Supplementary Material: Small Domain Estimation of Census Coverage – A Case Study in Bayesian Analysis of Complex Survey Data

A. Supplementary Figures

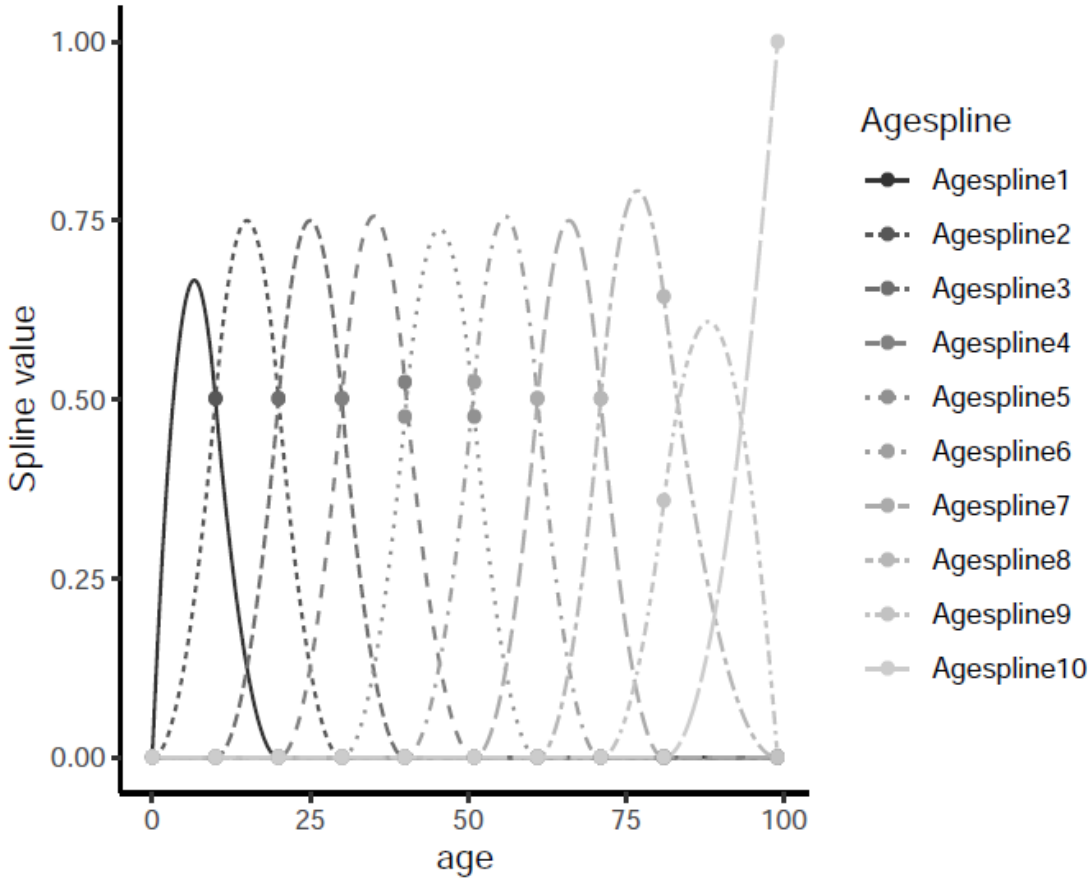

Fig. S1. Spline basis fitted to PES age variable. Dots show specified knots.

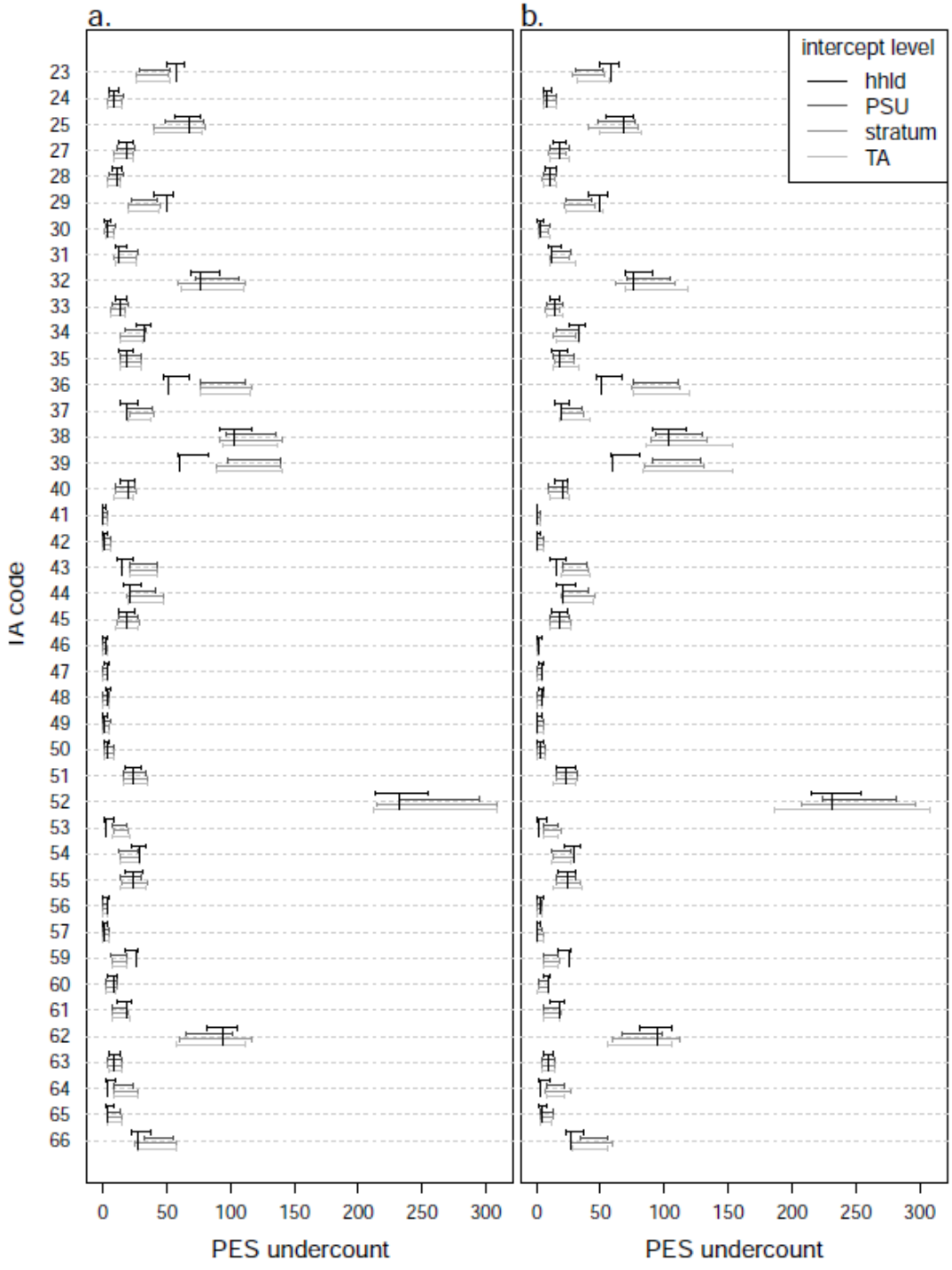

Fig. S2. Posterior predictive checks (PPCs) for model 1 (without household covariates, left) and model 2 with household covariates and their interactions, right). Undercounts by TA; presented for all TAs not represented in Figure 3 in the main text, all demographic categories pooled.

B. Equivalence of Predictions Obtained from Models with Household and PSU Covariates Included either at their Natural Level or at the Individual Level

In this appendix we verify that the model specification in which household and PSU level covariates are included in the model as covariates at the individual level yields the same predicted probabilities as the model given by Equations (2–7) in which covariates are included at their natural level.

Firstly, we provide some additional background on the equivalence of the model specification given by Equations (2–7) and the specification with household and PSU level covariates included at the individual level of the model. In the latter formulation all individuals in a household or PSU are assigned the covariate values for their household or PSU, in addition to their individual level covariate values. We used this formulation of the model for model fitting.

The household level model (4) implies the household effect for the h^{th} household can be written

$$\alpha_h^{\text{hh}} = \mu + \mathbf{X}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \delta_h^{\text{hh}},$$

for $\delta_h^{\text{hh}} \sim N\left(\alpha_{\text{psu}[h]}^{\text{psu}}, \sigma_{\text{hh}}^2\right)$. Similarly, the PSU level model implies the p^{th} PSU effect can be written

$$\alpha_p^{\text{psu}} = \mathbf{X}_p^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \delta_p^{\text{psu}},$$

for $\delta_p^{\text{psu}}, \sim N\left(\alpha_{strat[p]}^{\text{strat}}, \sigma_{\text{psu}}^2\right)$, so

$$\delta_h^{\text{hh}} = \mathbf{X}_{\text{psu}[h]}^{\text{psu}'}\boldsymbol{\beta}^{\text{psu}} + \lambda_h,$$

where $\lambda_h \sim N\left(\delta_{\text{psu}[h]}^{\text{psu}}, \sigma_{\text{hh}}^2\right)$. Consequently,

$$\alpha_h^{\text{hh}} = \mu + \mathbf{X}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{X}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \lambda_h \quad \text{(S1)}$$

and substituting Equation (S1) for α_h^{hh} in the individual level model (3) gives the alternative multilevel model specification

$$[Y_{hj}|p_{under_{hj}}] \overset{\text{indep}}{\sim} \text{Bernoulli}(p_{under_{hj}}), j = 1, \ldots, N_h^{\text{ind}};\ h = 1, \ldots, N_{\text{psu}[h]}^{\text{hh}}, \tag{S2}$$

$$\begin{gathered}\text{logit}\left(p_{under_{hj}}\right) = \mu + \mathbf{X}_{hj}^{\text{ind}\prime}\boldsymbol{\beta} + \mathbf{X}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{X}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \lambda_h, \\ j = 1, \ldots, N_h^{\text{ind}};\ \ h = 1, \ldots, N_{\text{psu}[h]}^{hh},\end{gathered} \tag{S3}$$

$$[\lambda_h|\delta_{\text{psu}[h]}, \sigma_{\text{hh}}^2] \overset{\text{indep}}{\sim} \mathcal{N}(\delta_{\text{psu}[h]}, \sigma_{\text{hh}}^2), h = 1, \ldots N_{\text{psu}[h]}^{\text{hh}}, \tag{S4}$$

$$[\delta_p^{\text{psu}}|\alpha_{strat[p]}^{\text{strat}}, \sigma_{\text{psu}}^2] \overset{\text{indep}}{\sim} \mathcal{N}(\alpha_{strat[p]}^{\text{strat}}, \sigma_{\text{psu}}^2),\ p = 1, \ldots, N_{strat[p]}^{\text{strat}}, \tag{S5}$$

$$[\alpha_s^{\text{strat}}|\mathbf{W}, \boldsymbol{\alpha}^{\text{ta}}, \sigma_{\text{strat}}^2] \overset{\text{indep}}{\sim} \mathcal{N}\left(\mathbf{W}_s\boldsymbol{\alpha}^{\text{ta}}, \sigma_{\text{strat}}^2\right), s = 1, \ldots, N^{\text{strat}}, \tag{S6}$$

$$[\alpha_t^{\text{ta}}|\mathbf{X}_t^{\text{ta}}, \boldsymbol{\beta}^{\text{ta}}, \sigma_{\text{ta}}^2] \overset{\text{indep}}{\sim} t_3(\mathbf{X}_t^{\text{ta}\prime}\boldsymbol{\beta}^{\text{ta}}, \sigma_{\text{ta}}^2), t = 1, \ldots, N^{ta}. \tag{S7}$$

To obtain the predicted under-coverage probability for a hypothetical individual with covariates $\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}$, residing in a household with covariates $\mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}$, located in a PSU with covariates $\mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}$ and in TA, *t*, under the model specification of Equations (S2–S7), we first marginalise over strata to obtain

$$\begin{gathered}\Pr(Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \boldsymbol{\xi}) = \\ \sum_s \Pr\ (Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \text{strat} = s, \boldsymbol{\xi}) \\ \times \Pr(\text{strat} = s|\text{TA} = t, \mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \boldsymbol{\xi}),\end{gathered} \tag{S8}$$

where ξ is the vector of coverage model parameters excluding the household effects.

To obtain the probability $\Pr(Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \text{strat} = s, \boldsymbol{\xi})$ we need to marginalise over the distribution of the household effects, conditional on stratum. Within stratum *s*, it

follows from Equations (S4) and (S5) the household effects distribution, can be obtained by marginalising over PSU effects, to give

$$p(\lambda_h|\alpha_s^{\text{strat}}, \sigma^2_{\text{hh}}, \sigma^2_{\text{psu}}) = \int \mathcal{N}(\lambda_h|\delta^{\text{psu}}_{\text{psu}[h]}, \sigma^2_{\text{hh}}) \times \mathcal{N}(\delta^{\text{psu}}_{\text{psu}[h]}|\alpha_s^{\text{strat}}, \sigma^2_{\text{psu}})\, \mathrm{d}\delta^{\text{psu}}_{\text{psu}[h]}$$
$$= \mathcal{N}(\lambda_h|\alpha_s^{\text{strat}}, \sigma^2_{\text{hh}} + \sigma^2_{\text{psu}}),$$

by the mixture property of the normal distribution (Gelman et al. 2014, 577). Consequently, the stratum-specific predicted under-coverage probabilities in the first term of the summand in Equation (S8) can be obtained as

$$\Pr(Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \text{strat} = s, \boldsymbol{\xi}) =$$
$$\int \text{expit}(\mu + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}^{\text{hh}'}\boldsymbol{\beta}^{\text{hh}} + \mathbf{x}^{\text{psu}'}\boldsymbol{\beta}^{\text{psu}} + \lambda_h)\mathcal{N}(\lambda_h|\alpha_s^{\text{strat}}, \sigma^2_{\text{hh}} + \sigma^2_{\text{psu}})\, \mathrm{d}\lambda_h. \quad \text{(S9)}$$

Note that the first term in the integrand on the right hand side of Equation (S9) does not depend on stratum or TA effects because, under the model given by Equations (S2) –(S7), conditional on the household effects, under-coverage probability does not depend on stratum or TA. The second term in the integrand does not depend on TA because the model implies household effects are independent of TA effects, conditional on stratum effects. That is, TA influences the household effects only through the stratum effects. Similar arguments justify Equation (10) in the main text.

From Equation (S1),

$$\lambda_h = \alpha^{\text{hh}}_{\text{h}} - \left(\mu + \mathbf{X}^{hh'}_h \boldsymbol{\beta}^{\text{hh}} + \mathbf{X}^{\text{psu}'}_{\text{psu}[h]}\boldsymbol{\beta}_{\text{psu}}\right). \quad \text{(S10)}$$

Moreover, due to the symmetry of the normal density in the argument and the mean parameter

$$\mathcal{N}(\lambda_h|\alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2) = \mathcal{N}\left(\alpha_h^{\text{hh}} - (\mu + \mathbf{X}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{X}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}})|\alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2\right)$$
$$= \mathcal{N}\left(\alpha_h^{\text{hh}}|\mu + \mathbf{X}_s^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{X}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2\right). \quad \text{(S11)}$$

Consequently, from Equations (S10) and (S11), (S9) can be written as

$$\Pr(Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \text{strat} = s, \boldsymbol{\xi}) =$$
$$\int \left(\text{expit}(\alpha_h^{\text{hh}} + \mathbf{x}^{\text{xind}\prime}\boldsymbol{\beta})\right) \mathcal{N}(\alpha_h^{\text{hh}}|\mu + \mathbf{x}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{x}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2)\, \text{d}\alpha_h^{\text{hh}},$$

where the straightforward change of variable from λ_h to α_h^{hh} follows from the simple relationship between them, Equation (S10). Therefore,

$$\Pr(Y = 1|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t, \boldsymbol{\theta}) =$$
$$\sum_s \left(\int \left(\text{expit}(\alpha_h^{\text{hh}} + \mathbf{x}^{\text{ind}\prime}\boldsymbol{\beta})\right) \mathcal{N}(\alpha_h^{\text{hh}}|\mu + \mathbf{x}_h^{\text{hh}\prime}\boldsymbol{\beta}^{\text{hh}} + \mathbf{x}_{\text{psu}[h]}^{\text{psu}\prime}\boldsymbol{\beta}^{\text{psu}} + \alpha_s^{\text{strat}}, \sigma_{\text{hh}}^2 + \sigma_{\text{psu}}^2)\, \text{d}\alpha_h^{\text{hh}} \right.$$
$$\left. \times \Pr(\text{strat} = \text{s}|\mathbf{X}^{\text{ind}} = \mathbf{x}^{\text{ind}}, \mathbf{X}^{\text{hh}} = \mathbf{x}^{\text{hh}}, \mathbf{X}^{\text{psu}} = \mathbf{x}^{\text{psu}}, \text{TA} = t) \right)$$

in agreement with Equation (10). That is, the predicted under-coverage probabilities under the model specifications given by Equations (2–7) and Equation (S2) to Equation (S7) are identical. This result is not restricted to normally distributed random effects but holds also for other densities symmetric in the argument and the mean, such as the t-distribution.

C. Ignorability Assumptions

In this appendix, we formalise the ignorability assumptions which justify our modelling approach to the PES data, which includes design variables in the model but does not make use of survey weights as in the design-based approach to the analysis of complex survey data. Our approach follows the framework outlined in Gelman et al. (2014, chap. 8), but, since specific details of the PES design differ from the general treatment given by Gelman et al. (2014), it seems useful to explicitly formalise our assumptions.

Two features of the PES design that differ from the Gelman et al. (2014) set up are that the covariates are observed only for the survey sample, rather than assumed known for the full population and the number of sampling units in the population, that is, households and individuals, is not known. Although households were selected from a household sampling frame this frame does not necessarily give an accurate count of the number of households by PSU at the time of PES fieldwork.

C.1. Notation

Although we have endeavoured to maintain consistency in notation between this appendix and the main text, we have made some slight notational changes in the interests of clarity for the current exposition. These are noted below. We consider a population of size N. We let N^{strat}, N_s^{psu}, N_{sa}^{hh} and N_{sah}^{ind} denote, respectively, the number of strata, the number of primary sampling units within stratum s, for $s = 1, \ldots, N^{\text{strat}}$, the number of households in PSU a in stratum s, for $\alpha \in \{1, \ldots, N_s^{\text{psu}}\}$, $s \in \{1, \ldots, N^{\text{strata}}\}$ and the number of individuals within household h in PSU a, in stratum s. This notation for the number of households in a PSU and the number of individuals in a household differs slightly from that used in the main text by explicitly reflecting the hierarchical structure of strata, PSU and household in the subscript notation. As noted above, the within-PSU household counts, N_{sa}^{hh} and the within household person count N_{sah}^{ind} cannot be assumed known in this application. The PES sample was obtained by an area-based sampling scheme, as described in Section 2. Note that although territorial authorities (TAs) were included in the model, they were not part of the survey design, and for the current discussion can be regarded as part of the auxiliary information included in the model.

In order to formalise the ignorability assumptions we need to consider the joint distribution of inclusion in the observed sample dataset and the survey variables. Consequently, we first establish notation to describe the pattern of inclusion in the sample. Note that inclusion in the observed sample dataset requires inclusion on the sampling frame, selection into the sample and response and is therefore not the same concept as selection into the sample.

Reflecting the design of the PES sample, we introduce inclusion indicators for each stage of inclusion,

within strata. Accordingly, let

$$\mathbf{I}_s^{\text{psu}} = \left\{ I_s^{\text{psu}}, \ldots, I_{sN_s^{\text{psu}}}^{\text{psu}} \right\}'$$

be a vector of indicators for inclusion or otherwise in the sample of each of the N_s^{psu} PSUs in the stratum *s*. Similarly, let

$$\mathbf{I}_{sa}^{\text{hh}} = \left\{ I_{sa1}^{\text{hh}}, \ldots, I_{saN_{sa}^{\text{hh}}}^{\text{hh}} \right\}'$$

denote a vector of household inclusion indicators for the a^{th} PSU in stratum *s* and

$$\mathbf{I}_{sah}^{\text{ind}} = \left\{ I_{sah1}^{\text{ind}}, \ldots, I_{sahN_{sah}^{\text{ind}}}^{\text{ind}} \right\}'$$

a vector of sample inclusion indicators, for individuals in the h^{th} household in the PSU *a* within stratum *s*

There are logical dependencies between the sets of inclusion indicators:

$$I_{sa}^{\text{psu}} = 0 \Rightarrow \mathbf{I}_{sa}^{\text{hh}} = \mathbf{0},$$

$$I_{sah}^{\text{hh}} = 0 \Rightarrow \mathbf{I}_{sah}^{\text{ind}} = \mathbf{0},$$

where bold-faced **0** denotes a vector with each element equal to zero.

It is also useful to define full PSU, household and individual inclusion vectors as

$$\mathbf{I}^{\text{psu}} = \left(\mathbf{I}_1^{\text{psu}}, \ldots, \mathbf{I}_{N^{\text{strata}}}^{\text{psu}} \right)',$$

$$\mathbf{I}^{\text{hh}} = \left\{ \mathbf{I}_{sa}^{\text{hh}}, a = 1, \ldots N_s^{\text{psu}}, s = 1, \ldots, N^{\text{strata}} \right\},$$

$$\mathbf{I}^{\text{ind}} = \left\{ \mathbf{I}_{sah}^{\text{ind}}, s = 1, \ldots, h = 1, \ldots, N_{sa}^{\text{hh}}, a = 1, \ldots, N_s^{\text{psu}}, s = 1, \ldots, N^{\text{strata}} \right\}.$$

$\mathbf{I}^{\text{hh}}$ and $\mathbf{I}^{\text{ind}}$ are both column vectors, structured by PSU, and PSU and household respectively. We have not written them explicitly in that form to avoid cumbersome notation.

Similarly we let

$$\mathbf{N}^{\text{psu}} = \left(N_1^{\text{psu}}, \ldots, N_{N^{\text{strata}}}^{\text{psu}} \right)',$$

$$\mathbf{N}^{\text{hh}} = \left\{ N_{sa}^{\text{hh}}, a = 1, \ldots, N_s^{\text{psu}}, s = 1, \ldots, N^{strata} \right\},$$

$$\mathbf{N}^{\text{ind}} = \left\{ N_{sah}^{\text{ind}}, h = 1, \ldots, N_{sa}^{\text{hh}}, a = 1, \ldots, N_s^{\text{psu}}, s = 1, \ldots, N^{strata} \right\},$$

where $\mathbf{N}^{\text{hh}}$ is a vector of household counts for the whole country, structured by PSU and strata. $\mathbf{N}^{\text{ind}}$ is a vector of counts of individuals, structured by household, PSU and stratum. Neither $\mathbf{N}^{\text{hh}}$ nor $\mathbf{N}^{\text{ind}}$ are fully observed.

Selection of PSUs is typically from a list and under the direct control of the sampler. Therefore, in typical area samples $\mathbf{I}_s^{\text{psu}}$, $s = 1, \ldots, N^{\text{strata}}$ is directly observed and this was the case in PES. The household inclusion indicator is observed and equal to one for all selected households that respond to the survey and is observed and equal to zero for selected households that do not respond. However, in area-based sampling there is a risk that some dwellings may not be represented on the sampling frame. Consequently, there are an unknown number of households that, though part of the target population of households, were erroneously given an effective selection probability of zero. The inclusion indicator for such households is clearly zero but the number of such households is unknown. Similarly, although households in non-selected PSUs clearly have an inclusion indicator equal to zero, the total number of such households will be unknown unless the sampling frame for households is perfect in those PSUs. A useful alternative representation of the household inclusion vector is $\mathbf{I}^{\text{hh}} \underline{\underline{\text{df}}} \left(\tilde{\mathbf{I}}^{\text{hh}}, \mathbf{N}^{\text{hh}} \right)$, where $\tilde{\mathbf{I}}^{\text{hh}}$ represents the vector of inclusion indicators for all selected households, and all known households that were not selected, by PSU and stratum. Given $\mathbf{N}^{\text{hh}}$ we could add the correct number of zero entries by PSU and stratum, to obtain the complete household inclusion vector containing an entry for every household in the target population. Similarly, an alternative representation of the vector of individual inclusion indicators is $\mathbf{I}^{\text{ind}} \underline{\underline{\text{df}}} \left(\tilde{\mathbf{I}}^{\text{ind}}, \mathbf{N}^{\text{ind}} \right)$, where $\tilde{\mathbf{I}}^{\text{ind}}$ is the vector of recorded inclusion indicators for all respondents, and for known non-respondents from responding households, structured by household, PSU

and stratum. Knowledge of $\mathbf{N}^{ind}$ would allow us to construct a completed individual level inclusion vector for the target population by appending the correct number of zero entries, by household, PSU and stratum, to $\tilde{\mathbf{I}}^{ind}$. Of course, if we had complete knowledge of $\mathbf{N}^{ind}$ we would have little need for a census post enumeration survey to assess the population coverage of a census. The fact that $\mathbf{N}^{hh}$ and $\mathbf{N}^{ind}$ are only partially observed means that the household and individual inclusion vectors are only partially observed, in contrast to the set-up assumed by Gelman et al. (2014, chap. 8).

C.2. Population Model: Model for the Joint Distribution of Outcome and Covariates

Let $\mathbf{Y}^{tot} = (Y_1, \ldots, Y_N)^{/}$ denote a vector of length N containing census under-coverage indicators for the full target population. Thus $Y_i = 1$ indicates that the i^{th} individual in the target population was included in the census, $Y_i = 0$, otherwise. In terms of the notation established in C.1 the index i represents the results of a mapping from the stratum, PSU, household and within-household individual indices to the positive integers. We also let $\mathbf{X}_i$ denote a $q \times 1$ vector of covariate values for the i^{th} individual in the population and let $\mathbf{X}^{tot} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ be an $N \times q$ matrix of covariate values for the population. In our analysis $\mathbf{X}$ includes the variables, sex, age, ethnic group, and binary indicators for Maori descent and country of birth (denoted by $\mathbf{X}^{ind}$ in the main text). In this appendix we also consider the identity of individuals' households and household level covariates aggregated from individual covariate values as part of $\mathbf{X}$. That is, the definition of $\mathbf{X}$ used in this appendix, includes $\mathbf{X}^{ind}$, $\mathbf{X}^{hh}$, as defined in the main text, as well as indicators for the household to which an individual belongs. $\mathbf{X}$ is observed through the interview process for PES respondents but is unobserved for non-respondents. We let $\mathbf{Z}^{tot}$ denote all additional available data relevant to the analysis. This includes the relationships between strata, TAs, PSUs, and households as well as covariates defined at each of these levels, except for covariates that are aggregated from the individual level covariates. Thus, in terms of the model given by Equations (1–6) we regard $\mathbf{X}^{psu}$, $\mathbf{X}^{ta}$ and the stratum to TA occurrence matrix $\mathbf{W}$ as part of $\mathbf{Z}^{tot}$. We regard $\mathbf{Z}^{tot}$ as, in principle, known before commencing fieldwork, and is therefore not part of the data collected by PES but rather prior or auxiliary information relevant to the analysis.

Our primary focus is on the relationship of census under-coverage, Y, to the covariates, $\mathbf{X}$, and TA which enters the estimation through the relationship of household to TA (via PSU and stratum). Given model parameters $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$ we model the joint distribution of under-coverage and covariates, conditional on the auxiliary information as

$$p(\mathbf{Y}^{\text{tot}}, \mathbf{X}^{\text{tot}}|\mathbf{Z}^{\text{tot}}, \boldsymbol{\eta}) = p(\mathbf{Y}^{\text{tot}}|\mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta})p(\mathbf{X}^{\text{tot}}|\mathbf{Z}^{\text{tot}}, \boldsymbol{\gamma}),$$

and assume *a priori* independence for parameters of the covariate distribution and the conditional distribution of the census under-coverage given the covariates, that is $p(\boldsymbol{\eta}) = p(\boldsymbol{\theta})p(\boldsymbol{\gamma})$.

Adopting the hierarchical structure of the under-coverage model described in Subsection 3.3 in the main text, we can write the joint conditional distribution of the under-coverage indicators as

$$p(\mathbf{Y}^{\text{tot}}|\mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}) = \prod_s \prod_a \prod_h \prod_j p\left(Y_{sahj,} \mid \mathbf{X}_{sahj}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right). \quad \text{(S12)}$$

where Y_{sahj} and $\mathbf{X}_{sahj}$ refer to coverage indicator and covariates for the j^{th} person in the h^{th} household within PSU a in stratum s.

The conditional independence assumptions embodied by equation (S12) are critically dependent on the content of the covariate and parameter vectors. For example, if $\boldsymbol{\theta}$ does not include household-level intercept terms as in Equation (3), or Equation (S3) and household is associated with tendency to respond to Census then conditional independence over individuals within households will not hold. In this case the final product term in Equation (S12) would need to be replaced by a model for the vector of within-household responses that explicitly modelled the associations between response for individuals in the same household. The conditional independence structure of Equation (S12) is, in fact, stronger than required for the development below but is implied by the model of Equations (2–7), or, equivalently, (S2–S7) and is assumed henceforth.

Our inferential focus is on the parameters of the hierarchical model relating undercoverage to covariates, and we therefore wish to compute the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}^{obs})$, where $\mathbf{D}^{obs}$ denotes the observed data. To define $\mathbf{D}^{\text{obs}}$ it is helpful to partition $\mathbf{Y}^{\text{tot}}$ and $\mathbf{X}^{\text{tot}}$ into components corresponding to

individuals included (denoted by superscript inc) and excluded (denoted by superscript exc) from the survey sample. Consequently, we write $\mathbf{Y}^{tot} = (\mathbf{Y}^{inc}, \mathbf{Y}^{exc})$ and $\mathbf{X}^{tot} = (\mathbf{X}^{inc}, \mathbf{X}^{exc})$, and note that the observable data, including inclusion indicators, are $\mathbf{D}^{obs} = (\mathbf{I}^{psu}, \tilde{\mathbf{I}}^{hh}, \tilde{\mathbf{I}}^{ind}, \mathbf{Y}^{inc}, \mathbf{X}^{inc})$. The auxiliary data $\mathbf{Z}^{tot}$ could also be regarded as part of $\mathbf{D}^{obs}$, but we have kept it as a separate entity, that is part of the conditioning information for our models but not observed through the PES survey process.

C.3. Model for Inclusion Given Outcome and Covariates

We let φ denote the parameters of the conditional distribution of the inclusion indicators given $(\mathbf{Y}^{tot}, \mathbf{X}^{tot}, \mathbf{Z}^{tot})$, $\boldsymbol{\psi} = (\phi, \boldsymbol{\theta}, \boldsymbol{\gamma})$ and model the joint distribution of all inclusion indicators, outcomes and covariates, conditional on the auxiliary information, $\mathbf{Z}^{tot}$, using the decomposition

$$p(\mathbf{I}^{psu},\mathbf{I}^{hh}, \mathbf{I}^{ind}, \mathbf{Y}^{tot}, \mathbf{X}^{tot}|\mathbf{Z}^{tot}, \boldsymbol{\eta}) =$$
$$p(\mathbf{I}^{psu},\mathbf{I}^{hh}, \mathbf{I}^{ind}, \mathbf{Y}^{tot}, \mathbf{X}^{tot}|\mathbf{Z}^{tot}, \phi)\, p(\mathbf{Y}^{tot}|\mathbf{X}^{tot}|\mathbf{Z}^{tot}, \boldsymbol{\theta})\, p(\mathbf{X}^{tot}|\mathbf{Z}^{tot}, \boldsymbol{\gamma}).$$

In view of the relationship between PSUs, households and individuals the conditional distribution of inclusion indicators can be modelled as follows

$$p(\mathbf{I}^{psu},\mathbf{I}^{hh}, \mathbf{I}^{ind}|\, \mathbf{Y}^{tot}, \mathbf{X}^{tot}|\mathbf{Z}^{tot}, \phi) = p(\mathbf{I}^{psu}|\mathbf{Y}^{tot}, \mathbf{X}^{tot}|\mathbf{Z}^{tot}, \phi)\times$$
$$p(\mathbf{I}^{hh}|\mathbf{I}^{psu}, \mathbf{Y}^{tot}, \mathbf{X}^{tot}|\mathbf{Z}^{tot}, \phi)\times$$
$$p(\mathbf{I}^{ind}|\mathbf{I}^{hh}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}\ \phi).$$

We say PSU inclusion is uninformative with respect to $\mathbf{Y}^{tot}$ if

$$p(\mathbf{I}^{psu}|\mathbf{Y}^{tot}, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi) = p(\mathbf{I}^{psu}|\, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi). \tag{S13}$$

Similarly, household inclusion is uninformative with respect to $\mathbf{Y}^{tot}$ if

$$p(\mathbf{I}^{hh}|\mathbf{I}^{psu}, \mathbf{Y}^{tot}, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi) = p(\mathbf{I}^{hh}|\mathbf{I}^{psu}, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi) \tag{S14}$$

and inclusion of individuals is ignorable with respect to $\mathbf{Y}^{tot}$ if

$$p(\mathbf{I}^{ind}|\mathbf{I}^{hh}, \mathbf{Y}^{tot}, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi) = p(\mathbf{I}^{ind}|\mathbf{I}^{hh}, \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \phi). \tag{S15}$$

The uninformative inclusion assumptions in Equations (S13)–(S15) are a version of the "independence assumption" that is commonly invoked in dual systems approaches to population estimation (Chandrasekar and Deming 1949; Brown et al. 2019). Simply put, by assuming uninformative inclusion, as defined above, we are assuming that the probability of inclusion in the PES does not depend on inclusion in the census. Equations (S13)–(S15) state this assumption for each stage of selection and response. As with the conditional independence assumption of Equation (S12) the validity of the uninformative inclusion assumptions depends critically on the conditioning information contained in ($\mathbf{X}^{tot}$, $\mathbf{Z}^{tot}$). For example, since PSUs in urban strata were sampled at higher rate than PSUs in non-urban strata, if census coverage varied between urban and non-urban areas, PSU inclusion would be informative since PSU inclusion would be associated with census coverage. However, within strata, inclusion of PSUs depends only on a measure of size that is related to the proportion of Pacific people in the PSUs, hence conditional on both stratum and the PSU size measure, it seems reasonable to assume PSU inclusion is uninformative. Therefore, stratum indicators and PSU size should be included in the model and we regard these variables as part of the auxiliary information, $\mathbf{Z}^{tot}$. In a multilevel model the natural way to include stratum indicators and PSU size is as covariates in the PSU level model.

An advantage of specifying the uninformative inclusion assumptions in the sequential manner of Equations (S13)–(S15), corresponding to the stages of selection and response, is that it facilitates consideration of the plausibility of the assumptions.

C.4. The Posterior Distribution of Model Parameters Given the Population and Inclusion models

Although our primary interest is in obtaining the posterior distribution for the parameters of the census coverage model, $p(\boldsymbol{\theta}|\mathbf{D}^{obs})$, we first derive the joint posterior distribution for parameters of the coverage, covariate and inclusion models, $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\phi})$. Given $p(\boldsymbol{\psi}|\mathbf{D}^{obs})$, $p(\boldsymbol{\theta}|\mathbf{D}^{obs})$ can be obtained as $p\left(\boldsymbol{\theta} \mid \mathbf{D}^{obs}\right) = \int p\left(\boldsymbol{\psi} \mid \mathbf{D}^{obs}\right) \, \mathrm{d}(\boldsymbol{\gamma}, \boldsymbol{\phi})$.

The likelihood function for the model parameters is obtained by integrating the complete data likelihood over the unobserved components of the complete data. The likelihood is therefore

given by

$$p\left(\mathbf{D}^{\text{obs}} \mid \boldsymbol{\psi}\right) = \int\int\int\int \left(p\left(\mathbf{I}^{\text{psu}}, \mathbf{I}^{\text{hh}}, \mathbf{I}^{\text{ind}} \mid \mathbf{Y}^{\text{tot}}, \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\phi}\right) p\left(\mathbf{Y}^{\text{tot}}, \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right) \right.$$

$$\left. \times p\left(\mathbf{X}^{tot} \mid \mathbf{Z}^{tot}, \boldsymbol{\gamma}\right) \mathrm{d}\,\mathbf{Y}^{\text{exc}}\, \mathrm{d}\,\mathbf{X}^{\text{exc}} \right) \mathrm{d}\,\mathbf{N}^{\text{ind}}\, \mathrm{d}\,\mathbf{N}^{\text{hh}}.$$

If inclusion is uninformative at each stage of selection and response we have

$$p\left(\mathbf{D}^{\text{obs}} \mid \boldsymbol{\psi}\right) = \int\int\int \left(p\left(\mathbf{I}^{\text{psu}}, \mathbf{I}^{\text{hh}}, \mathbf{I}^{\text{ind}} \mid \mathbf{X}^{tot}, \mathbf{Z}^{tot}, \boldsymbol{\phi}\right) p\left(\mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\gamma}\right) \right.$$

$$\left. \times \left(\int p\left(\mathbf{Y}^{\text{tot}} \mid \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\phi}\right) \mathrm{d}\,\mathbf{Y}^{\text{exc}} \right) \right) \mathrm{d}\,\mathbf{X}^{\text{exc}}\, \mathrm{d}\,\mathbf{N}^{\text{ind}}\, \mathrm{d}\,\mathbf{N}^{\text{hh}}$$

$$= \int\int\int p\left(\mathbf{I}^{\text{psu}}, \mathbf{I}^{\text{hh}}, \mathbf{I}^{\text{ind}} \mid \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\phi}\right) p\left(\mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\gamma}\right) p\left(\mathbf{Y}^{\text{inc}} \mid \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right) \mathrm{d}\,\mathbf{X}^{\text{exc}}\, \mathrm{d}\,\mathbf{N}^{\text{ind}}\, \mathrm{d}\,\mathbf{N}^{\text{hh}}$$

$$= p\left(\mathbf{Y}^{\text{inc}} \mid \mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right) \int\int\int p\left(\mathbf{I}^{\text{psu}}, \mathbf{I}^{\text{hh}}, \mathbf{I}^{\text{ind}} \mid \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\phi}\right) p\left(\mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\gamma}\right) \mathrm{d}\,\mathbf{X}^{\text{exc}}\, \mathrm{d}\,\mathbf{N}^{\text{ind}}\, \mathrm{d}\,\mathbf{N}^{\text{hh}}, \qquad \text{(S16)}$$

where the last equality follows from the conditional independence assumptions of our coverage model (S12) which implies that the coverage indicator for an individual can depend on that individual's covariates but not on the covariates of others. The latter assumption, which is weaker than conditional independence, would be sufficient to justify the last equality in (S16).

Thus, assuming uninformative inclusion at each stage of selection and response, in addition to the standard modelling assumption of conditional independence over individuals, leads to a likelihood function that separates into a component that involves only the parameters of the coverage model, $\boldsymbol{\theta}$, $p(\mathbf{Y}^{\text{inc}}|\mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta})$, and a component that involves the remaining parameter blocks, but not $\boldsymbol{\theta}$. Therefore, under the assumption of *a priori* independence for the parameters of the coverage model, covariate distribution and inclusion model, the joint posterior for the model parameters is

$$p\left(\boldsymbol{\psi} \mid \mathbf{D}^{\text{obs}}\right) \propto \left(p\left(\boldsymbol{\phi}\right) p\left(\boldsymbol{\gamma}\right) \int\int\int p\left(\mathbf{I}^{\text{psu}}, \mathbf{I}^{\text{hh}}, \mathbf{I}^{\text{ind}} \mid \mathbf{X}^{\text{tot}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\phi}\right) p\left(\mathbf{X}^{\text{tot}} \mid \mathbf{Z}^{\text{tot}}, \boldsymbol{\gamma}\right) \mathrm{d}\,\mathbf{X}^{\text{exc}}\, \mathrm{d}\,\mathbf{N}^{\text{ind}}\, \mathrm{d}\,\mathbf{N}^{\text{hh}} \right)$$

$$\times p(\boldsymbol{\theta})\, p\left(\mathbf{Y}^{\text{inc}} \mid \mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right). \tag{S17}$$

That is, the posterior for the coverage model parameters is

$$p(\boldsymbol{\theta}|\mathbf{D}^{\text{obs}}) \propto p(\boldsymbol{\theta})p(\mathbf{Y}^{\text{inc}}|\mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta})$$

and, since $p(\mathbf{Y}^{\text{inc}}|\mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta})$ does not involve the inclusion indicators or the parameters of the inclusion or covariate models, the inclusion model can be ignored in the analysis. Moreover, the conditional independence assumption in Equation (S12) implies that $p(\mathbf{Y}^{\text{inc}}|\mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta})$ has the same form as Equation (S12) but with the products restricted to PSUs, households and individuals included in the survey, that is

$$p\left(\mathbf{Y}^{\text{inc}} \mid \mathbf{X}^{\text{inc}}, \mathbf{Z}^{\text{inc}}, \boldsymbol{\theta}\right) = \prod_{s} \prod_{a:I_{sa}^{\text{psu}}=1} \prod_{h:I_{sah}^{hh}=1} \prod_{j:I_{sahj}^{\text{ind}}=1} p\left(\boldsymbol{Y}_{sahj} \mid \mathbf{X}_{sahj}, \mathbf{Z}^{\text{tot}}, \boldsymbol{\theta}\right).$$

If inclusion is informative at any of the sampling stages or there is dependence between the parameters of the inclusion and population models, then the inclusion model does need to be specified explicitly and included in the model fitting. Pfeffermann et al. (2006) consider a case where the inclusion model depends on the random intercepts of a two level population model. This renders the inclusion of the second level units informative and Pfeffermann et al. (2006) develop a conditional likelihood that explicitly incorporates inclusion indicators for each stage of inclusion and response.

Taken together, the assumptions of *a priori* independence between the parameters of the inclusion and population models, and uninformative inclusion defined by Equation (S13) to Equation (S15) imply that inclusion in the PES is "strongly ignorable" (Gelman et al. 2014, 203). This is, in fact, a stronger assumption than strictly needed to justify ignoring the inclusion indicators in the analysis. A weaker set of assumptions that allowed the joint distribution of inclusion indicators to depend on the observed outcomes $\mathbf{Y}^{\text{inc}}$ but not the unobserved outcomes, $\mathbf{Y}^{\text{exc}}$ would suffice. However, the stronger form of the uninformative inclusion assumptions

given in Equation (S13) to Eqaution (S15) is more readily interpretable and, as noted above, provides a link to the usual assumptions of the dual systems estimation literature.

We note that from Equation (S17) the posterior for the parameters of the covariate distribution does not follow straightforwardly from the observed covariate data, reflecting the impact of survey design and non-response on the observed covariate distribution of survey respondents.

D. Additional Information on Model Covariates

D.1. Hard-to-find Variable

Although presented in the main text as a household-level covariate, This binary variable is calculated at the meshblock level.

First, for meshblocks that are present in the PES sample, a dwelling undercoverage coefficient is calculated using the ratio of number of dwellings not in the PES complete list over number of dwellings in the census frame. It is then converted into a binary variable using a threshold of 0.2, chosen from the shape of the variable's distribution. A meshblock with a value of 1 for that variable means that the dwellings in this meshblock are more likely to be missed by PES (hard to find).

A logistic regression is then performed to predict this binary dwelling undercoverage variable from meshblock variables available in the whole country. After model selection, the selected variables are the following:

- Urban/rural indicator
- Proportion of addresses in the address register with an associated dwelling
- Proportion of residential dwellings

- Proportion of non-private dwellings
- Proportion of private dwellings at non-private dwellings
- Proportion of multi-dwelling addresses

The predictions from the logistic regression constitute the binary Hard-to-find variable.

References

Brown, J.J., C. Sexton, O. Abbott, and P.A. Smith. 2019. The framework for estimating coverage in the 2011 census of England and Wales: Combining dual-system estimation with ratio estimation. *Statistical Journal of the International Association of Official Statistics* 35: 481–499. DOI: https://doi.org/10.3233/SJI-180426.

Chandrasekar, C., and W.E. Deming. 1949. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* 44: 101–115. DOI: https://doi.org/10.1080/01621459.1949.10483294.

Gelman, A., J. Carlin, H. Stern, D. Dunson, D. Vehtari, and A. Rubin. 2014. *Bayesian Data Analysis.* Boca Raton, FL.: CRC Press.

Pfeffermann, D., F.A.D.S. Moura, and P.L.D.N. Silva. 2006. Multilevel modelling under informative sampling. *Biometrika* 93: 943–959. DOI: https://doi.org/10.1093/biomet/93.4.943.