Random probability vs quota sampling

Paul A. Smith
James Dawber

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.

Random probability vs quota sampling

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

Executive Summary

- Probability sampling has a well-developed, relatively straightforward, design-based estimation framework providing the best approach to making inference about a population.
- Non-probability sampling includes a diverse range of methods that are not easily described under a single framework, however model-based methods are required when making inference from a non-probability sample to adjust for differences between the sample and known population information. Inference from the non-probability sampling method is only as good as the model and assumptions that are used.
- Sampling in longitudinal studies requires a precise definition of the target population, which may not merely be a finite population, but instead a dynamic population or superpopulation.
- Problems with non-response, attrition and under-coverage should be anticipated and factored into the design of the longitudinal study using model-based methods, rather than addressed *post hoc*.
- Having a representative sample is an important aim for a national-level longitudinal study, as this ensures that the data will have a wider potential to be used in the distant future across different disciplines. A probability sample is the best starting point to ensure this.
- Non-probability sampling for longitudinal studies may be useful to supplement the main sample for specific populations which it is impractical to reach with a probability sample, but the ways to analyse such combinations of data from different sample types needs more research.

1 Introduction

Sampling is an essential process in obtaining data from which inference can be made about the wider population. Samples can be collected based on either probability or non-probability sampling methods. Probability sampling methods are characterised by the use of randomisation with known, non-zero probabilities of selection, whereas non-probability sampling methods do not have this property. Several approaches to non-probability sampling are used according to the specific requirements for data collection.

In this report we first briefly introduce probability and non-probability sampling and their respective strengths and weaknesses. We then introduce the idea of design- and model-based inference (plus a hybrid approach) and why this distinction is especially important in non-probability sampling. Next we focus on sampling problems specific to longitudinal studies and the potential to use nonprobability approaches. We briefly address the impacts of opt-in versus opt-out designs. In Annex A, we review existing longitudinal studies with specific emphasis on the use of non-probability sampling methods.

2 Probability sampling

Probability sampling has become the accepted method to make rigorous and reliable inference about a target population. In order to properly undertake probability sampling, a few conventions and assumptions are needed. First, a sampling frame is required which is assumed to exhaustively include all units of the target population. Second, each unit on the sampling frame has a non-zero inclusion probability, which is used to randomly select the units to be in the sample. Generally these inclusion probabilities are used for inference on the population parameter, by weighting observations by the inverse of this probability. Note that this applies in general to sampling designs, however more complex probability sampling designs such as multi-stage (cluster) sampling have additional factors to consider.

Probability sampling is when units are sampled through a randomisation process with the specified inclusion probabilities. When these inclusion probabilities are used as the basis for extending sample traits to the target population we have design-based inference (Koch & Gillings, 2006). Design-based estimators can generally be shown to be unbiased estimators (asymptotically unbiased in model-assisted estimation), as well as providing estimates of the sampling variance – a measure of how different from the true value estimates would be because of the sampling. Theoretically the design-based approach provides the gold standard strategy for social surveys (for a discussion see Lucas, 2016), but in practice there are two complications that can compromise it. The first is that the sampling frame may not have the entire target population, a problem known as under-coverage. The second is due to non-response, where individuals that were randomly selected in the sample are not contactable, or are unable or refuse to participate in the survey. The individuals associated with under-coverage or non-response can represent specific subgroups of the target population, in which case their exclusion would lead to biased estimates. However if the under-coverage and non-response are negligible then a probability sample is said to be representative of the population. A representative sample ensures that inference about the population using appropriate design-based estimators is unbiased. In practice under-coverage and non-response are not often negligible, and we then need to use statistical models in conjunction with the design-based approach to make inferences about the population.

Probability sampling typically often has considerable time and monetary costs to attempt to gather information from all the randomly selected individuals selected in the sample. This is particularly problematic when the interviewers are used and the target population is located over a large geographic area, such as a national household survey, because of the need to employ a large field force. Resources are needed to minimise nonresponse so that the design-based approach can be used. So although design-based probability sampling is guaranteed to be unbiased under perfect conditions, these are typically not met in practice and there is a trade-off of between the resources required to achieve as high a response as possible and the quality costs of additional adjustments.

3 Non-probability sampling

Unlike probability sampling, there is no single framework for non-probability sampling; it is perhaps most easily described as any sampling approach that is not based on the framework of probability sampling. Hence, one or more of a sampling frame, randomisation and known, non-zero inclusion probabilities are not available. Without randomisation the samples are selected based in some way on the researcher's judgement (which may be applied through a statistical model).

The strengths and weaknesses of non-probability sampling are essentially the opposite of probability sampling. Since a non-probability sample cannot be expected to be representative of the target population, then the estimates cannot be guaranteed to be unbiased. However, the costs can be considerably less, since the samples can be acquired more conveniently. Hence the non-probability sampling approach is intended to be more practical at the expense of being theoretically less rigorous.

Non-probability sampling encompasses a broad range of methods varying in sophistication. At its simplest non-probability sampling includes convenience sampling where attempts to reduce bias are not strongly considered. But non-probability sampling methods can be more methodical than this, for example, where auxiliary data can be used to adjust a non-probability sample in an attempt to minimise the potential bias, and these types of nonprobability design with a theoretical basis were considered in Baker et al. (2013). Inference can be made from a non-probability sample with a model-based approach, which utilises a model, which if it holds gives unbiased estimates. The risk of bias is controlled not through the randomness in the sample design (as in probability samples) but through assumptions that are used to build the model and that can only be validated by information external to the study design (Koch & Gillings, 2006). Building a suitable model can involve considerable statistical expertise, and moves the expenditure of resources more towards the estimation phase compared with design-based approaches. The assumed model cannot normally be assessed from the sample information, and if the model does not hold then estimation will be biased. These model-based assumptions are therefore stronger than the assumptions required in design-based sampling, but nevertheless provide a framework for inference from non-probability samples.

Shifts to web-based sampling as well as more accessible administrative data have brought about increased interest in non-probability sampling and corresponding model-based methods for inference. In response to this, Rivera (2018) reviews some non-probability sampling methods with guidelines to researchers about best practices for their implementation. Brick (2011) outlines future directions for survey sampling beyond probability sampling, highlighting that methods for making inference from web panels is of specific importance. He discusses how there are no clearly foreseeable paradigm shifts in sampling theory but suggests that the future of sampling will be dynamic, highlighting how model-based inference and multiple frame surveys may be of specific importance. An important report on non-probability sampling is provided by Baker et al. (2013). The purpose of this report was to examine under what conditions non-probability sampling will yield useful inferences about the population. Among the conclusions that are made are that non-probability sampling methods must be based on models that address challenges relating to both sampling and estimation. They also suggest that the reason why model-based methods are not more frequently used is the difficulty in testing the model assumptions, requiring significant statistical expertise. Lastly, they acknowledge that there is no universal framework for non-probability methods and that the current framework will need to be more coherent before it gets wider acceptance amongst researchers.

3.1 Design-based versus model-based inference

Rather than compare probability and non-probability sampling we focus on design and model-based approaches to inference. This is more relevant to studies which aim to make inference about a wider population, which is typically the case in social surveys. Sterba (2009) provide a good detailed overview of design and model-based frameworks.

Design-based inference is the more common and intuitive method where inference is made from a random sample to a population. This framework was developed by Neyman (1934) and is the basis for probability sampling.

Model-based inference is less commonly used in surveys, although widely used in general scientific inquiry, but with this framework it is possible to make inference from a non-probability sample under certain model assumptions. Note that model-based inference can also be used with probability samples, and this combination may be very efficient. The general model-based framework was introduced by Fisher (1922). For all intents and purposes model-based inference can be thought of as making adjustments to the non-representative sample in such a way that inference can still be justifiably made. Some of these methods will be discussed in the next section, but essentially the model overcomes (or measures) the potential bias that is otherwise induced by the non-probability sampling.

An important concept for model-based inference is the notion of a superpopulation. A superpopulation is a (hypothetical, unobserved) set of units from which the population values are assumed to be realised. Generally, model-based inference is made from the sample to this superpopulation, as opposed to the population as is done in design-based inference. It could be considered as a finding some general relationship rather than describing a relationship in a specific population.

It is important to realise that design and model-based approaches are not mutually exclusive, and in fact they are often used together. For example, a model can be used to adjust for non-response in a design-based survey; and if we have good information on the likely causes of nonresponse, we may be able to use this model information to adjust the design. In this case both the design and the model are required for inference. We discuss other hybrid approaches in Section 5.

Thompson (2015) provides a useful review of the approaches and methods used when using longitudinal complex survey data. She reduces the methods for analysing longitudinal data down to two broad types. The first is a design-based approach with model-assisted approaches to account for non-response and attrition, the second is a purely model-based approach which accounts for design features. She also notes that the latter option requires complex models which tend to exceed the capabilities of readily available statistical software.

4 Model-based methods for non-probability samples

Several methods are used which aim to make inference from non-probability samples using models. These methods include, sample matching, propensity scoring and quota matching.

Sample matching aims to reduce selection bias by matching a non-probability sample to a target population based on a reference data source such as the population census. Certain characteristics of the individuals must first be identified for matching on. For example, sex, age and ethnicity may be the selected characteristics and respondents will be sampled until the sample matches the population composition. In this case the model assumptions are that the selected variables account for all the variation in the outcome, which may be more likely to be true if a large number of characteristics is used.

Another method known as propensity score matching is used when comparing treatment versus control without randomisation. This method relies on modelled conditional

probabilities for each unit given their set of covariates, and assuring that these probabilities do not differ too much between the treatment and control groups.

Quota matching identifies important groups in the population which are related to the outcome measures. These groups are then sampled using convenience sampling and the composition of the groups is fixed to be proportional to known totals of these groups in the population. Like sample matching, it relies on a reliable reference data source to ensure the group proportions reflect those of the population. Selection biases within groups are assumed not to have an effect on the inference.

The common theme with these main methods for non-probability sampling is the assumption of certain relationships between covariates and the outcomes. The more appropriate these assumptions the more accurate the inference that can be drawn from the sample (Baker et al., 2013). In reality the model assumptions cannot account for the full complexity of the problem and biases do remain. The randomisation in probability sampling methods protects against these biases, but there are still sampling errors. As such, utilising design-based and model-based approaches together can be an effective strategy.

5 Hybrid approaches to sampling

Probability sampling is generally the best approach under ideal conditions where the assumptions are met. However, at what point do the coverage and response rate of the sample get so low that a non-probability sample would actually be better? Various studies have attempted to answer this, including Cumming (1990), Yang & Banamah (2014), and MacInnis et al. (2018). There are no simple guidelines about when non-probability sampling may be more appropriate but there is evidence that under non-ideal conditions non-probability sampling methods might offer cheaper or quicker estimates with an insignificant effect on accuracy. Nevertheless, Brick (2011) argues that a well-conducted probability sample with a low response rate is likely to be less biased than a volunteer sample.

Various studies have reported that response rates are declining (for example Curtin et al., 2005), which adds to the difficulty of inference under probability sampling. Furthermore, new challenges have emerged as technology advances. With large proportions of households without home telephones, and mobile phone and internet use growing, designing a probability sample has become challenging simply because there is no longer an effective frame from which the sample can be selected. Implementing a probability sampling design is seemingly becoming more difficult as response rates reduce, and this is leading to a reliance on model-based methods to deal with the non-response and under-coverage. These methods include imputation, weight adjustments and dual/multiple frame designs. With the practical and theoretical limitations of probability sampling and non-probability sampling respectively, a hybrid or "model-aided" approach may offer the best solution.

One of the first hybrid approaches was probability sampling with quotas (Sudman, 1966). This approach anticipates non-response being a problem and so is suggested for studies where call-back is too time- or cost-limiting. The method relies on stratification of the population with strata containing respondents assumed to have an equal probability of responding. The interviewer has a specified route to follow, and fills the quota with respondents as they are available. With probability sampling and the model assumption of equal probability of responding within strata, inference can be made on the population of interest. If the assumption of constant response probabilities within strata is true then no bias will be present. Stephenson (1979) showed that this method produced similar results to

a pure probability sample for most variables in one example.

Another hybrid approach is through a sequential sampling design, which includes multiple and general inverse sampling introduced by Chang et al. (1998) and Salehi & Seber (2004) respectively. This design is useful when targeting hard-to-reach subgroups, and basically simple random sampling (SRS) is undertaken repetitively until a desired sample size is reach for targeted strata. Similar to sequential sampling is adaptive sampling where more generally the repetitions in the sampling are conditional on what is already observed. For example adaptive cluster sampling will first undertake SRS followed by resampling around individuals with certain clustered characteristics. Sampling hard-to-reach subgroups is a challenging area with various techniques used, such as network sampling, snowball sampling, capture-recapture methods and respondent-driven sampling. Shaghaghi et al. (2011) and Heckathorn and Cameron (2017) provide an overview of these methods.

Berzofsky et al. (2009) present a summary of hybrid methods before applying a method to national business establishment surveys in the USA. This "model-aided sampling" technique is a hybrid between probability, quota and general inverse sampling which can be used without any introduction of bias. Occupations that have lower response are targeted through this method, ensuring that all groups are well represented, with appropriate weightings. They also argue that response burden is reduced overall by not needing larger samples than specified in occupations that do have high response rates. Elliott (2009) also presents a hybrid approach to estimation from a non-probability sample using "pseudo-weights" and a probability sample with similar predictive covariates.

6 Sampling in longitudinal studies

6.1 Population definition

Sampling in a longitudinal study has additional difficulties to cross-sectional studies. The two primary difficulties are around how the population should be defined, as well as how to manage attrition (Smith et al. 2009).

Defining a population from which to sample is more complicated for longitudinal studies, because the population is not necessarily fixed at a selected time point. Hence the population must not only be defined geographically, but also temporally. Furthermore, it may be that the superpopulation needs to be defined as well, for model-based inference. The representativity of a sample can only be defined relative to a suitable population or frame. Defining a (super)population is important for the design of the study, the probabilities used in the sampling design, the weighting methods and the focus of the analysis. For example, a sample can be representative of the population at the start, end or throughout the study period, each requiring different implications to the study design.

6.2 The need for representativeness

Goldstein et al. (2015) present a discussion on the importance of representativeness, particularly for longitudinal studies. Representativeness is a key factor, and challenge, when assessing the sampling design of a study. Some of the key ideas from this paper are discussed below.

An important distinction between scientific and population inference is highlighted. For the

latter it is important to have a representative sample, whereas in the former it is more important to cover the range of characteristics in the population. Scientific inferences are concerned with establishing causal relationships, which do not necessarily have to be population-wide or population-specific relationships. In such cases a purposive sampling approach could be as or more appropriate rather than probability sampling. For example randomised control trials do not require probability sampling, only a random assignment of cases to groups. The purpose here is to establish causality from a treatment, not to generalise to a specific population. Hence the purposes of the longitudinal study must be considered, and what kind of inference will be best suited to those purposes.

Generally the purpose of a longitudinal study is to enable longitudinal estimation, so the population must be longitudinal also. A longitudinal population is dynamic in nature, meaning that people will join and leave the population over time. Making inference on a dynamic population creates difficulties which is why the population may be restricted to be the cross-sectional population at the beginning of the study. This requires the assumption that the population will not change much, however over long periods this assumption becomes hard to justify. Lynn in Goldstein et al. (2015) argues that when defining a population the representativeness of the sample should be considered not just in relation to a study population but also a policy population at which the study is targeted, that is, the population which will be affected by the study findings. The target population should then be carefully defined at the beginning of the study, from which the appropriateness of design versus model-based approaches can be assessed.

It is worth pointing out that the population defined in the study has immediate consequences on how the parameters should be defined. Furthermore, if the statistical bias is defined as the difference between a sample estimate and the population parameter, then the way the population is defined has implications on how the bias is defined. If probability sampling is optimal for reducing this bias, and it is not well defined then perhaps non-probability sampling may be considered. These factors should all be considered in conjunction when designing a study.

A case is made in the discussion by O'Muircheartaigh in Goldstein et al. (2015) that probability sampling should be used in longitudinal studies to maximise the cross-disciplinary collaboration of the data. This is because in no discipline is a probability sample inferior to a non-probability sample. This is a strong indication that a design-based approach should be attempted, and then model-based methods used to supplement the unavoidable limitations of the sample representativeness. It is also suggested that these model-based approaches with their corresponding model assumptions should be testable and well-justified in practice.

6.3 Attrition

Attrition is unavoidable in longitudinal studies, and various weight adjustment approaches have been suggested, but all require the use of models. For example, Schmidt and Woll (2017) suggest an approach based on logistic regression where probabilities of drop-out are predicted using key auxiliary variables, and the weights are inversely proportional to these probabilities. Cumming and Goldstein (2016) argue that reweighting methods can be inefficient and propose multiple imputation methods.

6.4 Non-probability samples in longitudinal studies

The key sampling decisions in a longitudinal study are made with respect to the initial recruitment. In this section we provide some thinking on two topics – on-line panels, and network samples for scarce populations. A few examples of nonprobability samples in longitudinal surveys are given in Annex A.

On-line panels have become an important way to gather information on a range of topics, and many research organisations have invested in them. "On-line panel" however is not a generic term describing how they are recruited (Baker et al. 2013), and recruitment can vary from opportunist through advertising, which presents almost no control over the composition of the panel, through quota-based methods where recruits are only added when they have characteristics which are relatively underrepresented in the panel, through to a panel which is recruited through a probability mechanism (Blom *et al.* 2016). The NatCen panel (see Annex B) is an example of the latter, recruited from participants in the British Social Attitudes Survey (BSA).

Clearly an on-line panel does not cover those who do not use the internet; but for the remaining population an adjustment to known control totals can be made using a suitable model. Then the question is how well the model can account for differences between panel members and those who did not participate. In the NatCen panel the previous responses from the BSA are available to examine these differences, and an evaluation of the characteristics of the panel (particularly as they change longitudinally over several waves) would be valuable. Otherwise, the efficacy of the model-based adjustments cannot be evaluated from study itself, and would need to be compared with an external source. Potentially the population spine suggested by Davis-Kean et al. (2017) to support future longitudinal infrastructure could act as this external source; depending on which variables are included, it may however not have sufficient information to distinguish the characteristics that are most relevant. In a general purpose survey the available variables could not be expected to be good predictors for all the characteristics of interest.

The risk that an on-line panel would over-represent certain population subgroups and thus affect population inference means that it is unlikely to be suitable for a large longitudinal study, even if the sample could thereby be much larger. But could it provide a supplementary sample and thereby improve scientific inferences? This situation is not unlike the Life Study's probability-based national sample plus centre-based samples. In either case there is a need to combine the information ("borrow strength") to use the unbiasedness of the probability sample but also take advantage of the larger nonprobability sample. This can be approached in a way which seeks to minimise the mean squared error, but this requires complex models, and providing tools to support regular analyses of this type seems challenging (Goldstein in Goldstein et al. 2015).

A further challenge in longitudinal studies is to have sufficiently large samples in a series of subpopulations of policy interest; how best to achieve this is considered in work package 2. If it is important to include them in the main longitudinal study, we may be faced with sampling enough cases (or indeed *over*sampling enough to account for expected attrition, which is often higher in subpopulations of interest). For populations with specific characteristics this may perhaps be achieved using network sampling approaches such as adaptive cluster sampling (Thompson 1990), provided that "neighbouring" units can be defined (eg, but not necessarily, geographically) or respondent driven sampling (Heckathorn 1997, Heckathorn & Cameron 2017) for socially connected populations.

Both these approaches have less control over the sample characteristics than a probability-based approach. They have been widely used for specialist populations where there is no frame, but as a supplementary sample suitable models would be needed to borrow strength, as discussed for on-line panels. Further research would be needed to investigate how these approaches could be integrated, but this does not seem a priority area currently. We did not find any examples of nonprobability samples being used as boosts in major longitudinal studies, and it may well be that studies of specialist populations are better undertaken and analysed separately.

There is perhaps a need for a pilot investigation of the longitudinal stability of samples collected by network sampling approaches (particularly respondent driven sampling where it is not the researcher doing the recruitment). The first response is the most difficult, and in a longitudinal setup contact information would be available for further waves; but the attrition in these groups might be high.

6.5 Conclusions

The ESRC-funded longitudinal studies are a key component of the UK's data infrastructure, and therefore have multiple uses, including both scientific and population inference. To support both approaches, probability samples are the only currently available practical approach which provides an uncontested basis for both approaches, and additionally offers a relatively straightforward design-based analytical pathway to a wide range of researchers. The availability of this approach does not preclude the use of model-based approaches, and indeed the discipline of probability sampling may offer additional protection against bias in the modelling by producing samples with known properties.

Further research on the appropriate models and tools for combining probability and nonprobability elements would be valuable (Skinner in Goldstein et al. 2015). If sufficiently generic approaches can be found, this may open the way to making nonprobability components of the major longitudinal surveys viable and thereby increase the available information at relatively low cost.

7 Opt-in versus Opt-out

A probability sampling approach relies on response rates being high enough to ensure sample representativeness. Recent findings from the Life Study revealed that response rates were too low to provide a representative sample (Dezateux et al., 2016). A factor that was believed to cause this was the opt-in recruitment model used, as opposed to an opt-out model. The former being when consent to participate is affirmed, rather than withdrawn in the case of the opt-out.

Previous studies have assessed the response rate differences between opt-in and opt-out approaches. Junghans et al. (2005) performed a randomised control trial in two general practices in England, where patients with angina were recruited. They found that the recruitment rate was 38% (96/252) for opt-in and 50% (128/258) for the opt-out approach. In Pennsylvania, USA a similar trial was done on diabetic patients, where a recruitment letter for opt-in got 12.7% (63/496) of patients enrolling compared to 38.3% (28/73) for opt-out (Aysola et al., 2018). More relevantly to longitudinal studies, Bray et al. (2015) compared opt-in and opt-out approaches aiming to reengage dropped out participants in the Avon Longitudinal Study of Parents and Children (ALSPAC). They found that only 3% (4/150) in the opt-in arm consented to continue, compared to 31% (46/150) in the opt-out

arm. There is also evidence that opt-out relative to opt-in significantly improves responses rates in children (Spence et al., 2015) and adolescents (Severson and Ary, 1983).

There is evidence that opt-out approaches to recruitment will receive higher response rates in a number of settings. In the examples provided, the response rates are below 50% for both approaches. This relatively low percentage creates challenges in making the sample representative. But given these challenges it seems sensible to aim for opt-out approaches to recruitment if it is ethical to do so. Then adjustments to non-response can be made using auxiliary data and appropriate weighting or model-based methods.

8 Recommendations and further research

- For a national-level longitudinal study probability sampling should be used to ensure representativeness, and provide a robust resource which can be used for general purposes into the future. Non-probability samples do not currently offer sufficient protection against the risk of bias in population inference; scientific inference does not deteriorate if based on a large probability sample.
- Models for nonresponse and attrition rates should be used to adjust the sample requirements in the design stage of the longitudinal study; model-based methods will still be needed to make adjustments for the actual outcomes. Auxiliary data sources will be useful for the modelling, potentially including linked administrative data.
- Longitudinal study designs should consider precise definitions of the population, allowing for a clear understanding of what should be sampled.
- Non-probability sampling for longitudinal studies should only be considered in situations where probability designs are not feasible (for example if there is no reliable sampling frame, or if the target population is hard to reach). There may be potential to use nonprobability samples to boost a probability sample for certain population subgroups, but the challenges of producing suitable models to combine these two types of data are not yet worked out, and more research in this area is needed.

9 References

Aysola, J., Tahirovic, E., Troxel, A. B., Asch, D. A., Gangemi, K., Hodlofski, A. T., … Volpp, K. (2018). A Randomized Controlled Trial of Opt-In Versus Opt-Out Enrollment Into a Diabetes Behavioral Intervention. American Journal of Health Promotion, 32(3), 745–752. https://doi.org/10.1177/0890117116671673

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. Journal of Survey Statistics and Methodology, 1(2), 90-143. https://doi.org/10.1093/jssam/smt008 (see also the full report at https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf).

Berzofsky, M., Williams, R., & Biemer, P. (2009). Combining probability and non-probability sampling methods: Model-aided sampling and the O*NET data collection program. Survey Practice, 2(6), 2984. https://doi.org/10.29115/SP-2009-0028

Blom, A., Bosnjak, M., Cornilleau, A., Cousteaux, A.S, Das, M., Douhou, S & Krieger, U.

(2015) A comparison of four probability-based online and mixed mode panels in Europe. *Social Science Computer Review*, 34, 8-25. https://doi.org/10.1177/0894439315574825.

Bray, I., Noble, S., Boyd, A., Brown, L., Hayes, P., Malcolm, J., ... & Molloy, L. (2015). A randomised controlled trial comparing opt-in and opt-out home visits for tracing lost participants in a prospective birth cohort study. BMC medical research methodology, 15(1), 52. https://doi.org/10.1186/s12874-015-0041-y

Brick, J. M. (2011). The Future of Survey Sampling. Public Opinion Quarterly, 75(5), 872–888. https://doi.org/10.1093/poq/nfr045

Cumming, R. G. (1990). Is probability sampling always better? A comparison of results from a quota and a probability sample survey. Community health studies, 14(2), 132-137.

Cumming, J. J., & Goldstein, H. (2016). Handling attrition and non-response in longitudinal data with an application to a study of Australian youth. Longitudinal and Life Course Studies, 7(1), 53-63. http://dx.doi.org/10.14301/llcs.v7i1.342

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. Public opinion quarterly, 69(1), 87-98. https://doi.org/10.1093/poq/nfi002

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., Colson, D., Brocklehurst, P., & Elias, P. (2016). Life after Life Study. In Report of a Scientific Meeting held at The Royal College of Physicians, London, UK, 14th January. http://discovery.ucl.ac.uk/1485681/

Elliott, M. (2009). Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. Survey Practice, [online] 2(6), 1-7. https://doi.org/10.29115/SP-2009-0025

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A 222, 309-368.

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.

Heckathorn, D. D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. Social Problems, 44, 174–199.

Heckathorn, D. D., & Cameron, C. J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. Annual Review of Sociology, 43, 101-119. https://doi.org/10.1146/annurev-soc-060116-053556

Junghans, C., Feder, G., Hemingway, H., Timmis, A., & Jones, M. (2005). Recruiting patients to medical research: double blind randomised trial of "opt-in" versus "opt-out". BMJ, 331:940. https://doi.org/10.1136/bmj.38583.625613.AE

Koch, G. G. and Gillings, D. B. (2006). Inference, Design-Based vs. Model-Based. In Encyclopedia of Statistical Sciences (eds S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic

and N. L. Johnson). https://doi.org/10.1002/0471667196.ess1235.pub2

Lucas, S. R. (2016). Where the Rubber Meets the Road: Probability and Nonprobability Moments in Experiment, Interview, Archival, Administrative, and Ethnographic Data Collection. Socius. https://doi.org/10.1177/2378023116634709

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension. Public Opinion Quarterly, 82(4), 707-744.

Nærde, A., Janson, H., & Ogden, T. (2014). BONDS (The Behavior Outlook Norwegian Developmental Study): A prospective longitudinal study of early development of social competence and behavior problems. ISBN 978-82-93406-00-6. Oslo, Norway: The Norwegian Center for Child Behavioral Development.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97(4), 558-625.

Rivera, J. D. (2018): When attaining the best sample is out of reach: Nonprobability alternatives when engaging in public administration research, Journal of Public Affairs Education. https://doi.org/10.1080/15236803.2018.142982

Salehi, M., & Seber, G. A. (2004). A general inverse sampling scheme and its application to adaptive cluster sampling. Australian & New Zealand Journal of Statistics, 46(3), 483-494.

Schmidt, S. C., & Woll, A. (2017). Longitudinal drop-out and weighting against its bias. BMC medical research methodology, 17(1), 164.

Severson, H. H., & Ary, D. V. (1983). Sampling bias due to consent procedures with adolescents. Addictive Behaviors, 8(4), 433-437. https://doi.org/10.1016/0306-4603(83)90046-1

Shaghaghi, A., Bhopal, R. S., & Sheikh, A. (2011). Approaches to recruiting 'hard-to-reach'populations into research: a review of the literature. Health promotion perspectives, 1(2), 86. https://doi.org/10.5681/hpp.2011.009

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.

Spence, S., White, M., Adamson, A. J., & Matthews, J. N. (2015). Does the use of passive or active consent affect consent or completion rates, or dietary data quality? Repeat cross-sectional survey among school children aged 11–12 years. BMJ open, 5(1). https://doi.org/10.1136/bmjopen-2014-006457

Sterba S. K. (2009). Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. Multivariate behavioral research, 44(6), 711–740. https://doi.org/10.1080/00273170903333574

Stephenson, B. C. (1979). Probability Sampling with Quotas: An Experiment, Public Opinion Quarterly, 43, 477–496. https://doi.org/10.1086/268545

Sudman, S. (1966). Probability Sampling with Quotas. Journal of the American Statistical Association, 61(315), 749-771. https://doi.org/10.2307/2282785

Thompson, M. E. (2015). Using Longitudinal Complex Survey Data. Annual Review of Statistics and its Application, 2, 305–320. http://dx.doi.org/10.1146/annurev-statistics-010814-020403

Thompson, S.K. (1990). Adaptive cluster sampling. Journal of the American Statistical Association 85(412): 1050–1059. https://doi.org/10.1080/01621459.1990.10474975.

Thurber, K. A., Banks, E., Banwell, C., & LSIC Team (2015). Cohort Profile: Footprints in Time, the Australian Longitudinal Study of Indigenous Children. International journal of epidemiology, 44(3), 789–800. https://doi.org/10.1093/ije/dyu122

Wilson, I., Huttly, S. R., & Fenn, B. (2006) A Case Study of Sample Design for Longitudinal Research: Young Lives, International Journal of Social Research Methodology, 9:5, 351-365. https://doi.org/10.1080/13645570600658716

Yang, K., & Banamah, A. (2014). Quota Sampling as an Alternative to Probability Sampling? An Experimental Study. Sociological Research Online, 19(1), 1–11. https://doi.org/10.5153/sro.3199

Annex A: Longitudinal study examples

In this section we review national level longitudinal studies from around the world, focussing on whether probability or non-probability sampling was used and how they were implemented. We reviewed longitudinal studies from many countries, including the UK, USA, Australia, New Zealand, Ireland, Denmark, South Korea, Netherlands, Germany, and South Africa. With the exception of four studies, all designs that were reviewed were based on probability sampling, most commonly with cluster and stratified sampling and oversampling of targeted subgroups such as ethnic minority groups. Very often the reasoning for probability sampling was to ensure the sample was nationally representative. In some cohort studies no sampling was undertaken as the entire population were asked to be recruited to the study.

Wilson et al. (2006) present a non-probability sampling design for the Young Lives longitudinal study covering four developing countries, in which "sentinel sites" were chosen purposively, based on evidence of their characteristics. Within these sites, probability sampling was used. The foremost reason for a non-probability design was the lack of an accessible and accurate sampling frame. The way in which sites were chosen and local recruitment undertaken attempts to make the study as representative as possible of the populations in the four countries considered, but there was no formal framework to ensure representativity.

Thurber et al. (2015) discuss a two-stage purposive sampling design for the Australian Longitudinal Study of Indigenous Children (LSIC). Due to the limited accessibility of the target population 11 sites were purposively selected across the country based on rurality and willingness for the community to participate. Within these sites, families were recruited to the study based again on purposive sampling as well as snowball sampling. It is noted that the purposive sampling is a limiting factor which inhibits inference on the wider Indigenous population, but does not prevent internal comparisons and longitudinal analyses.

The Behaviour Outlook Norwegian Development Study (BONDS) is a longitudinal study focussing on developmental pathways of children (Nærde et al., 2014). The aims of the study were to identify risk factors for a child's social and behavioural development and how these factors relate to childcare and early schooling. Five sites in different municipalities of south-east Norway were selected with demographics "approximating those at the national level". Children were recruited non-randomly and a power analysis was performed based on a growth mixture model, suggesting a model-based sampling framework. There were also quotas based on gender and the demographics of the sample was compared to the national proportions. While representativeness was discussed in the design, no probability-based sampling methods were used. Importantly, none of the aims of the longitudinal study were to make inference about a wider population. This would justify the non-probability sampling design, which is understandable given that the focus is on finding specific associations in developmental pathways, i.e. making scientific inferences, rather than attempting to estimate at a national level, i.e. making statistical inferences.

The recent Life Study in the UK used two components, one aiming at a representative sample of babies based on a clustered probability sample and the other a purposive sample of pregnant women and their partners from participating maternity units. The maternity units were selected to capture a range of ethnic and social backgrounds, with over-representation of births to mothers from black and minority ethnic groups. The two components of the study were designed to allow for data to be combined. Nevertheless, it

is clear that the two components can be used together to strengthen any inference made, using similar methods to Elliot (2009) for example. The Life Study was stopped prematurely for reasons discussed in Dezateux et al. (2016).

These four longitudinal examples highlight three different scenarios where non-probability sampling is favourable to probability sampling: when there is no reliable sampling frame, when targeting hard-to-reach populations, and when scientific inference rather than inference about population quantities is the aim of the study.

Annex B

Random versus non-random probability sampling in longitudinal surveys

NatCen
Social Research that works for society

Alina Carabat

1. Overview

In survey design and implementation, and the analysis of survey data for research and policy purposes, the issue of sampling approach is of key importance for several reasons which have theoretical and practical implications. In statistics and survey methodology, sampling refers to the selection of a subset of individuals from within a statistical population to estimate characteristics of the entire population of interest. Primarily, such sampling methods are divided into two broad categories – random probability and non-random probability sampling.

Random probability sampling refers to a sampling method in which all the members of a specified population have a known and non-zero chance to be a part of the sample (and that the sample may be reliably replicated through this sampling methodology). This sampling strategy helps to reduce the possibility of sample/population bias.

When it comes to non-random probability, rather than a random selection of sample members, selection relies on the subjective specification of the researcher. As a result, conclusions drawn by the researcher cannot be inferred from the sample to the whole population.

The following sections will discuss the strengths, advantages, challenges and limitations of random probability sampling versus non-random probability sampling in relation to longitudinal surveys with examples.

2. Random probability sampling in longitudinal surveys

Longitudinal surveys are generally characterised by a need for population representative sampling as a main strategy. According to Lynn (in Goldstein et al., 2015) such features include the limited advance knowledge about estimation parameters, the inability to specify all estimation requirements to flow from the study in advance as well as the long scale timeline between data collection and analysis/policy implementation.

More often than not, longitudinal surveys aim to have a stable sample with high response rates across multiple waves. Commonly, the initial sample of such longitudinal surveys is representative of the population of interest. One of the advantages of having a representative initial sample is that the initial respondents can then be used to adjust analysis at subsequent waves to account for any bias arising from attrition (Goldstein et al., 2015).

Because the purpose of longitudinal studies is to be used as data resources for primary/secondary analysis it is important that the initial sample is representative of the population. Lynn (in Goldstein et al., 2015) argues that this can be regarded as a safety net ensuring that the population distribution of (multiple) topics of interest can be covered therefore allowing for the analysis of various research topics not envisioned at the design stage.

Further benefits of probability sampling, according to Lehtonen (in Goldstein et al., 2015) are in the flexibility it allows for the control selection of participants as well as its ability to "provide a basis for proper statistical inference under the actual sampling design" (p.469).

However, one of the disadvantages that longitudinal studies can't go around is that outcomes measured longitudinally can only be based on those people who were in the 'population' at both points in time (i.e. baseline and outcome). Those people who 'enter the population' after the baseline measure (people being born, people emigrating into the country) as well as people who leave the 'population' before the outcome measurement (through death or immigration) cannot contribute to the estimate (Lynn in Goldstein et al., 2015).

Muniz-Terrera and Hardy (in Goldstein et al., 2015) highlight the complication that even in situations where a random, representative sample is selected at the onset of a longitudinal survey, this representativeness is not likely to last over time. The original sample is bound to change, while the study population will also change due to loss at follow-up. They use the example of the MRC National Survey of Health and Development (NSHD), for which the initial sample was made up of:

- All babies born to women with husbands in non-manual and agricultural employment
- One in four births by women with husbands in manual employment

Wadsworth (1991) explains how the aim of employing this sampling scheme was to achieve similar numbers of children in both groups. With the sample aged 69 at the latest (24th) data collection, the participants are no longer representative of the entire population aged 69 in England, Scotland and Wales. Because of population demographics changing naturally but also due to immigration and emigration occurring over the lifetime of the cohort, any estimates coming from this study can only be representative of the British-born population in 1946. This constitutes a disadvantage of random samples for longitudinal surveys in the long-term.

One way of dealing with this would be for national cohort studies to boost the sample in such a way as to maintain representativeness such as the 1958 British Birth Cohort (NCDS) and the 1970 British Birth Cohort which topped up the sample with immigrants born in the cohort member reference birth week using school records. The original birth sweep for NCDS managed to collect information about 17,415 (98%) of all new-borns in Great Britain in that week. In three instances at ages 7, 11 and 16, the sample was topped up with children (and subsequently teenagers) who had been born overseas in the relevant week and had subsequently moved to Great Britain. The new augmented sample was made up of 18,558 cohort members (Cls.ucl.ac.uk, 2019).

3. Non-random probability sampling in longitudinal surveys

It is generally accepted that random sampling approaches are preferred to non-probability sampling because they can be generalised to population characteristics. However, non-random samples offer cost savings if their lack of representativeness can be overcome.

Ebrahim and Davey-Smith (2013) argue that results coming from non-random probability samples are often different compared to results coming from random probability samples. In their opinion one of the disadvantages of non-random samples is the element of volunteer bias that comes into play for non-random probability studies has the effect of distorting such surveys/studies. One such example is the American Cancer Society volunteer cohort. The advantages of using such samples include the ease to follow up, participants motivation to stay in the study and the fact that, when it comes to cancer research, events are likely to be easy to count. One of the results coming out of this cohort showed that high alcohol consumption was associated with a reduced risk of stroke, which is surprising considering that alcohol is known to increase blood pressure which is, in turn, a main cause of stroke.

This begs the questions of what type of heavy drinkers volunteer for studies about the health effects of their lifestyle? The argument here is that they are unlikely to be representative of the heavy drinkers in the population (they may not smoke, may exercise vigorously) and the factors that make them non-representative will tend to render them at lower risk of stroke. Having such a non-representative cohort leads to potentially spurious results, especially when factors that are associated with the outcome of interest are also likely to be linked to self-selection into a study.

Further disadvantages of non-probability samples (in longitudinal studies and not only) as described by O'Muircheartaigh (in Goldstein et al., 2015), include lowering the acceptability of data and conclusions across different fields of study. In his opinion data coming from non-probability samples lack the capacity to maximise potential for cross-disciplinary work and publication, as non-probability samples are generally perceived as questionable and treated with caution. He goes on to argue that the allocation of resources and acceptability of conclusions drawn from such samples are likely to be lower than those based on probability samples which are commonly perceived as producing generalisable inferences. His argument concludes that when it comes to both science and policy "a probability sample is superior to a non-probability sample, [and] representation trumps convenience" (p.464). At the same time, the argument is that probability methods/samples are the best way to obtain estimates that can be passed as representative of the population.

4. Integrating random and non-random probability samples in longitudinal studies

While probability and non-probability samples have both been a big part of social research for a long time, few attempts seem to have been made to actually combine the two. O'Muircheartaigh (in Goldstein et al., 2015) argues that the extent to which combined samples can be used to draw conclusions regarding the entire population rests in great part on either:

- Both being probability samples (not of interest here)
- Making model-based assumptions in order to allow inferences to be made from the non-probability sample to the combined sample and consequently the population

According to Elliott (2009), the most obvious reasons to make use of non-probability samples when probability samples are already available are:

- The non-probability sample contains the outcome of interest
- The probability sample is not large enough and needs to be topped-up with non-probability sample

Furthermore, an argument made by Goldstein et al. (2015) claims that the idea that longitudinal studies should always be representative of 'real' populations can be problematic. The argument here is that for the purpose of most analysis a high degree of heterogeneity is needed and this can generally be achieved by sampling purposively. Such an example is the Life Study longitudinal survey which combines purposive sampling with random sampling which will be discussed later.

Survey methodologists are starting to propose various approaches to improve the quality of analyses obtained from such combined datasets, as further detailed in this section.

4.1 A Partially Successful Attempt to Integrate a Web-Recruited Cohort into an Address-Based Sample

The data used in this paper by Kott (2019) came from a 2015 web-and-mail survey conducted in Oregon. Two thirds of the (respondent) sample was randomly selected from Oregon addresses, however, about half of these actually responded by web, while the rest responded by mail questionnaire. The other third of the (respondent) sample was non-probability, recruited via

Facebook and responded by web. Therefore, the survey was made up of three respondent groups: a web group (n=640), a mail group (n=722) and a non-probability recruit group (n=627). According to Kott, exploratory analysis of the unweighted data suggested the non-probability recruit group was akin to the web respondents but fairly different from the mail group.
The randomly selected respondents (web and mail groups) were calibrated and scaled to variable totals from the 2014 American Community Survey. The non-random Facebook recruit cohort was then calibrated and scaled to the randomly selected web respondents using the same variables with the addition of 'political affiliation'. The SUDAAN 11 procedure WTADJX was used for both calibrations.

Once both groups were calibrated to the corresponding population proportions, the DIFFVAR statement in WTADJX was used to evaluate any differences in the calibrated estimates between the two groups, as proposed by Kott (2006). For these purposes 40 variables were investigated based on 20 survey items. Half of these were to do with whether (or not) there was a valid response to each of the 20 items, while the other half were the responses to the 20 items when valid. To determine the significance of differences between the two groups (random probability vs. non-probability), a Holm-Bonferroni procedure was used (Holm, 1979).

The way this procedure works is by sorting the 40 differences in ascending order by their p-values. According to Kott "The smallest difference is deemed significant at the 10% level when its p-value is less than 0.1/40. If it is, the second smallest difference is deemed significant at the 10% level if its p-value is less than 0.1/39, and so forth until a difference is not deemed significant" (p.97). Differences were assessed at the 10% level rather than the conventional 5% level as the Bonferroni procedure is deemed 'conservative'.

For this data, the Bonferroni-adjusted p-values were not used to analyse the combined data as such but rather to determine the stability of the assumption that the two samples coming from different groups were measuring the same things. The conclusion of Kott's analysis is that only "one or two (out of 40 or 20) variables had significant differences at the 10% level" (p.98). However, even a single significant difference between the two groups is enough for the overall Bonferroni test of equivalence of the two groups to fail.

It is worth noting Kott's highlight that certain items had frequent missing data instances which would have been useful to make the calibration weighting more accurate, concluding that more thought needs to go into the analysis/weighting to be done with such samples before the surveys are design and the data collected.

4.2 Life Study

The British birth cohort known as the 'Life Study' is made up of 60,000 mothers (recruited during pregnancy between 2014 - 2018) within small but heterogeneous geographic clusters (treated as a random sample for those geographic strata) as well as a UK random sample (recruited during the same time period) of 20,000 live births (treated as a random sample for the whole country). The pregnancy component is defined as all the pregnant women attending a set of maternity units between 2014 - 2018. Maternity units were selected to represent births to mothers from a range of ethnic and social backgrounds, with over-representation of births to mothers from black and minority ethnic groups.

Goldstein (in Goldstein et al., 2015) notes the advantage of this design is that for population estimates making use of variables collected in the birth component there is additional information available from the (considerably larger) birth component to improve their accuracy. This can be achieved by using appropriate weights (computed from nationally available birth data). The possibility to carry out combined analysis on these two samples has the potential advantage of increasing the power of detecting differences in the population. He goes on to acknowledge that in order to be able to successfully exploit such a design, appropriate tools that can 'borrow strength' across the two components are needed, although he accepts such 'tools' might be challenging to find.

He also argues that one way to deal with this integrated design would be to weight inversely by the probability of selection (for those who these applies to), with members of the maternity unit sample being given selection weights equal to one. Skinner (in Goldstein et al., 2015) disagrees, assuming this is only likely to increase the effective sample size of the 'birth sample' by a small fraction.

4.3 The Great British Class Survey

A discussion on a recent review of non-probability sampling conducted by the American Association of Public Opinion Research can be found in Baker et al. (2013). The review notes that combining nationally probability samples with geographically clustered non-representative samples is not a common approach. In the same paper Baker et. al. cover the topic of weighting highlighting that "the main concern with model-based inferences from non-

probability samples is that population estimates are highly dependent on model assumptions" (p.97).

The Great British Class survey is an example of such a combination of non-probability sample (approx. 161,000 web respondents) and a smaller but representative quota sample (approx. 10,000 respondents). The authors themselves, (Savage et al., 2014) accept the criticism that their design is 'unorthodox' and highlight that their effort should be regarded as an 'experiment'. In light of this, Skinner (in Goldstein et al., 2015) concludes that the reliability and efficiency of national estimates coming out of such integrated designs is questionable and represents a current topic in need of further study.

5. Probability and non-probability web panels

While web surveys are a known and viable model for well-defined populations, for which exhaustive sampling frames exist (such as university students or company employees), surveying general populations in this way is not possible. One solution to this, adopted by commercial and non-profit organisations alike, is recruiting participants into online panels (Göritz, 2004a). Such samples, coming from online panels have been shown to produce fewer break-offs (O'Neil, Penrod, & Bornstein, 2003) and higher response rates than other online recruitment approaches (Schillewaert & Meulemeester, 2005). However, it should be noted that such high response rates can be explained by the fact that non-respondents have already selected themselves out during the panel recruitment stage and therefore, this can't simply be regarded as a good indicator of sample quality.

The main disadvantage of online panels has to do with non-probability selection, whereby respondents volunteer for panels, mainly for monetary rewards. The obvious risk here is that such panels will mainly attract consistent internet users who are primarily financially motivated. Bethlehem and Biffignandi (2012) actually go as far as to show that the worst-case bias of a large self-selection survey can be up to "a factor 13 x larger" (p. 313) than the bias of a small probability survey. In order to offset this, Couper (2000) advises against using such volunteer panels and focus instead on pre-recruited, probability panels which contact potential members using established offline probability-based techniques.

6. The NatCen Panel

Such an example of a probability-based panel is the NatCen Panel, the only probability-based research panel in Great Britain open for the social research community to use. The approach to recruiting panellists was on the back of a pre-existing survey, in this instance the British Social Attitudes survey (BSA). This is a random probability face to face survey of people aged 18+ across Great Britain and by recruiting from BSA, the NatCen Panel was able to maintain a probability-based design. Those interviewed as part of the BSA were asked to join the panel at the end of the BSA interview. For each panel fieldwork wave, all active panellists are invited to take part, sustaining the principle that the population has a known and non-zero chance of being selected, enforcing the random probability design. This random probability design, alongside the use of telephone fieldwork to include people who are not avid internet users, has the potential to reduce the risk of bias and allow the panel to maintain quality (Jessop in the-sra.org.uk, 2019).

6.1 Longitudinal response to the NatCen Panel

Although the NatCen Panel was designed with cross-sectional research in mind, its design is more akin to a longitudinal study, and as such longitudinal analysis is possible. However, for this to work, a high 're-interview' rate is required to ensure estimates are representative of the general population. (Natcen.ac.uk, 2019). Considering the seven waves conducted between November 2016 and October 2017 (waves using sample recruited from BSA 2015 and 2016), between 82% and 92% of panellists taking part in any one wave took part in any one other, whilst 66% of panellists interviewed in November 2016 took part in all seven waves (to October 2017), suggesting longitudinal analysis is feasible with the NatCen Panel design (Jessop in the-sra.org.uk, 2019).

6.2 Lifetime Gifting: blending probability and non-probability samples

For the Lifetime Gifting survey, the research objectives required two separate samples to be interviewed:

- the general population (to understand the levels of gifting)
- people who gifted in the two years before the interview (to understand gifting in detail)

To achieve this, a general population sample of adults aged 18 and over in Britain was recruited from the NatCen Panel. The sample of gifters was made up of those from the general population identified as gifters during the first stage of data collection and a 'boost' sample coming from the PopulusLive panel, a non-probability opt-in panel. The sample was designed to be representative of the gifter population, with quotas set on sex, age, region and wealth (Assets.publishing.service.gov.uk, 2019).

The data was then weighted to be representative of the general or gifter population as appropriate. Here we will only cover the boosted gifter sample weight which illustrates blending probability and non-probability samples and interpreting the results. As discussed above, the boosted gifter sample was made up of the general population fieldwork with the NatCen Panel and the Boost fieldwork with the PopulusLive panel. A set of weights were created to account for selection bias of the Populus boost (a non-probability sample) and non-response of both samples. Because the subsample of gifters from the NatCen panel was too small to be treated as source of population totals for the combined subsample, data from gifters in the NatCen panel and Populus Omnibus surveys were used for this purpose.

Propensity Score Matching (PSM) was used to align the Populus Omnibus sample with the NatCen panel weighted sample profile. The modelling aimed at minimizing both selection and non-response bias in the Populus Omnibus sample. Variables observed in both surveys were tested, and the final PSM model included: region, age and sex grouped, banded income, tenure type, household type and total wealth. The combined and weighted NatCen panel and Populus Omnibus samples were then treated as a source of population totals for the adjustment of the boosted gifter subsample. Calibration weighting was used to adjust the boosted gifter subsample within: region, age and sex, income banded, tenure type, household type, whether the participant has given a single gift of more than £1,000, and whether the participant has given multiple gifts that add up to more than £3,000.

The weighting scheme aimed to remove the selection bias, but while it was effective at reducing observed bias it should not be assumed that all of the bias was removed since it is impossible to account for unobserved differences and assess the accuracy of the estimates since there are no reliable sources of profiling information for gifters. It is therefore simply assumed that the weighting minimised the selection and non-response bias introduced through addition of a non-random sample.

7. Conclusion

This review has looked into the advantages and challenges specific to the use of random versus non-random population sampling strategies, particularly in relation to the acquisition of longitudinal studies data. The available evidence to date makes strong arguments in favour of using random probability samples over non-probability samples. However, in the context of longitudinal studies, even samples selected with random probability to begin with are bound to become unrepresentative as the population changes, and additional top-ups are recommended.

When it comes to combining probability and non-probability samples, few attempts seem to have been made, and where attempts have been made as discussed above, the evidence in their favour lacks strength.

Even when it comes to online recruitment, the general consensus is that the best approach is to focus on pre-recruited, probability panels which contact potential members using established offline probability-based techniques, especially if such panels are to be used for longitudinal analysis purposes.

References:

Assets.publishing.service.gov.uk. (2019). [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/799577/Lifetime_Gifting_-_Reliefs__Exemptions__and_Behaviours_Research.pdf [Accessed 6 Jun. 2019].

Baker, R., Brick, M.J., Bates, N., Battaglia, M., Couper, M.P. & Dever, J.A. (2013) Summary report of the AAPOR Task Force on non-probability sampling (with discussion). *Journal of Survey Statistics and Methodology*, *1*, 90-143. http://dx.doi.org/10.1093/jssam/smt008

Bethlehem, J., & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Cls.ucl.ac.uk. (2019). *CLS | 1958 National Child Development Study*. [online] Available at: https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/ [Accessed 3 Jun. 2019].

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Ebrahim, S. & Davey-Smith, G. (2013). Commentary: should we always be deliberately non-representative? *International Journal of Epidemiology, 42*, 1022-1026. http://dx.doi.org/10.1093/ije/dyt105

Elliott, M. (2009). Combining Data from Probability and Non- Probability Samples Using Pseudo-Weights. *Survey Practice*, [online] 2(6), pp.1-7. Available at: https://www.surveypractice.org/article/2982-combining-data-from-probability-and-non-probability-samples-using-pseudo-weights [Accessed 29 May 2019].

Goldstein, H., Lynn, P., Muniz-Terrera, G. & Hardy, R., O'Muircheartaigh, C., Skinner, C. & Lehtonen, R. (2015). Population sampling in longitudinal surveys debate. *Longitudinal and Life Course Studies, 6, 447 – 475.* http://dx.doi.org/10.14301/llcs.v6i4.345

Göritz, A. S. (2004a). Recruitment for online access panels. *International Journal of Market Research*, *46*, 411–425.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. Survey Methodology, 32(6), 133–142.

Kott, P. (2019). *A Partially Successful Attempt to Integrate a Web-Recruited Cohort into an Address-Based Sample*. [online] Ojs.ub.uni-konstanz.de. Available at: https://ojs.ub.uni-konstanz.de/srm/article/view/7222 [Accessed 29 May 2019].

Natcen.ac.uk. (2019). [online] Available at: http://www.natcen.ac.uk/media/1484228/Developing-the-NatCen-Panel-V2.pdf [Accessed 6 Jun. 2019].

O'Neil, K. M., Penrod, S. D., & Bornstein, B. H. (2003). Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods, Instruments, & Computers*, *35*, 217–226.

Savage M., Devine, F., Cunningham, N., Friedman, S., Laurison, D., Miles, A. … & Taylor, M. (2014) On Social Class, Anno 2014. *Sociology*, *48* (forthcoming). http://dx.doi.org/10.1177/0038038514536635

Schillewaert, N., & Meulemeester, P. (2005). Comparing response distributions of offline and online data collection methods. *International Journal of Market Research*, *47*, 163–178.

The-sra.org.uk. (2019). [online] Available at: http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-06-summer-2018.pdf [Accessed 6 Jun. 2019].

Wadsworth, M.E.J. (1991). *The imprint of time: Childhood, History and Adult Life.* Oxford University Press.