KANTAR PUBLIC=

Five methods of within-household sampling: does it matter which one we use?

An empirical study using Community Life Survey data

January 2019

Joel Williams

Contents

Abstract

This paper describes a sensitivity test carried out to assess how Community Life Survey estimates might change if we used a different method of selecting individual(s) to survey within sampled households. At present, *all* resident individuals aged 16+ may participate but alternatives include (i) sampling one at random (the gold standard but very difficult to implement properly), (ii) allowing *any* individual to take part, (iii) allowing any *two* individuals to take part, and (iv) selecting the youngest individual in the household.

The results suggest that the current method – allowing all individuals to participate – yields estimates that are very close to a simulated 'random-one' gold standard while being easier to implement and costing less. However, in general, the choice of within-household selection mechanism is not critical to the Community Life Survey estimates. With a small number of exceptions, effect sizes are small and inference would be largely unaffected whichever method was adopted.

1. Background to this study

The Community Life Survey utilises the ABOS method – address-based online surveying – to collect official statistics on levels of community cohesion and engagement. The core ABOS design is a simple one: a stratified random sample of addresses is drawn from the Royal Mail's postcode address file and an invitation letter is sent to 'the residents' of each one, containing four usernames and passwords plus the URL of the survey website. Every individual aged 16+ who is resident in the sampled household can log on using this information and complete the survey as they might any other online survey.[1] Once the questionnaire is complete, the specific username and password cannot be used again, ensuring data confidentiality from others with access to this information.

Paper questionnaires are used as a supplementary mode of data collection for those who are offline or otherwise unwilling to complete the questionnaire online. Two reminders are sent to each address and a £10 shopping voucher is given to all who take part.[2] The (design-weighted) estimated response rate for the 2016-17 survey was 23%. If we count only the online completed questionnaires, the estimated response rate was 17%.[3]

As noted above, *all* individuals aged 16+ and resident in sampled households are invited to take part in the Community Life Survey. This design was adopted due to strong evidence that within-household random sampling is very difficult to implement accurately if left to the residents themselves.[4]

However, the 'all-individuals' design has its own weaknesses:

(i) the first respondent's experience of the survey may influence the responding behaviour (whether and how) of other adults in the household; and

(ii) because each completed questionnaire is incentivised with a £10 shopping voucher, some individuals may be tempted to complete the questionnaire multiple times, adopting different identities each time. Efforts are made at the processing stage to identify and exclude these cases but these are likely to be at least partially fallible.

The 'all-individuals' design is not the only alternative to (attempted) random sampling of a single eligible individual per sampled household. Other methods include (i) asking *any one* eligible individual to complete the questionnaire, (ii) asking *any two* eligible individuals to complete the questionnaire, and (iii) asking the *youngest* eligible individual to complete the questionnaire.

If we select the 'any-one' design we make the assumption that it does not matter which eligible individual takes part but that within-household conditioning is a risk to be avoided.

[1] Four logins are provided in the letter but more can be supplied on request for households with 5+ eligible individuals.

[2] See http://the-sra.org.uk/wp-content/uploads/social-research-practice-journal-issue-03-winter-2017.pdf for a thorough introduction to ABOS methods in general.

[3] See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/651589/Community_Life__Online_and_Paper_Survey_Technical_Report_-_2016-17_v4_FINAL.pdf for details.

[4] See www.gov.uk/government/uploads/system/uploads/attachment_data/file/466925/The_Community_Life_Survey_Investigating_the_Feasibility_of_Sampling_All_Adults_in_the_Household_FINAL.pdf and www.natcen.ac.uk/media/541183/ess-mixed-mode-natcen-report.pdf for some examples.

If instead we select the 'any-two' design we are asserting that self-selection within household *is* a risk and a greater one than within-household conditioning. However, rather than inviting *all* individuals to take part, we limit the risk of incentivising false completions by setting a maximum of two vouchers per household. In short, 'any-two' is a hybrid design, allowing non-random selection in households with three or more eligible individuals (c15% of the total in the UK) in exchange for a reduced risk of false completions.

The 'youngest-one' design shares the assumptions of the 'any-one' design that (i) there is no need for *random* sampling within sampled households but that (ii) within-household conditioning is a risk to be avoided. However, it additionally makes the assumption that younger individuals will be less willing to take part in the survey than older individuals and that, consequently, sampling only the youngest individual might improve the overall age profile of the responding sample. This design differs from the others by explicitly ruling out the participation of 'non-youngest' individuals living in sampled households. In all the other designs considered here, every eligible individual has a chance of being selected.

In short, there are at least five within-household sample designs that might be used for the Community Life Survey (or any other ABOS study): 'all-individuals', 'random-one' (if a way can be found to implement it properly), 'any-one', 'any-two' and 'youngest-one'.[5] Other designs might be employed but we consider just these five in this paper.

[5] Two of these designs – 'random-one' and 'youngest-one' – have selection rules that might be broken by sampled households. However, the questionnaire might include items to detect if this has happened, giving research commissioners the option of discarding these cases (or at least treating them differently).

2. Hypothesis motivating this study

An obvious question that follows is: what difference does the within-household sample design make to the data? Our hypothesis is: "it makes very little difference", and we explain why below.

For the Community Life Survey, the sample of households is controlled and random. This limits the range of possible samples of individuals *within* these households, regardless of which design is employed.

For a start, all designs 'sample' the same individual in households with just one eligible individual. Secondly, in a household with two eligible individuals – the most common scenario for most surveys - there are only two possible permutations: individual A or individual B (or only one permutation if the 'any-two' design is adopted). In these households, we should not expect even the most self-selecting design – 'any-one' – to yield a radically different sample from the least self-selecting design ('random-one', properly implemented). Regardless of the individuals' relative probabilities of self-selection under the 'any-one' design, we should expect half the sample – across *n* households of two individuals - to be identical to what a 'random-one' design would deliver.

It is really only in households with three or more eligible individuals that the choice of within-household sampling method may yield substantially different samples. Even here, the impact on the data may be limited. Households of individuals tend to be relatively homogeneous compared to the broader population of individuals. This fact may limit the effect of within-household selection on the survey outcomes even when two methods produce a wholly different sample of individuals within the same sample of households.

In short, our hypothesis is that the systematic effect of each within-household sample design is likely to be very small. In other words, it does not matter greatly which design is adopted, the survey estimates would be approximately equal over the long run. If this hypothesis has empirical support, then the best method is the one that maximises the effective sample size per unit of cost.

3. An empirical test using Community Life Survey data

The Community Life Survey– with its 'all-individuals' design - provides an opportunity for an empirical test of the hypothesis set out in section 2 (that the choice of within-household sampling mechanism has no substantial systematic effect on the survey estimates).

As part of the survey, household composition data is collected, providing gender and age information for every individual in the household. Furthermore, for the online questionnaires, we have precise data about the time of data collection. This allows us to reasonably identify who would have been sampled under each of the five designs:

- 'all-individuals': all respondents under the 'all-individuals' design
- 'random-one': post hoc simple random selection of one individual from the household, with data retained if he/she responded under the 'all-individuals' design
- 'any-one': the first respondent under the 'all-individuals' design[6]
- 'any-two': the first two respondents (or the first respondent if only one) under the 'all-individuals' design
- 'youngest-one': the youngest individual from the household grid, with data retained if he/she responded under the 'all-individuals' design

Each design *d* sub-sample from the 'all-individuals' respondent set ought to be a good proxy for the sample that would have been obtained had the same addresses been selected but design *d* been used instead of the 'all-individuals' design. However, there are two constraining assumptions underpinning this assertion.

The first is that all who would have responded under design *d* would also have responded under the 'all-individuals' design, or at least have had approximately the same response probability. This strikes us as plausible enough, even if not strictly verifiable.

Our second assumption is that the risk that a case is an unidentified false completion and/or subject to within-household response conditioning would be the same under design *d* as it is under the relevant simulation described above.

This seems a reasonable assumption to make about the 'any-one' and 'any-two' simulations but a stronger assumption to make about the 'random-one' and 'youngest-one' simulations. In both of these simulations, the selected case may or may not have been subject to within-household response conditioning but the risk of such conditioning would be close to zero if either design was implemented for real.

Furthermore, the validity of this assumption relies on the efficacy of the data processing algorithm that has been used to identify – and remove - 'false completion' cases. We would expect the case-level risk of false completion to be correlated with the number of completions allowed per household, so higher for a simulation based on unprocessed 'all-individuals' data than it would be for unprocessed 'random-one' or

[6] We make an unverifiable – but we think reasonable - assumption that the early responders under the 'all individuals' design would also self-select under the 'any-one' or 'any-two' designs. For the other designs, no such assumptions are required to identify the relevant samples.

'youngest-one' data. The processing of the 'all-individuals' data needs to reduce the false completion risk to the same level as processed 'random-one' or 'youngest-one' data. This also is not verifiable.

For analysis to proceed, we must either accept these two assumptions or alternatively not fully accept them but treat the resulting error as small and quasi-random in character. While we think it reasonable to treat the proxy samples as reflective of the kinds of sample we would obtain under the five different designs, we acknowledge that others might not.

For this empirical work, we use the 2016-17 Community Life Survey dataset, restricted to the online completions (72% of the total) so that we can categorise each response as the first, second etc. within the household. Naturally, that means ignoring all the paper data obtained through the ABOS method, so– strictly speaking - the conclusions that follow are specific to a (hypothetical) online-only version of the Community Life survey.

The analysis sample sizes (plus other data discussed below) are shown in table 3.1.

Table 3.1: Sample sizes and other data for this empirical test (online respondents only)

Design *d*	Sample size	Design effect due to weighting	Estimated (design-weighted) response rate
All-individuals	7,365	1.60	16.9%
Any-one = first respondent only	4,902	1.83	n/a
Any-two = first two respondents only	6,784	1.62	n/a
Random-one = post hoc random selection from household grid	3,726	1.89	16.2%
Youngest-one = youngest from household grid	4,700	1.81	20.5%

Each subset of cases has been weighted in a consistent manner, following the Community Life Survey protocol where applicable. The weighting process is described in Appendix 1 and the estimated design effects due to weighting are shown in table 3.1. The smallest effective sample size for any design is around 2,000 so we have plenty of statistical power to detect the impact of each design on the survey estimates.[7]

[7] It is worth noting that the 'all-individuals' and 'any-two' designs also have design effects due to household clustering. Consequently, the net design effects (taking both weighting and clustering into account) are similar for all designs (averaging at around 1.8 to 1.9).

4. Analysis of the data

Most survey practitioners would regard the 'random-one' design as the gold standard against which the other designs should be compared. If implemented correctly, it avoids the theoretical weaknesses of (i) sampling bias ('any-one', 'any-two', 'youngest-one') and (ii) response conditioning ('all-individuals' and 'any-two'). However, it is the most difficult design to implement and also leads to higher costs than the 'all-individuals' and 'any-two' designs. Consequently, evidence that other designs produce the same survey estimates should be taken as evidence *against* the random-one design.

To maintain relevance, we use the 28 principal variables used by DCMS when reporting on the Community Life Survey 2016-17 data.[8] [9]

For six of the 28 variables, the range between the highest and lowest design *d* survey estimate is one percentage point or less. This range is less than two percentage points for 18 of the 28 variables and less than three percentage points for 25 of the 28 variables. This alone suggests that the majority of the principal survey estimates are robust to the choice of within-household sampling method. The exceptions – a range greater than three percentage points - are:

1. "How important is it personally for you to be able to influence decisions in your local area?" (Very/quite important): Range = 58.2%-61.9% (3.8%pts); simulated 'random-one' = 58.2%
2. Any civic participation in the last 12 months (derived variable): Range = 43.3%-47.8% (4.5%pts); simulated 'random-one' = 45.6%
3. "Generally speaking, would you like to be more involved in the decisions your council makes that affect your local area?" (Yes): Range = 53.2%-59.0% (5.8%pts); simulated 'random-one' = 54.0%

The 'random-one' survey estimate is at the low end of the range for valuing active involvement in local decisions (items 1 and 3) but is around the mid-point for the behavioural measure, the reported civic participation rate (item 2).

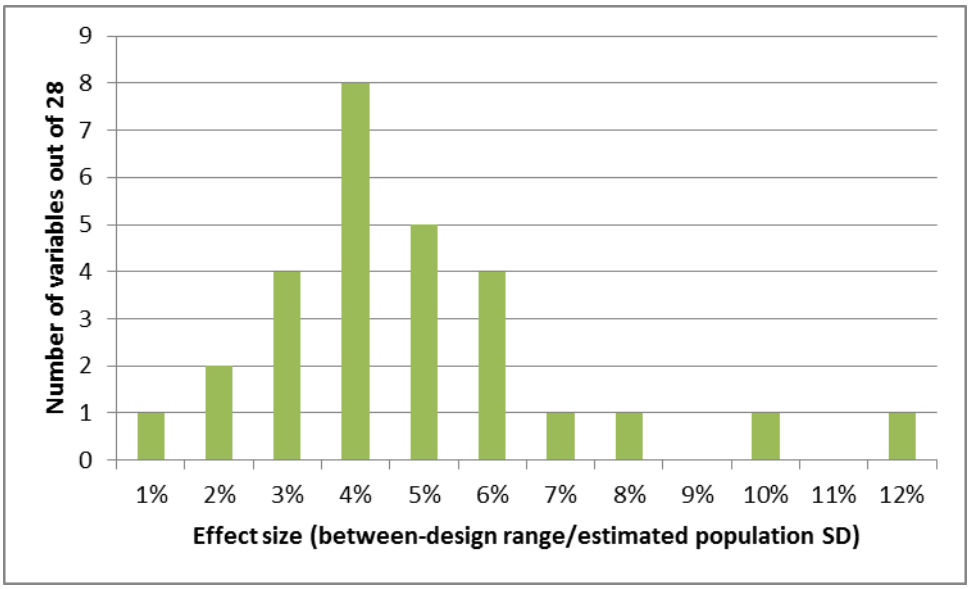
These ranges can also be standardised by dividing each one by the estimated population standard deviation[10]. This may be called the 'effect size'. The modal effect size is around 3-4% of the population standard deviation. It is reasonable to describe this effect size as 'small'. The largest effect size is 12%. A histogram of these effect sizes is shown in Chart 4.1.

[8] For binary variables with 'yes/no' response options, we use 'yes' as the sole response; for categorical variables, we simply use the modal response category.

[9] See this source for an example of how the data is used for official statistics: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/638534/Community_Life_Survey_-_Statistical_Release_2016-17_FINAL_v.2.pdf

[10] This estimate is taken from the 'random-one' design simulation - as the nominated 'gold standard' design – but is virtually identical under all designs.

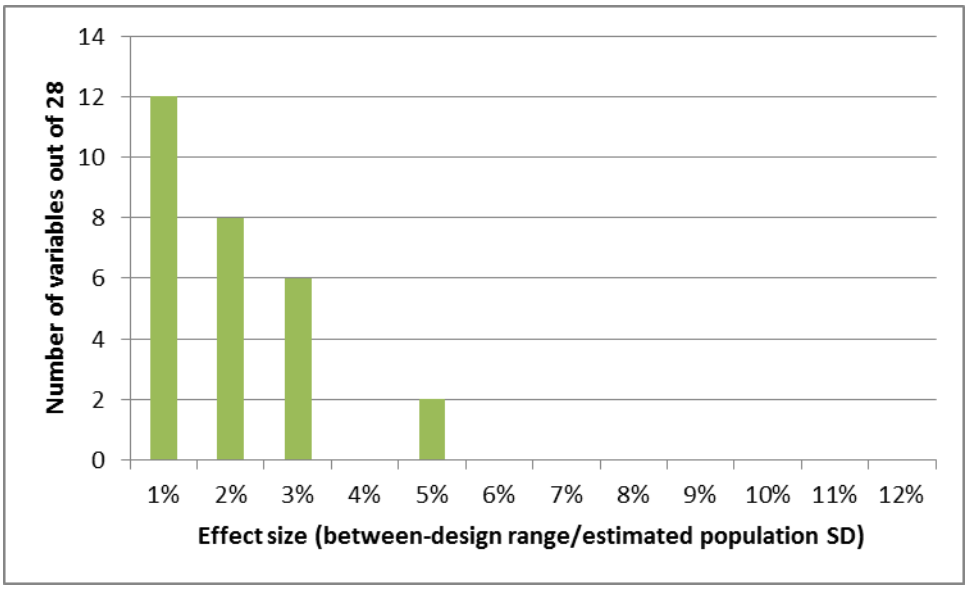
Chart 4.1: Standardised between-design ranges in survey estimates (28 principal variables)



Focusing now on the survey's implemented 'all-individuals' design against the simulated benchmark 'random-one' design, we find that these two designs produce *very* similar results. This suggests (but does not prove, given the caveats around this analysis) that the relative risks of (i) within-household response conditioning, and (ii) unidentified false completions are small. For 24 of the 28 variables, the between-design difference in survey estimates is one percentage point or less. The biggest gap is only 2.3 percentage points: any civic participation in the last 12 months is estimated at 43.3% under the 'all-individuals' design but 45.6% under the simulated 'random-one' design.

Using the 'effect size' metric described above, we find that the maximum effect size is 4-5% of the population standard deviation and that 26 of the 28 effect sizes are 3% or less. In other words, *all* the between-design differences could be classified as 'small'.

Chart 4.2: Standardised difference in survey estimates between the 'all-individuals' and simulated 'random-one' designs (28 principal variables)



We also carried out 'between-design' significance tests for each of the 28 survey estimates. One was significant at the 5% level[11]. However, random sampling theory tells us that we should expect one or two 'significant' differences to be observed if 28 tests are carried out[12] *even if there is no systematic mechanism to produce such differences.* Consequently, we have no strong evidence against the hypothesis that the 'all-individuals' design produces approximately the same survey estimates as the simulated benchmark 'random-one' design.

But how do the other designs perform in comparison? Referring back, we have shown some modest evidence of between-design effects, albeit limited to a small subset of variables. One way to rank the designs is to calculate the average (mean) effect size against the simulated random-one benchmark, across all 28 variables. This produces the following rank order for the within-household sampling designs:

1. 'all-individuals': average effect size = 1.4% (range: 0.1%-4.6%)
2. 'any-two': average effect size = 1.6% (range: 0.1%-4.1%)
3. 'youngest-one': average effect size = 2.1% (range: 0.2%-6.0%)
4. 'any-one': average effect size = 3.0% (range: 0.2%-10.2%)

This suggests that the multiple-response designs ('all-individuals' and 'any-two') produce the most accurate estimates against the benchmark and are the best proxies if the random-one design proves too difficult or expensive to implement. From a broader survey design perspective, these methods might be better than the random-one method because they both cost less and the design effect of the data due to weighting *and* household clustering (averaging between 1.8 and 1.9) is comparable to the design effect due to weighting alone of the 'random-one' design. In other words, the cost per statistically valid case is lower under the 'all-individuals' and 'any-two' designs and lowest of all under the 'all-individuals' design.

Conclusion

Based on this study the 'all-individuals' design adopted for the Community Life Survey appears to be fit for purpose. However, this conclusion rests upon a few untestable - if reasonable – assumptions with respect to our ability to fairly simulate alternative designs. In other words, the evidence we present in support of this conclusion is strong but not the final word on the matter.

While our conclusions with regard to within-household sampling designs are survey-specific, it is reasonable to think that it may carry across in a *general* fashion for other surveys of this type (ABOS). Replicating this study in other contexts would be a useful next step to take.

[11] The same tests found 8 significant differences between the simulated 'random-one' and 'any-one' survey estimates,1 between the 'random-one' and 'any-two' survey estimates, and 2 between the 'random-one' and 'youngest-one' survey estimates. However, these totals are not quite comparable because each set of t-tests has a different average standard error. The smaller the standard error of the difference, the more likely that a fixed sized systematic effect will be found 'significant'.

[12] If a zero effect has a 5% probability of producing a false positive test result, then we can expect 28*5% = 1.4 false positives from 28 tests even if the true effect is zero in every case.

5. Appendix 1: Weighting protocol

Firstly, for simplicity, the cases in each design *d* simulation subset share the same 'address base weight' as the equivalent case in the 'all-individuals' responding sample. This address base weight b_i is equivalent to:

b_i = 1/(address *i* sampling probability * (expected number of responses | address *i* sampled, address-level geodemographic characteristics = $\mathbf{X_i}$, all-individuals design implemented))

The second component in this base weight utilises a predictive model suitable for the 'all-individuals' design. Because this is a minor part of the overall weighting protocol, we retain this model for the other four within-household sampling designs on the assumption that there would be a strong correlation between the fitted values under each design.

This address base weight is multiplied by a within-household sampling weight for the 'random-one' simulation. This within-household sampling weight is equivalent to $1/N_i$ where N_i is the number of eligible individuals in the household. No such weight is applied to any of the other design simulations because either the within-household selection is non-random ('any-one', 'any-two' and 'youngest-one' designs) or there is no selection (the all-individuals design).

Finally, each design simulation subset is *calibrated* to the same vector of population totals using a generalised regression method. This replicates what would have been done had any of the alternative within-household sampling designs been employed for the Community Life survey.

The population totals in the calibration vector cover (i) gender crossed by age group, (ii) region, (iii) housing tenure, (iv) degree-level education crossed by age group, (v) broad ethnic group and (vi) number of residents in the household. This last constraint - on the share of the sample that is resident in single-person households – partly compensates for the non-random selection of single respondents under the 'any-one' and 'youngest-one' designs.

The standard error of the difference in *y* (a survey estimate such as a weighted mean or proportion) between two designs *d* and *e* is estimated as:

$\sqrt{(var(y_d) + var(y_e) – [2*cov(y_d,y_e)*(n^2_{d+e}/(n_d*n_e))])}$

In this case, the common elements of each design simulation subset also have common data, so $cov(y_d,y_e)$ is approximately equivalent to $var(y_{d+e})$. Due to the different weights applied under each design, the sampling covariance of the weighted means/proportions ($cov(y_d,y_e)$) will not be *exactly* equal to the sampling variance of the common mean/proportion ($var(y_{d+e})$) but will be so similar that this simplification is quite acceptable when estimating the standard error.[13]

Table 5.1 shows the bivariate Pearson correlation matrix of every weight. The smallest correlation is .73 and the highest is .99.

Table 5.1 also shows the component $n^2_{d+e}/(n_d*n_e)$ for each pair of design simulation subsets. This is multiplied by 2 times the sampling variance of *y* in the common data ($var(y_{d+e})$) and then subtracted from the sum of the sampling variances of *y* under each design simulation.

[13] Furthermore, any bias in the estimate of the standard error will be towards zero so will work against our hypothesis that the within-household sampling method has little impact on the survey estimates.

Table 5.1: Correlation matrix for each of the five weights, plus other information about each pair of design simulation subsets

DESIGN *d*		All-adults	Any-one	Any-two	Random-one	Youngest-one
All-adults	Pearson Correlation	1.00				
	N	7,365				
	$n^2_{d+e}/(n_d*n_e)$	100%				
Any-one	Pearson Correlation	.92	1.00			
	N	4,902	4,902			
	$n^2_{d+e}/(n_d*n_e)$	67%	100%			
Any-two	Pearson Correlation	.99	.96	1.00		
	N	6,784	4,902	6,784		
	$n^2_{d+e}/(n_d*n_e)$	92%	72%	100%		
Random-one	Pearson Correlation	.81	.89	.85	1.00	
	N	3,726	2,805	3,585	3,726	
	$n^2_{d+e}/(n_d*n_e)$	51%	43%	51%	100%	
Youngest-one	Pearson Correlation	.80	.73	.78	.73	1.00
	N	4,700	3,274	4,422	2,733	4,700
	$n^2_{d+e}/(n_d*n_e)$	64%	47%	61%	43%	100%