed on sampling and estimation with multiple frame surveys (see for example Yu & Lohr (2010) who apply dual frame estimation in a longitudinal context to estimate gross flows). These are useful for rare populations if the frames contain some identifying or indicative information on the subpopulation of interest, but do not individually present complete coverage, because it allows sampling to be more efficient. The analysis of such datasets is quite involved, however. The data collection process must be able to identify when a respondent belongs on more than one frame, so that appropriate adjustments can be made in the estimation to account for their additional chance of selection.

It is possible to combine multiple frame with dual (or multiple) system estimation to estimate the size of the part of the subpopulation which does not appear on either frame. In order to do this it is important to match accurately, which means identifying when sampled units are in multiple frames; any errors in this identification will tend to inflate the estimate of the size of the missed population, as well as biasing the estimated values derived from the responses (so creating a bias through two mechanisms).

General framework and evaluation of different approaches

Zhang & Patone (2017) show that many of the sampling approaches described above can be fitted into a framework based on graph sampling, and Lavallée (2014) presents a similar framework through indirect sampling; these unifying frameworks suggest some scope to compare the properties of the different design options in different cases. There are few examples where different designs have been compared in practise, but McKenzie & Mistiaen (2009) compared snowball and location sampling with a standard probability sample, and discovered that the former two methods oversampled people more closely connected with the subpopulation of interest. Kendall *et al*. (2008) compared respondent driven, snowball and time location sampling, and found that respondent driven sampling gave better coverage more quickly than the other methods. Both these examples derive from S America, and it would be interesting to undertake such studies in a UK context to broaden the evidence base for how to sample in targeted studies.

5. Periodic boosts

Returning to the idea of designing the main survey to provide adequate representation of some key groups, we already saw that there may be a need for some disproportionate stratification. The impacts of differential attrition may suggest targeting the sample even more towards particular groups (Smith *et al*. 2009). To counteract the effect of extreme oversampling in the initial design of a study, an alternative possibility is to introduce a boost at a later wave, perhaps when the attrition has been sufficient to trigger a quality threshold. Bianchi & Biffignandi (2017) propose the use of R-indicators to assess whether particular subpopulations are over- or under-represented in a panel sample, and apply their approach to data from Understanding Society; this provides a useful indicator for whether a boost might ned to be considered, although context and user needs will also influence a decision.

When a boost sample is introduced, baseline information must be collected for the new cases at the first contact, which is during the boost recruitment. This data will therefore refer to a different time period to the baseline data from the main survey, and in general complicate analysis, though how much depends in part on the stability of the characteristics or relationships being investigated (Lynn in Goldstein *et al*. 2015).

An alternative view is that the different baseline periods in successive boosts are useful for investigating the stability of a phenomenon. In that case the different periods act at least in part as different subpopulations, and therefore suffer from some of the same issues of small samples size. But if there are no substantial differences in the baseline measures at different times (akin to convergence in accelerated longitudinal designs, see work package 6), then the data may sensibly be pooled. The logical extension of this approach is a series of boosts approximating to a rotating panel design, though for a longitudinal study there will always be a reluctance to actually drop cases as in a genuine rotating panel.

6. Responsive design and booster samples

The persistent reduction in response rates in social surveys (de Leeuw *et al*. 2018) has led to a substantial body of research on understanding the reasons for non-response and its impact on survey estimates. Survey organisations continue to adopt a range of strategies to monitor and maximise response. One class of strategies has become known as *responsive design*, where in broad terms some indicators of the quality of the survey are monitored during the data collection phase, and where they show that certain subpopulations show signs that the data collected will be low quality (generally because of poor response), then action is taken to increase the information availability and therefore quality for these subgroups.

The types of actions available include targeting field work at particular groups or in particular areas (which usually means reduced activity for other parts of the sample), use of the most experienced interviewers in these cases, attempts at refusal conversion, etc. For a review of responsive (adaptive) design strategies see Tourangeau *et al*. (2017).

The availability of R indicators (Schouten *et al*. 2009) as an indicator of the bias due to non-response (in a more direct way than response rates, which indicate the risk of bias rather than the size of the bias) has led to adaptive design strategies which are targeted at improving the R indicators, and this often involves targetting collection and conversion activities at subpopulations where the R indicators suggest there is most contribution to the variation in response rates.

There is an analogy here with data availability for minority groups (as shown by Bianchi & Biffignandi (2017)). We can introduce an adaptive sampling strategy, where we introduce additional sample blocks in response to indicators of low quality (in this case small sample size, which has a contribution to *variance* rather than bias). This approach could be developed to suggest regular boosts in targeted parts of the population distribution, maintaining the usefulness of the data and providing an opportunity to update at least part of the sample in response to changes in the population.

In a true responsive design, there is a stopping rule to determine when the minimum quality criteria for the survey have been met. In a longitudinal survey there may be many waves, so the work of recruitment to maintain the size and composition of the sample can be expected to continue as long as the panel does.

7. Coordinated sampling

Sample coordination is a suite of techniques designed to allow two (or more) samples to contain common units or to avoid common units. This can operate at different levels. Watson & Lynn (2020) suggest that fieldwork for refreshment surveys can be more efficient if addresses within PSUs already selected for fieldwork are used, although this increases the clustering and therefore has higher variance than choosing new PSUs. They also suggest using information from previous surveys, or shared between surveys (in the case of screening), and all these approaches are types of positive coordination.

A variety of methods has been proposed to coordinate samples (see Nedyalkova *et al*. 2006 for an overview), and they are typically used in business surveys for controlling survey burden and inducing overlaps from period to period, neither of which is particularly relevant in a longitudinal survey context. But there is scope to ensure that a programme of longitudinal surveys (or surveys and boosts) is coordinated. Positive coordination of areas would make sample units closer together and possibly improve fieldwork efficiency; negative coordination would give wider coverage of areas and improve accuracy (through reduced clustering). Coordination of social surveys is used in the Swiss Federal Statistical Office, and there would be scope to use it in the design of a programme of longitudinal surveys to select good complementary samples across several designs.

Geographical coordination need not be restricted to PSUs as defined in samples designs – grouping PSUs and coordinating samples on the groups may give many of the benefits of fieldwork efficiency but still spread the sample sufficiently to reduce clustering effects.

8. Analysis

The network sampling approaches are non-probability designs, even though in some cases there are unbiased estimators of the population size under suitable assumptions about a model for the sample structure. But this means that even standard regression analyses are complicated. Beckett *et al*. (2017) consider regression analysis of a *cross-sectional* dataset collected by respondent driven sampling, and say that there is no clear multivariable regression approach for data derived from RDS. They examine two logistic regression approaches. One uses weights and introducing a simple covariance structure with recruiters (clusters) nested with trees (strata) to deal with the correlation between individuals selected by the same recruiter. The other approach is a mixed logistic regression model with the variance structure derived from the form of the tree. In this latter model there were sometimes convergence problems, for which a simpler autoregressive covariance structure was substituted.

The point estimates of the regression parameters were quite similar between the two models, but there were some differences in the variances, which would lead to slightly different inferences. Parameter estimates for two of the variables failed to converge in the multilevel model.

Extending such a model to the longitudinal situation should in principle be handled by the addition of a random effect to account for the correlation of measurements at different times for the same individual. This generates a more complex model, and given the convergence problems with the cross-sectional model, it would be interesting to see whether such a model would converge.

More research on the best types of models to use to make inference about relationships between variables derived from RDS and similar approaches is needed, including how to specify the covariances, and what sorts of sample size and spread is needed for convergence of the modelling algorithms.

9. Discussion

The general tenor of the arguments above suggests that methodologically there are good reasons to do separate, focused longitudinal studies for subpopulations, rather than trying to incorporate all the specialist requirements into one large survey. Questions and approaches can be tailored more to the population of interest. If the survey instruments can be standardised as much as possible that will facilitate integration with the main study. There are some clear exceptions where the subpopulation of interest is sufficiently large and well-defined that it can be included as part of the main data collection, such as ethnic minorities. Nonetheless, where attrition is disproportionately large there may be a need for a boost sample to return to a minimum achieved sample size.

In practice, however, smaller studies are vulnerable to changes in funding, so despite this tendency there is a case for having a larger central survey with oversampling for subpopulations of interest. This will increase the complexity of design and processing, and also increase the variance) of estimates from the central survey.

Recommendations

- More research on how to combine smaller focused studies with a main panel survey in order to do comparative inference and make the best use of samples in minority groups from both sources.
- Investigate how to integrate a series of top-up samples selected with network designs at different times into an ongoing panel survey.
- Undertake a systematic review of the evidence for the effectiveness of focused enumeration in a general longitudinal survey context.
- Undertake an experimental small-scale test of non-probability sampling methods alongside probability sampling methods for a longitudinal survey (ie with at least one follow-up wave) in the UK, to supplement existing evidence for the impacts of non-probability designs.
- Investigate the best form of models under network-type and other non-probability designs, including the best ways to model the covariance structures and obtain convergence in the fitting algorithm. Provide guidance and tools for researechers.

References

Baio, G., Blangiardo, G.C. & Blangiardo, M. (2011) Centre sampling technique in foreign migration surveys: a methodological note. *Journal of Official Statistics* **27** 451-465.

Beckett, M., Firestone, M.A., McKnight, C.D., Smylie, J. & Rotondi, M.A. (2018) A cross-sectional analysis of the relationship between diabetes and health access barriers in an urban First Nations population in Canada. *BMJ open* **8** e018272.

Bianchi, A. & Biffignandi, S. (2017) Representativeness in panel surveys. *Mathematical Population Studies* **24** 126-143.

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.

DCLG (2007) *English House Condition Survey, Technical Report* (2005 Edition). Communities and Local Government Publication Centre, Wetherby.

de Leeuw, E., Hox, J., & Luiten, A. (2018) International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data. *Survey Methods: Insights from the Field* Retrieved from https://surveyinsights.org/?p=10452.

Goldstein, H., Lynn, P., Muniz-Terrera, G., Hardy, R., O'Muircheartaigh, C., Skinner, C.J. & Lehtonen, R. (2015) Population sampling in longitudinal suverys. *Longitudinal and Life Course Studies* **6** 447-475.

Heckathorn, D.D. & Cameron, C.J. (2017) Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual Review of Sociology* **43** 101-119.

Iacovou, M. & Lynn, P. (2013) Implications of the EU-SILC following rules, and their implementation, for longitudinal analysis. ISER Working Paper Series 2013-17.

Kalton, G. (2009) Methods for oversampling rare subpopulations in social surveys. *Survey Methodology* **35** 125-141.

Kalton, G. (2014) Probability sampling methods for hard-to-sample populations. Ch. 19 in R. Tourangeau, B. Edwards, T.P. Johnson, K.M. Wolter & N. Bates (eds) *Hard-to-survey populations*. Cambridge University Press: Cambridge.

Kendall, C., Kerr, L.R., Gondim, R.C., Werneck, G.L., Macena, R.H.M., Pontes, M.K., Johnston, L.G., Sabin, K. & McFarland, W. (2008) An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. *AIDS and Behavior* **12** S97-S104.

Kish, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics* **7** 339-347.

Lavallée, P. (2014) Indirect sampling for hard-to-reach populations. Ch. 21 in R. Tourangeau, B. Edwards, T.P. Johnson, K.M. Wolter & N. Bates (eds) *Hard-to-survey populations*. Cambridge University Press: Cambridge.

Lynn, P., Nandi, A., Parutis, V. & Platt, L. (2017) Design and implementation of a high quality probability sample of immigrants and ethnic minorities: lessons learnt. *Demographic Research* **38** 513-548.

McKenzie, D.J. & Mistiaen, J. (2009) Surveying migrant households: a comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society, Series A* **172** 339-360.

Nedyalkova, D., Pea, J. & Tillé, Y. (2006) *A review of some current methods of coordination of stratified samples. Introduction and comparison of new methods based on microstrata*. Rapport technique, Université de Neuchâtel.

Salganik, M.J., & Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* **34** 193-240.

Schouten, B., Cobben, F. & Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Survey Methodology* **35** 101-113.

Smith, P., Ashworth, K. & Lound, C. (2011) Using administrative data sources to target sampling in the Wealth and Assets Survey in the UK. *Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data* 135-140. http://publications.gc.ca/collections/collection_2014/statcan/CS11-522-2010-eng.pdf.

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.

Thompson, S.K. (1990) Adaptive cluster sampling. *Journal of the American Statistical Association* **85** 1050–1059.

Watson, N. & Lynn, P. (2020) Refreshment sampling for longitudinal surveys. In P. Lynn (ed) *Advances in longitudinal survey methodology*. Wiley: Chichester.

Wu, Y., & Lohr, S. (2010) Gross flow estimation in dual frame surveys. *Survey Methodology* **36** 13-22.

Zhang, L.-C. & Patone, M. (2017) Graph sampling. *Metron* **75** 277–299. doi: 10.1007/s40300-017-0126-y.

1 Literature Review

James Yarde, NatCen Social Research

Survey research is based upon the principle of representativeness; either of the population in its entirety, or some specific sub-population of interest. Increasingly heterogeneous populations pose some important questions for the survey researcher. Not least in ensuring that the populations to which survey findings are attributed are meaningful, both in theoretical and practical terms. The escalating demand for sub-group analysis, as noted by Kalton (2009), necessitates a recognition of the inherent trade-offs associated with ensuring that marginal groups are sufficiently sampled for meaningful statistical analysis.

For longitudinal survey research the challenge is yet more acute. As "representativeness is not a static concept that is preserved indefinitely over time" (Goldstein, et al., 2015, p. 458), the temporal aspect of representativeness is especially pertinent here. The risk is that, without due consideration, samples in longitudinal surveys may drift away from the populations which they purport to be representative of. In 2019, respondents to the National Child Development Study (NCDS) are due to turn 61. But respondents to this survey cannot be said to be representative of all 61-year olds in the UK population given factors such as migration. This remains the case despite attempted mitigations while the cohort was still of school age, with top-up samples being drawn for some marginal groups through the use of administrative data.

1.1 Representativeness: national vs. sub-groups

Where sub-group analysis of a sample is to be conducted, the question of representativeness relates to a given sub-group and the corresponding part of the population. Kish (1987) uses the term *subclass* to refer to a subdivision of the sample, while *domain* refers to the corresponding subdivision of the population. Further to this, four types of domain are set out: (1) major domains (>10% of population), (2) minor domains (1-10% of population), (3) mini-domains (<1% of population), (4) rare types, meanwhile, comprise less than 1/10,000 of the population. The smaller of these types of domain (from minor domains downwards) will ordinarily require measures to be adopted in the sampling design to ensure that a sufficient number are identified and drawn into the sample (Kalton, 2009).

The overriding statistical concern here is the size of the domain of interest, rather than the underlying factor(s) which define membership. With this in mind, domains can refer to geographic areas, socio-economic groups, or other sub-populations (Rao & Molina, 2015). Irrespective of which of these characterisations applies, research suggests that there is an inherent trade-off between ensuring that both national and domain level estimates, for a given variable, are precise. Target sample sizes for domains can be determined by following one of a variety of formulations (Bankier, 1988; Kish, 1976; 1987). Even still, these procedures still risk that they "may not allocate sufficient sample to small domains to produce estimates at the required level of precision" (Kalton, 2009, p. 127). This point is echoed elsewhere in the literature, where "random population samples are often insufficient to accumulate large enough samples of hard-to-reach groups" (Bonevski, et al., 2014, p. 17).

Subsequent approaches have sought to build on the work of Kish and Bankier, such as Costa et al (2004) and Longford (2006). The former approach (Costa, Satorra, & Ventura, 2004) suggests a compromise between proportional and equal allocations which takes account of the convex function of mean squared errors for samples of the population and sub-populations. Meanwhile, Longford (2006) developed a procedure wherein different domains are assigned inferential priorities, through which a compromise allocation is again achieved. Subsequent analysis by Choudhury et al. suggests that, while both of these approaches "perform reasonably well in terms of reliability requirements" (Choudhry, Rao, & Hidiroglou, 2012, p. 28), that allocations which use nonlinear programming (NLP) can be more effective at minimising the ratio of the standard deviation and the mean (the coefficient of variation).

One contrasting approach has previously been used in the Canadian Labour Force Survey (LFS), where a core sample of 42,000 households was selected to produce reliable estimates at the national level. A supplemental sample was then allocated to optimise the sub-provincial estimates. In empirical terms, this approach was found to lead to a small increase in the coefficient of variation (CV) for Canada (1.51% from 1.36%). But at a sub-provincial level, the maximum CV declined from 17.7% to 9.4%. Thus, the approach appears to have fulfilled its rationale: to "produce reliable estimates for almost all planned domains" (Singh, Gambino, & Mantel, 1994, p. 8).

The effect of an increased sample size for a marginal group on the precision of estimates (as measured by the coefficient of variation) is not constant. Indeed, increasing the oversampling ratio results in diminishing returns in terms of post-weighting effective sample sizes (Kalsbeek, 2003). But, equally, the inverse is also true. In the context of Canadian provinces, "small reductions in sample sizes for larger provinces usually [have] little effect on the reliability of data […] but the corresponding sample increase in smaller provinces has significant impact on the reliability of their data" (Singh, Gambino, & Mantel, 1994, p. 6).

For studies which look to analyse both a population and specific sub-divisions of it, this can be done without too much compromise in terms of the precision of population-level estimates. The optimal balance can, in turn, be sought by using one of the approaches enumerated here. Use of booster samples is therefore a valid means through which to analyse specific marginal groups. So long as a compromise allocation is reached, an individual survey remains capable of simultaneously producing precise estimates at population and domain levels. Consideration still needs to be awarded to the specific sampling strategies required to identify and engage members of marginal groups in the survey. The need for this to be considered remains just as potent where a study is designed solely around the specific marginal group.

In a similar vein, the anticipated response rate among different groups needs to be considered in the sample design. If survey response among the marginal group is expected to be comparatively low, then this needs to be factored in so that the achieved sample is sufficiently large. In the context of longitudinal surveys, matters are further complicated when the rate of response from wave to wave differs between groups. In cases where the rate of attrition is higher for one group than another, this needs to be considered in the survey design at the outset. This is explored in greater detail in Section 1.5.

1.2 Sampling methods for marginal groups

Irrespective of whether a specific marginal group is to be oversampled to ensure that the sample size is sufficiently large to produce reliable estimates, or if a separate study is to be designed, the appropriateness of the chosen sampling method is of paramount importance. Without due diligence, there is a risk of bias affecting the estimates – especially if the group of interest (or part thereof) is hard to reach.

In the first instance, the characteristics of the domain of interest should inform the sampling strategy. Kalton (2009) outlines a number of considerations which should be made before deciding which sampling method is most appropriate. Firstly, for standard sampling approaches to be effective, a suitable sampling frame for the marginal group – which is sufficiently up to date – needs to be available. Failing this, an alternative approach should be used. Where, for instance, the population of the marginal group is concentrated (and the concentrations in different areas are known), then disproportionate stratification can be used. Areas with a high concentration of the group of interest can be assigned a higher probability of selection, therefore boosting the sample of the marginal group. By contrast, where the group is more thinly spread, then screening for members of the group is likely to be more appropriate. Finally, in cases where members of the population of interest are identifiable, then a network method could instead be used.

Other authors have also reach similar conclusions. For instance, Sudman, Sirken and Cowan (1988) suggest that where special populations are geographically clustered, existing data can be used to sample these. Equally, where populations are dispersed and/or mobile, network sampling may be more appropriate. Ultimately, to conduct surveys among specific minority groups, "a sampling frame must be available of these specific populations, or people have to be identified through screening questions in a general survey, or another sampling strategy" (Stoop, 2014, p. 226).

The quality of the available sampling frames should inform the sampling design in the first instance, in any case. If the group of interest is particularly rare, then it can be advantageous to use multiple (rather than single) sampling frames, if the population is likely to be under-represented (Binder, 1998). This can be particularly useful where no single frame provides coverage of the entire population of interest. Issues potentially arise, however, where there is overlap between the sampling frames, i.e. an individual appears in more than one frame. In such a case, variation in the probability of selection needs to be accounted for during survey weighting.

1.2.1 Administrative data

To be able to take a simple random sample of individuals, given a known profile of characteristics, a suitable sampling frame needs to be available. Where a particular marginal group is of interest, this typically requires high-quality administrative data to be available – wherein the information provided is both accurate and up-to-date. One of key promises of administrative data, with reference to marginal groups, is that it has scope to provide "detailed data on hard-to-reach populations" (Harron, et al., 2017, p. 1). When of sufficiently high quality, administrative data may enable a simple random sample of the population(s) of interest to be drawn – whether this be within a national survey, or as part of a separate study.

The use of administrative data to provide a single or multiple sampling frames for survey designs is more common internationally than it is in the UK. Stoop (2014) gives the example of how in the Netherlands, the availability of administrative data enables working women with young children to be identified and sampled. This is made possible because the population register (held by Statistics Netherlands) can be linked to administrative data, including employment and social welfare records. Because this data is available at an individual rather than a household level, sampling can be a relatively straightforward exercise once the sampling fractions for different sub-groups have been determined.

In the UK-based Families and Children's Study (FACS), administrative data was used to sample Child Benefit recipients. In Wave 9, sub-sampling of the panel took place, wherein priority groups (including lone parents and families with an equivalised income below 70% of the median) were selected automatically; all remaining non-priority cases were selected randomly (Conolly, Maplethorpe, & D'Souza, 2009). Furthermore, a booster sample was issued to reflect families moving into the postcode sectors selected in the original waves of the study. Because the majority of characteristics relevant to the sample selection are known – with limited screening required for those that are missing – it avoids much of the uncertainty associated with sampling households, for instance.

In cases such as this, where the population is known and has specified characteristics, the use of administrative data to select a sample has scope to be extremely effective. In one sense, FACS is a separate study of a marginal group (families in receipt of Child Benefits) in itself. But, where further sub-group analysis is required, the administrative data from which the sample was derived could, in theory, be leveraged again.

Problems primarily arise, however, where the sampling frame used is insufficient – with bias the likely result. Indeed, one potential disadvantage of sampling from administrative data is that it risks capturing only a subset of the marginal group (Kalsbeek, 2003). Consider, for instance, the Hispanic population in the United States. The likelihood of sampling the *settled* Hispanic population is different to the *mobile* population, and the degree to which it differs will depend on the quality of the sampling frame. A number of other problems related to the use of administrative data are also outlined in Harron et al. (2017). Particularly relevant here are issues relating to data quality. If, for instance, data is either missing or is no longer accurate, this may have implications for the sample. If there are systematic reasons why some groups are less likely to have inaccurate data held about them, then this is again likely to manifest itself in terms of bias.

1.2.2 Disproportionate stratification

Where members of a marginal group are not known to the survey researcher, a different approach needs to be adopted. One such approach is to employ disproportionate stratification, wherein some strata are sampled at a higher rate that others. If, for instance, the group of interest is known be concentrated in one area – and account for a large proportion of the population there – then this area can be oversampled. By so doing, there is an increased likelihood that the sub-group will be sufficiently represented within the achieved sample.

Disproportionate stratification has commonly been used in household surveys, where there is a need to boost marginal groups and the geographical distribution is known or can be approximated. Where, for instance, there is a requirement to boost the number of ethnic minority households sampled, census data can be used to identify areas where there is a high concentration of this population. In studies, such as Understanding Society (USoc), data from the Annual Population Survey (APS) has also been leveraged to estimate the population composition of small areas (Berthoud, Fumagalli, Lynn, & Platt, 2009). Such an approach helps to avoid any potential issues related to data regarding the concentrations of minority groups being out of date.

But for such a method to be effective, certain conditions must be met (Kalton, 2009, p. 132). Firstly, the marginal group should be "much more prevalent in the oversampled strata". As such, the likelihood of an address with a member of the marginal group present is more likely to be identified than in a stratum with low prevalence of that group. Secondly, "the oversampled strata must contain a high proportion of the rare population", else the gain in terms of precision of estimates for the sub-group will likely be subdued. This is the case given that "substantial oversampling can have important negative implications on the statistical quality of estimates, particularly when area clusters are selected and the oversampled group is scattered geographically" (Kalsbeek, 2003, p. 1528). Finally, "the cost of the main data collection per sampled unit must not be high". Should this final condition not hold, other sampling methods may be more efficient.

One significant drawback of disproportionate stratification is that it remains a high cost approach. The need to sample a set number within a highly concentrated population still cannot justify sampling these in one area. Taking the example of the Bangladeshi population, it "would not [be] legitimate simply to interview 1,000 Bangladeshis in the most densely concentrated area of Tower Hamlets" (Berthoud, Fumagalli, Lynn, & Platt, 2009, p. 9). Against this backdrop, the consideration of cost may lead to researchers using non-random methods instead. But "using such methods leads to questions about the validity of the survey's findings for the target population of inference" (Kalton G. , 2014, p. 417). The cost of disproportionate stratification therefore needs to be balanced against this.

1.2.3 Screening

Where an individual's membership of a marginal group is not known at the point of sample, screening must be used to identify them if membership of said group is the inclusion criterion for the survey. The purpose of screening, in such a case, is to determine an individual or household's eligibility to take part. This is especially relevant in surveys where (1) the marginal group is the only population of interest, or (2) a sampling boost of a specific group is required due to population size or the expected response.

An extension of screening is a technique called focused enumeration, wherein those included in the address sample are first screened, before being asked about their neighbours' eligibility. Where a neighbour is known to be eligible, they are subsequently added to the sample. Examples of surveys where this sampling approach has been used include the Health Survey of England (HSE) (Becker, et al., 2006, p. 15) and the British Crime Survey (BCS) (Bolling, Grant, & Sinclair, 2008). But, while such an approach is theoretically more efficient, such gains may be outweighed by the underreporting of marginal groups that it has been found to lead to (Smith, Pickering, Williams, & Hay, 2010).

Expert screening, where interviewers are trained to identify households which are likely to contain members of the marginal group, has similarly been used to sample hard to reach groups. In one such example, Cambodian immigrants were identified in Long Beach, California by training interviewers to identify observable external characteristics of households that suggest members of that population lived there. Cues included footwear outside the door and Buddhist altars (Elliott, McCaffrey, Perlman, Marshall, & Hambarsoomians, 2009). The successful implementation of such an approach depends, however, on visual cues which relate to the likelihood of the marginal group of living in the property being available.

1.2.4 Network sampling

Network sampling, like focused enumeration, relies on individuals eligible for a survey being identifiable. Individuals/households that are sampled "serve as proxy informants to provide the screening information for persons who are linked to them in a clearly specified way" (Kalton, 2009, p. 135). In one such variation of network sampling, members of a rare population are asked to identify other members of the population – i.e. the sample snowballs (Kalton & Anderson, Sampling Rare Populations, 1986).

The disadvantage of such an approach is that, without modification of the design, it does not account for the probability of selection. Alternative forms of network sampling, such as respondent driven sampling (RDS) (Volz & Heckathorn, 2008) are designed to take account of the probability of selection. It can do so because the total number of eligible connections to each individual is known, the relationships are reciprocal, and recruitment is random while following a Markov chain.

In either case, networked sampling is best used in cases where a marginal group is recruited for a specific study, given that the underlying method is so distinct from, say, random probability sampling. In some cases, however, it may be the only viable means through which a rare population can be sampled. This applies in cases where the population of the marginal group is extremely small and geographically dispersed, but members of the population are known to each other and can therefore be identified.

1.3 Domain specific considerations

In some cases, the high costs which are associated with drawing a probability sample of marginal groups often leads researchers to use non-random methods. While "using such methods leads to questions about the validity of the survey's findings for the target population of inference" (Kalton G. , 2014, p. 417), in some cases a probability sample may not be permissible. Such a judgement depends entirely on the specific characteristics of the marginal group. If the group of interest is mobile and the topic of the proposed survey is particularly sensitive, then non-random methods may be the only practical option. In such a case, a specific study of that group would be most appropriate, given the likely difficulties in assimilating the study into a random-probability sample.

The problem posed by a marginal group being hard to reach is different in cases where this is primarily because the population of the group is very small. Where the population is extremely rare, screening within a disproportionate stratification design may still be too costly. In this case, it may be better to concentrate on an even more restricted number of strata and risk non-coverage, than it is to use a non-random approach (Kalton, 2009).

This example differs from cases where the culture and characteristics of a group lead to response being less likely. Indeed, for some marginal groups, special measures will have to be taken to account for the subject matter of the survey. Where a survey deals with sensitive issues (such as sexual health), some groups may need to be approached differently to others, due to varying perceptions of sensitivity. This, in turn, means that "formulaic approaches to handling sensitivity and ethnicity in research studies are not appropriate" (Elam & Fenton, 2003, p. 21). The practicalities that this necessitates, as well as whether the researcher is concerned with sub-group comparisons, are all considerations that will need to be made before determining whether a separate focussed study is appropriate.

More broadly, some groups are less likely to respond to surveys than others, be this because they are (i) less likely to be contactable or (ii) more likely to refuse response upon contact. The characteristics of a given group may have an effect on either one of these. One such example is the ethnic minority population in Western Europe, which "tend[s] to have below-average response rates" (Feskens, Hox, Lensvelt-Mulders, & Schmeets, 2006, p. 285). In the Netherlands specifically, research drawing on the Continuous Survey on Living Conditions (POLS) suggests that "ethnic minorities do not respond as well as the native population [in the Netherlands], but the explanations […] have less to do with divergent response behaviour among ethnic minorities, and more to do with living conditions" (Feskens, Hox, Lensvelt-Mulders, & Schmeets, 2007, p. 405). If the ethnic minority population is the marginal group of interest, then this would therefore need to be considered in the sample design, either through increasing the number contact attempts, or by factoring the likelihood of response of different groups into the sampling ratio.

Where a specific and identifiable reason underlies non-response, steps can be taken to mitigate the effects of this. One such group that may be affected in this way are linguistic minorities, who are “often excluded from surveys due to cost and other factors” (Harkness, Strange, Cibelli, Mohler, & Pennell, 2014, p. 248). Should the marginal group of interest be a linguistic minority, measures such as questionnaire translation and matching an interviewer who speaks the same language can be applied in an attempt to minimise the risk of non-response. Such measures are potentially costly, however. In such circumstances, it may be more appropriate for a separate study specific to the linguistic minority to be undertaken.

1.4 Marginal groups in longitudinal surveys

The study of marginal groups within a longitudinal survey adds yet another dimension to the considerations that need to be made with respect to the survey design. Change over time needs to be factored into the design, in addition to all the issues concerning the balance of representativeness between the population and the marginal group, as well as coverage.

In the existing body of longitudinal surveys in the UK, Understanding Society (USoc) is one of the most prominent examples where a boost has been applied to ensure that a marginal group is sufficiently represented. As part of USoc, an ethnic minority boost has been applied to ensure that “at least 1,000 adults from each of five communities: Indians, Pakistanis, Bangladeshis, Caribbeans and Africans” were added to the survey sample (Berthoud, Fumagalli, Lynn, & Platt, 2009). This was a direct response to an ONS review of longitudinal data resources in 2000 which found that the number of ethnic minority respondents to the British Household Panel Survey (BHPS, USoc’s predecessor survey) was too small for a serious analysis of ethnicity.

The design of the ethnic minority boost in USoc used disproportionate stratification in addition to screening in order to identify members of the specified ethnic minorities. The rationale for doing so was that “neither of these ingredients is efficient on its own” (Berthoud, Fumagalli, Lynn, & Platt, 2009, p. 4). Further to this, efforts were made to ensure that the reference values for geographical concentrations of different ethnic minority groups by using micro-data from the 2007 Annual Population Survey to estimate change in ethnic composition between then and the census data from 2001. As such, this addressed one of the potential shortcomings of using a disproportionate stratification design. This design, in ensuring a sufficient sample of ethnic minority respondents to enable meaningful statistical analysis, has enabled subsequent academic research on topics such as attitudes to household work (Kan & Laurie, 2018) and the intersectional effects of becoming “not in employment, education or training (NEET)” (Zuccotti & O'Reilly, 2019).

The Millennium Cohort Study (MCS) is another example of a UK-based longitudinal survey that has a boost to specified marginal groups built into the design. In the past there had been limited need for birth cohort studies to take account of ethnic minorities given that these accounted for a small proportion of the population. As the UK population has become more heterogenous, the need for an ethnic boost, to ensure that meaningful statistical analysis of this sub-group could be conducted, has grown more acute. Like in the case of Understanding Society, this was achieved through disproportionate stratification – the key difference was that no screening was applied.

Scope to conduct sub-group analysis was a key tenet of the sample design for MCS. In addition to applying the ethnic boost in England, further measures were taken to take account of deprivation in each country within the UK. Further to this end, “sample sizes in the smaller countries were boosted to yield sufficient cases for within-country analysis” (Joshi & Emla, 2016, p. 420). This has, in turn, enabled research to be undertaken in each of the devolved nations: Scotland (Connelly, 2011), Wales (Welsh Assembly Government, 2011) and Northern Ireland (Sullivan, Joshi, Ketende, & Obolenskaya, 2010).

The success of the MCS with respect to the subsequent feasibility of sub-group analysis can be attributed to the considerations made at the outset of the survey design. While not directly comparable, the design of the NatCen Panel Survey is more limited, insofar as the core sample is derived from respondents to the British Social Attitudes survey. As a result, should there be an insufficient number of respondents in the marginal group of interest due to a combination of (i) the sampling design of BSA and (ii) non-consent to join the panel, then this is fed through to the base of the NatCen panel. While there are ways in which this can be addressed – such as through the use of supplementary quota samples, which can then be incorporated into the analysis through propensity score matching – these tend not to be parsimonious.

The design of these national longitudinal studies, which have boosts applied, can be contrasted with international examples of longitudinal work which relate solely to a specific marginal group. The Longitudinal Survey of Immigrants to Canada (LSIC), for instance, looks specifically at all landed immigrants aged 15+, who arrived in Canada from abroad between 1 October 2000 and 30 September 2001 (Statistics Canada, 2005). This survey was, in turn, conducted using a random probability sample taken using administrative data from Citizenship and Immigration Canada. Within the survey design, a two-stage sample was drawn to allow for oversampling of some sub-groups.

The advantage of such a design is that it can be executed efficiently, without the need to screen the sample – except where the administrative data held on individuals is either inaccurate or missing. However, the feasibility of the LSIC study was predicated on the availability of a high-quality sampling frame. Without this, other methods – such as those used in Understanding Society for the ethnic boost – would have needed to have been used (assuming the geographic concentrations of the migrant population were known). Given that the study was exclusively concerned with individuals who migrated to Canada within a set time period, an absence of high-quality administrative data would likely have been prohibitive.

1.5 Retention of marginal groups in longitudinal surveys

In cases where a longitudinal survey sample is boosted to account for marginal groups, consideration also has to be given to the rate of retention of different groups. Where marginal groups are recruited to longitudinal surveys, it does not necessarily mean that they will be retained. If, for instance, the marginal group in question is a particularly mobile subdivision of the population, then follow-up may not be possible. Equally, some groups may be less cooperative. Evidence from MCS suggests that there was disproportionate dropout by minority groups once they had been recruited (Joshi & Emla, 2016, p. 418). This experience was repeated for the Understanding Society ethnic minority boost, similarly, and led to a new sample being introduced (the Immigrant and Ethnic Minority Boost, IEMB) to compensate for sample attrition (Lynn, Nandi, Parutis, & Platt, 2018).

Annex A

There are two approaches which can be taken to ensure that sample attrition does not prevent sub-group analysis being conducted. The first of these is to design a monotonic sample where the marginal group is initially oversampled to account for attrition, as outlined by Singh, Petroni and Allen (1994). When the lifespan of the longitudinal survey is known and an assumption regarding the rate of attrition can be made, then it is feasible to design the sample so that it is sufficiently large in each wave, even as survey respondents leave the survey. This is sometimes referred to as a “funnel” design. Such a design was used in (LSIC), where respondents were required to respond to all waves, due to possible issues with recall if they missed an intervening wave. On this basis, 20,322 immigrants were selected in the initial sample, to achieve a minimum sample of 5,755 in Wave 3, based on several sample attrition hypotheses (Statistics Canada, 2005, p. 97).

The second approach that can be used is to undertake supplemental sampling in response to observed attrition across waves of the survey. Such an approach has been taken in Growing Up in Scotland (GUS) Sweep 9, in response to uneven rates of attrition across the sample. Two groups, in particular, were found to have grown under-represented: children born to mothers aged 16-24 at time of birth and children living in the 15% most deprived areas (according to SIMD). To draw an additional sample from these two groups, Child Benefit records were used. While this is no longer a universal benefit, it was deemed to be a suitable sampling frame given that the likelihood of those ineligible for Child Benefit (due to high income) being in the target groups was relatively low.

One of the benefits of supplemental sampling is it allows for a more dynamic approach which can respond to differences in attrition between groups over time. But it is not without its challenges. Primarily, there is a risk that “such a strategy can distort the representativity of the cohort” (Binder, 1998, p. 104). In the context of longitudinal research, opportunities for temporal analysis will also be limited. In GUS Sweep 9, for instance, only cross-sectional weights were assigned to boost cases. For some surveys this may be adequate – if, for instance, cross-sectional estimates of marginal groups compared to the rest of the population are of interest. In the Family and Children’s Study (FACS), the supplemental sampling that was undertaken was justified by the need to “approximate to a representative sample of Child Benefit recipients in each year” (Conolly, Maplethorpe, & D'Souza, 2009, p. 6).

A monotonic design can therefore be beneficial, insofar as this approach avoids the need to integrate top-up or refreshment cases into the main sample. All cases have scope to be assigned a longitudinal weight and can therefore be analysed as such, so long as the hypothesised rate of attrition is consistent (or lower) than the observed rate of attrition. If there are insufficient numbers of respondents in the marginal group across all waves of the survey, the resulting estimates may not be of the requisite precision. On the other hand, if the assumptions made regarding attrition were too risk averse – i.e. attrition is over-estimated – then this will have cost implications for the survey. Where the marginal group of interest is hard-to-reach – and requires special measures to be taken to encourage response – then the marginal cost would be higher still.

Irrespective of the methodological implications of the approach taken to manage the effects of attrition for marginal groups, practical issues are also an important consideration. Should screening be required due to there being no individual level sampling frame, a monotonic design may be impractical due to the high front-end cost this would entail. Where availability of high-quality administrative data means that there is an appropriate sampling frame, the options available are more open-ended. Just as this enabled a monotonic design in the case of LSIC, it can allow for adjustment to account for losses due to attrition in the case of the Medicare Current Beneficiary Survey (Calderwood & Lessof, 2009, p. 61).

1.6 Implications for survey design

One of the challenges implicit here is that it is difficult to anticipate all analytical needs that will need to be met by a survey – “the client will always require more than is specified at the design stage” (Fuller, 1999, p. 344). This is especially relevant to longitudinal surveys, where the perceived purpose may be liable to change over time.

Key to determining the sampling design for a survey is considering the individual characteristics of the different sub-groups which are of analytical interest. This must then be balanced against the need to draw a sample which is sufficiently large to produce precise estimates for these groups. Without such provision, comparisons across groups may not be possible.

Once the relevant characteristics of the marginal group have been identified, these should inform any decisions concerning how individuals are selected and whether any special measures are required to enable response. If there is a significant distinction between these for the marginal group and the rest of the sample, then a separate study may be necessary. If supplementary samples are likely to be required at a later stage, similar considerations apply here.

Annex A

References

Bankier, M. D. (1988). Power allocations: Determining sample sizes for subnational areas. *American Statistician, 42*, 174-177.

Becker, E., Boreham, R., Chaudhury, M., Craig, R., Deverill, C., Doyle, M., . . . Zaninotto, P. (2006). *Health Survey for England: The health of minority ethnic groups (Volume 1).* Leeds: The Information Centre.

Berthoud, R., Fumagalli, L., Lynn, P., & Platt, L. (2009). *Design of the Understanding Society ethnic minority boost sample.* University of Essex, Institute for Social and Economic Research.

Binder, D. A. (1998). Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodology, 24*(2), 101-108.

Bolling, K., Grant, C., & Sinclair, P. (2008). *2006-07 British Crime Survey (England and Wales). Technical Report. Volume I.*

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., . . . Hughes, C. (2014). Reaching the hard-to-reach: a systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology*, 14-42.

Calderwood, L., & Lessof, C. (2009). Enhancing Longitudinal Surveys by Linking to Administrative Data. In P. Lynn, *Methodology of Longitudinal Surveys.* Chichester: John Wiley & Sons.

Choudhry, G., Rao, J., & Hidiroglou, M. (2012). On sample allocation for efficient domain estimation. *Survey Methodology, 38*(1), 23-29.

Connelly, R. (2011). *Drivers of Unhealthy Weight in Childhood: Analysis of the Millennium Cohort.* Scottish Government Social Research Report. Edinburgh: Scottish Government.

Conolly, A., Maplethorpe, N., & D'Souza, J. (2009). *Families and Children Study (FACS) 2007, Wave 9: Technical report.* Retrieved from http://doc.ukdataservice.ac.uk/doc/4427/mrdoc/pdf/4427facs2007_technical_report.pdf

Costa, A., Satorra, A., & Ventura, E. (2004, January-June). Improving both domain and total area estimation by composition. *SORT, 28*(1), 69-86.

Elam, G., & Fenton, K. (2003). Researching sensitive issues and ethnicity: Lessons from sexual health. *Ethnicity and Health, 8*(1), 15-27.

Elliott, M., McCaffrey, D., Perlman, J., Marshall, G., & Hambarsoomians, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly, 73*, 56-73.

Feskens, R., Hox, J., Lensvelt-Mulders, G., & Schmeets, H. (2006). Collecting data among ethnic minorities in an international perspective. *Field Methods, 18*(3), 284-304.

Feskens, R., Hox, J., Lensvelt-Mulders, G., & Schmeets, H. (2007). Nonresponse among ethnic minorities: A multivariate analysis. *Journal of Official Statistics, 23*(3), 387-408.

Fuller, W. A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological, and Environmental Statistics, 4*(4), 331-345.

Goldstein, H., Lynn, P., Muniz-Terrera, G., Hardy, R., O'Muircheartaigh, C., Skinner, C. J., & Lehtonen, R. (2015). Population sampling in longitudinal surveys. *Longitudinal and Life Course Studies 6(4)*, 447-475.

Harkness, J., Strange, M., Cibelli, K. L., Mohler, P., & Pennell, B. (2014). Surveying cultural and linguistic minorities. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates, *Hard-to-survey Populations* (pp. 245-269). Cambridge: Cambridge University Press.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data and Society, July-December 2017*, 1-12.

Joshi, H., & Emla, F. (2016). The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies, 7*(4), 409-430.

Kalsbeek, W. (2003). Sampling minority groups in health surveys. *Statistics in Medicine, 22*, 1527-1549.

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. *Survey Methodology 35(2)*, 125-141.

Kalton, G. (2014). Probability sampling methods for hard-to-sample populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates, *Hard-to-survey Populations* (pp. 401-423). Cambridge: Cambridge University Press.

Kalton, G., & Anderson, D. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society, 149*, 65-82.

Kan, M.-Y., & Laurie, H. (2018). Who is doing the housework in multicultural Britain? *Sociology, 52*(1), 55-74.

Kish, L. (1976). Optima and proxima in linear sample design. *Journal of the Royal Statistical Society, A, 139*, 80-95.

Kish, L. (1987). *Statistical design for research.* New York: J. Wiley & Sons.

Longford, N. T. (2006). Sample size calculation for small-area estimation. *Survey Methodology, 32*, 87-96.

Lynn, P., Nandi, A., Parutis, V., & Platt, L. (2018). Design and implementation of a high-quality probability sample of immigrants and ethnic minorities: Lessons learnt. *Demographic Research, 38*, 513-548.

Rao, J., & Molina, I. (2015). *Small area estimation* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Singh, M. P., Gambino, J., & Mantel, H. J. (1994). Issues and strategies for small area data. *Survey Methodology, 20*(1), 3-22.

Singh, R. P., Petroni, R. J., & Allen, T. M. (1994). Oversampling in panel surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, (pp. 675-679). Retrieved from http://www.asasrms.org/Proceedings/papers/1994_114.pdf

Smith, P., Pickering, K., Williams, J., & Hay, R. (2010, May). The efficacy of focused enumeration. *Presentation to the Royal Statistical Society Social Statistics Section meeting on 'Special issues in sampling ethnic minorities and migrants'*. Retrieved from https://www.ipsos.com/sites/default/files/migrations/en-uk/files/Assets/Docs/Publications/The_Efficacy_of_Focused_Enumeration.PDF.

Statistics Canada. (2005). *Longitudinal survey of immigrants to Canada: A portrait of early settlement experiences.* Ottawa: Statistics Canada. Retrieved from http://publications.gc.ca/Collection/Statcan/89-614-XIE/89-614-XIE2005001.pdf

Stoop, I. (2014). Representing the populations: what general social surveys can learn from surveys among specific groups. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates, *Hard-to-survey Populations* (pp. 225-244). Cambridge: Cambridge University Press.

Sudman, S., Sirken, M. G., & Cowan, C. D. (1988). Sampling Rare and Elusive Populations. *Science, 240*(4855), 991-996.

Sullivan, A., Joshi, H., Ketende, S., & Obolenskaya, P. (2010). *The consequences at age 7 of early childhood disadvantage in Northern Ireland and Great Britain.* Northern Ireland Office of the First Minister and Deputy First Minister. London: Institute of Education.

Volz, E., & Heckathorn, D. (2008). Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics, 24*(1), 79-97.

Welsh Assembly Government. (2011). *2011 Children and Young People's Wellbeing Monitor for Wales.* Welsh Assembly. Cardiff: Welsh Assembly Government.

Zuccotti, C., & O'Reilly, J. (2019). Ethnicity, gender and household effects on becoming a NEET: An intersectional analysis. *Work, Employment and Society, 33*(3), 351-373.