Design of the Millennium Villages Project Sampling Plan: a simulation study for a multi-module survey

Shira Mitchell, Rebecca Ross, Susanna Makela, Elizabeth A. Stuart, Avi Feller, Alan M. Zaslavsky, Andrew Gelman

arXiv:1507.02739v1 [stat.AP] 9 Jul 2015

Abstract

The Millennium Villages Project (MVP) is a ten-year integrated rural development project implemented in ten sub-Saharan African sites. At its conclusion we will conduct an evaluation of its causal effect on a variety of development outcomes, measured via household surveys in treatment and comparison areas. Outcomes are measured by six survey modules, with sample sizes for each demographic group determined by budget, logistics, and the group's vulnerability. We design a sampling plan that aims to reduce effort for survey enumerators and maximize precision for all outcomes. We propose two-stage sampling designs, sampling households at the first stage, followed by a second stage sample that differs across demographic groups. Two-stage designs are usually constructed by simple random sampling (SRS) of households and proportional within-household sampling, or probability proportional to size sampling (PPS) of households with fixed sampling within each. No measure of household size is proportional for all demographic groups, putting PPS schemes at a disadvantage. The SRS schemes have the disadvantage that multiple individuals sampled per household decreases efficiency due to intra-household correlation. We conduct a simulation study (using both design- and model-based survey inference) to understand these tradeoffs and recommend a sampling plan for the Millennium Villages Project. Similar design issues arise in other studies with surveys that target different demographic groups.

1 Background

The Millennium Villages Project (MVP) is an economic development project that targets rural populations across ten countries in sub-Saharan Africa, implementing a multi-sector package of interventions at a village level (Sachs and McArthur, 2005; Sanchez et al., 2007). See Mitchell et al. (2015a) for background on the project, study site selection, outcomes of interest, and a comprehensive description of the plan to evaluate its effectiveness. Mitchell et al. (2015b) describe our plan for causal inference about the MVP's effect on a variety of development outcomes measured in different demographic groups. These outcomes will be measured via survey modules administered in both treatment and comparison villages.

A design analysis described in Mitchell et al. (2015b) was used to recommend the number of control villages and magnitude of sampling in each. Next, we must determine how to select households and individuals within households. We propose a two-stage sample: households will be sampled in stage I, followed by individuals within households in stage II (Lohr, 2010; Särndal et al., 1992). In the first stage, we must decide between simple random sampling and probability

proportional to size sampling of households. Because the project operates at the village level, a sampling plan that efficiently estimates outcome means per village is an efficient sampling plan for the overall causal evaluation. In this paper we conduct a simulation study to decide on a sampling plan for estimating finite population village means.

We aim to minimize the design effect, the ratio between the actual and effective sample sizes. One factor in determining the efficiency of a sampling design is the intraclass correlation, i.e. the correlation among individuals within a household. If more than one individual is sampled per household, the intraclass correlation increases the design effect, reducing the effective sample size relative to the actual sample size.

Another factor in the efficiency of a sampling design is the distribution of individuals' sampling probabilities. Sampling probabilities can be optimized for a specific outcome, e.g. by sampling with probability approximately proportional to the outcome (Särndal et al., 1992, p.88). However, with many outcomes of interest, such tailored optimization is difficult or impossible. Therefore, a *self-weighted* sample design is preferred, such that all individuals are sampled with equal probability (Kish, 1992; Lohr, 2010, p.287). Such samples are representative without weighting adjustments, and unbiased point estimates can be obtained from standard statistical procedures.

Given a fixed precision, we aim to minimize time and resources for the survey enumerator teams. This includes minimizing the numbers of people surveyed (i.e. the actual sample size), but also considering the number of households visited, and the effort required to prepare a sampling frame. To conduct the first stage of sampling, a scheme that samples households with equal probability only requires a list of all households with GPS coordinates identifying their locations. However, a scheme which samples households with probability proportional to size requires some measure of household size (e.g. the total number of household members). This additional piece of information requires more effort for enumerators, especially for larger villages with many households. After either method of first stage sampling, we will conduct a demographic census in the sampled households to create the sampling frame for the second stage.

In this paper we conduct a simulation study to understand the tradeoffs between simple random sampling and probability proportional to size sampling of households in the context of the MVP evaluation. Additionally, our simulations explore design-based versus model-based inference, a dichotomy which has implications for our the analysis of our outcome data.

2 Outcomes and survey modules

The Millennium Villages Project (MVP) defines 51 outcomes of interest, including measures of poverty alleviation, agriculture, education, gender equality, health, environmental sustainability, and infrastructure (Mitchell et al., 2015a). These outcomes are measured in six different survey modules, whose content is discussed in Section **??** of Mitchell et al. (2015a). These modules include:

- a **household survey**, administered to all household heads (or other knowledgeable household

members) within the sampled households;

- a sex-specific **adult survey**, administered to men and women of reproductive age (15 to 49 years) within the sampled households;
- within the adult-female survey, a **birth history** section, administered to women of reproductive age (15 to 49 years) both in the sampled households and in additional sampled households to reach sample size sufficient for estimating child mortality;
- a **nutrition survey**, administered to men and women age 15 to 49 years in sampled households;
- **blood (malaria and anemia)** testing, administered to four age-sex groups in sampled households: children age 6 to 59 months, school-aged children (5 to 14 years old), men age 15 to 49 years, women age 15 to 49 years; and
- **anthropometry** measurements, administered among children age 6 to 59 months in sampled households.

For each module and age-sex group combination, the project has budgeted a target sample size based on a combination of budget, logistics, and relative importance of different vulnerable populations and intervention beneficiaries.

3 Sampling plans considered

For the purpose of our simulation study, we consider all survey modules except for birth history and the nutrition survey. Our sampling will be performed in two phases. First, we will sample households using either simple random sampling (SRS, without replacement) or probability proportional to size sampling (PPS, with replacement), with household size defined as $\mathsf{N}_{\mathfrak{h},\text{total}}$, the number of household members under 50 years old in household $\mathfrak{h}$. Let $\mathfrak{s}_\text{I}$ be the set of (unique) sampled households. In the PPS scheme, we use $\mathfrak{r}_\text{I}$ to denote the set of sampled households with repeats. Let $\mathfrak{n}_\text{I} = |\mathfrak{s}_\text{I}|$ be the number of households sampled without replacement in the SRS scheme and let $\mathfrak{m}_\text{I} = |\mathfrak{r}_\text{I}|$ be the number of households sampled with replacement in the PPS scheme. We let $\mathfrak{n}_\text{I} = \mathfrak{m}_\text{I} = 300$ based on the project's previous survey rounds and budget for the final survey round.

To describe the within-household sampling plans for each survey module, we use the following notation. Let $\mathsf{N}_\mathfrak{h}$ be the total number of people in household $\mathfrak{h}$ that are in the target age-sex group for a particular module. Let $\mathfrak{n}_\mathfrak{h} \leqslant \mathsf{N}_\mathfrak{h}$ be the number of people in household $\mathfrak{h}$ that we sample and survey. For example, if considering the anthropometry module, then $\mathsf{N}_\mathfrak{h}$ is the number of children under five years of age in household $\mathfrak{h}$ and $\mathfrak{n}_\mathfrak{h}$ is the number of those sampled for the anthropometry module. Let $\mathsf{N} = \sum_\mathfrak{h} \mathsf{N}_\mathfrak{h}$ be the total number of people in the sampling frame (an MV1 or a control village) that are in the module's target age-sex group.

We now outline the within-household sampling plans considered in our simulation study.

Adult, anthropometry, and blood modules

For each module and age-sex group combination, the project has budgeted a target sample size, n_{target}:

- **adult survey** - $n_{target} = 400$ men and $n_{target} = 400$ women of reproductive age (15 to 49 years);
- **blood (malaria and anemia)** - $n_{target} = 300$ children age 6 to 59 months, $n_{target} = 100$ school-aged children (5 to 14 years old), $n_{target} = 100$ men age 15 to 49 years, $n_{target} = 100$ women age 15 to 49 years; and
- **anthropometry** - $n_{target} = 300$ children age 6 to 59 months.

The second stage sampling schemes we consider in this simulation study are, for a given module and age-sex group:

- For **SRS** sampling of households - combine all $N_{s_I} = \sum_{h \in s_I} N_h$ people in the sampled households in the target age-sex group. If $N_{s_I} \leqslant n_{target}$, then survey all. Otherwise, we consider two options:
  - **stratify** by sampling n_h individuals from each household, where n_h is proportional (up to rounding) to N_h, and the constant of proportionality is determined by the total in the sampled households,

$$n_h = \text{round}\left(N_h * \frac{n_{target}}{N_{s_I}}\right); \text{ or}$$

  - take an equal-probability **systematic sample** of n_{target} people. We order the households randomly, and people (in the module's target age-sex group) within households randomly, so that the people within a household are listed consecutively. We then take a sample using the fractional interval method described in Särndal et al. (1992, p.77) and Appendix A. This procedure enables us to control sample sizes and spread the sample across households such that the sample size in a household is always either the ceiling or the floor of the expected sample size in that household under simple random sampling (see Appendix A). Conceptually, this is similar to stratifying on household, except that there is dependence of the samples between strata (i.e. households).
- For **PPS** sampling of households - if $n_{target} \geqslant m_I$, sample a fixed number of people, $n_h = 1$, per household (regardless of household size) if available.[a] If $n_{target} < m_I$, take a simple random sample of n_{target} households from r_I to obtain a smaller PPS sample of households. Then sample $n_h = 1$ per household if available.

[a]It is possible that a household is sampled without any members of the target age-sex group. Therefore, if

Household survey

For both the SRS and PPS schemes, the household survey module is administered to the head of household in each sampled household.

3.1 Simulated data

For each survey module, we simulate one outcome measured by that module. For the household survey we use the total household consumption; for the adult male survey we use the number of days after illness began when the man first sought advice or treatment; for the adult female survey we use the number of times a woman received antenatal care during her most recent pregnancy; for malaria and anemia testing we use hemoglobin blood concentration; for the anthropometry module we use the weight for age z-score. In generating simulated data, we make the simplifying assumption that all individuals in a target age-sex group have non-missing outcomes. For example, we generate antenatal care outcomes for all women of reproductive age.

To generate data, we use the multilevel model

$$\begin{aligned} y_i &\sim \text{Normal}(\alpha_{h[i]}, \sigma_y) \text{ for individuals } i \\ \alpha_h &\sim \text{Normal}(\mu + \beta_1 N_{h,\text{total}} + \beta_2 N_{h,\text{total}}^2,\ \sigma_\alpha) \text{ for households } h, \end{aligned} \tag{1}$$

For total household consumption we use a model analogous to model 1:[b]

$$t_h \sim \text{Normal}(\mu + \beta_1 N_{h,\text{total}} + \beta_2 N_{h,\text{total}}^2,\ N_{h,\text{total}}\sigma_t) \text{ for households } h. \tag{2}$$

We also use a model for the log total consumption (which in our data is more Normally distributed than the total consumption),

$$\log(t_h) \sim \text{Normal}(\mu + \beta_1 N_{h,\text{total}} + \beta_2 N_{h,\text{total}}^2,\ \sigma_t) \text{ for households } h. \tag{3}$$

We use the demographic information from the census and the multilevel model with estimated parameter values (from the survey data) to generate simulated populations. If when models 1, 2 or 3 are fit to past survey data, the 50% posterior interval of β_1 or β_2 contains 0, we set the parameter to 0 when simulating populations. This prevents us from using very noisy estimates of coefficients. Within each simulated population, we randomly sample according to the sampling plans described above, and estimate the finite population mean using either model-based or design-based inference.

$n_h = 1$, then the PPS scheme will result in a smaller sample size than the SRS scheme. Additionally, for the adult module (where $n_{\text{target}} = 400$), if $n_h = 1$ then the PPS scheme will at most sample only 300 adults, one per sampled household.

[b] Model 2 can be motivated by assuming that model 1 holds for individual-level consumption (this would assume that within a household consumption is identically distributed, not taking into account age-sex differences). This model implies that $t_h = \sum_{i|h[i]=h} y_i \sim \text{Normal}(N_{h,\text{total}}\alpha_h, N_{h,\text{total}}\sigma_y)$ and that the marginal variance of t_h is $N_{h,\text{total}}^2(\sigma_\alpha^2 + \sigma_y^2)$.

4 Bayesian model-based inference

To generalize from the data to the population, both design-based and model-based inference must take into account how the data are collected. Let $y = (y_1, ..., y_N)$ denote data for the population of interest, and $I = (I_1, ..., I_N)$ indicators of the observation of y, where $I_i = 1$ if y_i is sampled,[c] and $I_i = 0$ if y_i is missing. Let 'obs' $= \{i : I_i = 1\}$ and 'mis' $= \{i : I_i = 0\}$. Thus, the information available is y_{obs}, I, and the likelihood is $p(y_{obs}, I|x, \theta, \phi) = \int p(y|x, \theta)p(I|x, y, \phi)dy_{mis}$, where x are observed covariates. Bayesian inference computes the posterior distribution $p(\theta, \phi|x, y_{obs}, I)$ (superpopulation inference) and $p(y_{mis}|x, y_{obs}, I, \theta, \phi)$ (finite population inference). Under the *ignorability* condition, these inferences can be simplified to $p(\theta|x, y_{obs})$ and $p(y_{mis}|x, y_{obs}, \theta)$. Ignorability is satisfied if both the *missing at random* and *distinct parameters* conditions are satisfied (Gelman et al., 2014, p.202, 206-211). Missing at random requires that the missingness be independent of the missing values conditional on observed variables and a parameter ϕ: $p(I|x, y, \phi) = p(I|x, y_{obs}, \phi)$. The distinct parameters condition requires that the parameters of the missingness mechanism (ϕ) be independent of the parameters of the data generating process (θ), conditional on covariates: $p(\phi|x, \theta) = p(\phi|x)$.

We include design variables such that the data collection mechanism is ignorable with respect to this model. For example, in our SRS-stratified sampling plan, the data collection mechanism is:

$$p(I|x, y, \phi) = 1/\left[\sum_{\substack{s_I \subseteq \{1,...,N_I\} \\ |s_I| = n_I}} \prod_{h \in s_I} \binom{N_h}{n_h}\right] \text{ where } n_h = \text{round}\left(N_h \frac{n}{N_{s_I}}\right)$$

if $\exists s_I \subseteq \{1, ..., N_I\}$ s.t. $|s_I| = n_I$ and $\sum_{i:h[i]=h} I_i = \text{round}\left(N_h \frac{n}{N_{s_I}}\right)$ for all $h \in s_I$. Otherwise, the probability of missingness pattern I is zero.

Thus, we include as design variables the household identifiers and the N_h (e.g. the number of women per household, if the survey module targets women). Similar computations show that under the SRS-systematic sampling scheme these variables are also sufficient to satisfy missing at random. For the PPS scheme, we also will need the measure of household size used to select the households (e.g. the total number of household members under 50) (Gelman et al., 2014, p.211). For simplicity, we fit the same ignorable model for both the SRS and PPS schemes. For the anthropometry, blood, and adult survey modules we fit

$$\begin{aligned} y_i &\sim \text{Normal}(\alpha_{h[i]}, \sigma_y) \text{ for individuals } i \\ \alpha_h &\sim \text{Normal}(\mu + \beta_1 N_{h,\text{total}} + \beta_2 N_{h,\text{total}}^2 + \beta_3 N_h + \beta_4 N_h^2,\ \sigma_\alpha) \text{ for households } h, \end{aligned} \tag{4}$$

For the household survey, we fit models 2 and 3.

Our parameter of interest is the finite population mean $\bar{Y} = \frac{1}{N}\sum_{h=1}^{N_I} N_h \bar{y}_h$, where $\bar{y}_h =$

[c] We assume that all units that are sampled are observed.

$\frac{n_h}{N_h}\overline{y}_{h,obs} + \frac{N_h - n_h}{N_h}\overline{y}_{h,mis}$ (Gelman et al., 2014, p.205). We obtain posterior simulations of $\overline{Y}$ as follows: if household h is sampled, we use a simulation of α_h to generate $N_h - n_h$ simulated y_i's. If household h is not sampled, we use simulations of μ and σ_α to simulate a new α_h, then generate N_h simulated y_i's.

5 Frequentist design-based inference

We use the `survey` package to compute design-based estimates and variances (Lumley, 2004). Though we perform our SRS schemes without replacement, we compute all variances without finite population corrections, using the Horvitz Thompson (Hajek) ratio estimator and its with-replacement variance (Lohr, 2010, p.247).

Our SRS schemes are two-phase rather than two-stage designs, since the sampling within a household depends on which households were sampled in the first stage (Särndal et al., 1992, p.134-135). This dependence is reflected in the design weights we compute, see below. For the SRS-systematic sampling scheme, the independence assumption of two-stage sampling is also violated, with the sampling in each household dependent on the sampling in other households. Our design-based analysis approximates these two-phase designs with a two-stage analysis. In contrast, in model-based inference the details of the design do not matter in the analysis once we include design variables in our model (Gelman et al., 2014, p.202, 206-211).

5.1 Design weights

For the SRS-systematic design, the inclusion probabilities are:

$$\begin{aligned}
\pi_{hi} &\equiv P[\text{person } i \text{ in household } h \text{ is sampled}] \\
&= P(h \in s_I)P(i \in s_h | h \in s_I) \\
&= \frac{n_I}{N_I} \sum_{s_I | h \in s_I} P(i \in s_h | h \in s_I, s_I) * P(s_I | h \in s_I) \\
&= \frac{n_I}{N_I} \sum_{s_I | h \in s_I} \underbrace{\min\left(\frac{n}{N_{s_I}}, 1\right)}_{(*)} * \frac{1}{\binom{N_I - 1}{n_I - 1}}
\end{aligned}$$

For SRS-stratified, we replace $(*)$ with $\min\left(\frac{\text{round}\left(\frac{n}{N_{s_I}} N_h\right)}{N_h}, 1\right)$. In the simulations, instead of computing this precisely, we estimate it by randomly sampling s_I such that $h \in s_I$. This avoids the computationally intensive loop over all $\binom{N_I - 1}{n_I - 1}$ such sets. Although these weights are not equal for all individuals, because the distributions of household sizes (from the MVP demographic data) have no extreme outliers, in our simulations the weights are nearly equal.

For the PPS scheme, the inclusion probabilities are:

$\pi_{hi} = E_k\left[P\left(\text{person } i \text{ in household } h \text{ is sampled} \mid \text{household } i \text{ is chosen } k \text{ times}\right)\right]$

$= E_k[1-(1-n_h/N_h)^k]$ since we independently subsample a household as many times as it is drawn.

Since $k \sim \text{Bin}(m_I, p_h)$, by its probability generating function, we obtain

$= 1-(p_h(1-n_h/N_h)+(1-p_h))^{m_I}$

$= 1-\left(1-p_h\frac{n_h}{N_h}\right)^{m_I}$

if $p_h\frac{n_h}{N_h}$ is small, we can approximate this as:

$= m_I p_h \frac{n_h}{N_h}$

In PPS sampling, $p_h \propto x_h$, where x_h is a measure of household size (Särndal et al., 1992, p.97). So the PPS weights are:

$$w_{hi} = \frac{\sum_{h\in U_I} x_h}{m_I x_h}\frac{N_h}{n_h}.$$

If $x_h \propto N_h$, and $n_h \propto 1$, then the design is self-weighted. We take $n_h = c$, a constant, but we cannot choose x_h such that $x_h \propto N_h$ for all modules, since the target age-sex groups differ from module to module. We chose $x_h = N_{h,\text{total}}$, the number of household members under 50 years of age, because it represented a compromise between the different target age-sex groups. Thus, our weights are $w_{hi} \propto \frac{N_h}{N_{h,\text{total}}}$.

6 Comparisons between sampling schemes: variances and design effects

We want to compare the PPS and SRS designs (in either the Bayesian model-based or the design-based paradigms). In general, the two schemes will have slightly different sample sizes, making direct comparisons of variances less relevant. For the household survey module, we fix the sample sizes to be equal, and for the adult, anthropometry, and blood modules, we adjust for the differing sample sizes by computing a design effect, defined below.

The household survey module is administered to the heads of households only, not individual members. Therefore, the time cost of the household module is mostly determined by the number of households surveyed. We set up our simulations such that the number of household heads to be interviewed (i.e. sample size) is the same for the SRS and PPS sampling schemes. We first perform a PPS sampling of households. Then, we use the number of unique sampled households to obtain the number of households to sample for the SRS scheme. We then directly compare the variances in estimating $\overline{Y}$, the finite population mean consumption per person.

For the remaining modules, we compute design effects. To define the design effect (often abbreviated as "deff"), we first introduce the following notation. Let $\widehat{\theta}_\pi = \widehat{\theta}_\pi(I, y_{obs})$ be the estimator of θ (in our case, $\theta = \overline{Y}$) where π is the sampling distribution assumed to have been used in drawing sample S. Let $V_{\pi_1}(\widehat{\theta}_{\pi_2}; y)$ be the sampling variance of an estimator of θ that assumes sampling distribution π_2, and π_1 is the distribution with respect to which we want the variance. Let $\widehat{V}_{\pi_1}(\widehat{\theta}_{\pi_2}; \pi_3; I, y_{obs})$ be an estimator where π_3 is the sampling distribution assumed to have been used in drawing sample S. The population design effect is defined as $= V_p(\widehat{\theta}_p; y)/V_{SRS}(\widehat{\theta}_{SRS}; y)$. The estimated design effect is defined as $= \widehat{V}_p(\widehat{\theta}_p; p; y_{obs})/\widehat{V}_{SRS}(\widehat{\theta}_{SRS}; p; y_{obs})$.

In the design-based setting, we compute design effects assuming sampling with-replacement in both numerator and denominator variances. This is done in the `survey` package by specifying `deff = 'replace'`.

For the model-based simulations, we estimate the numerator of the deff with the posterior variance for $\overline{Y}$ from fitting a model that includes enough design variables such that the data collection mechanism is ignorable with respect to this model. This posterior variance includes an implicit finite population correction, so we compute a denominator variance that also includes such a correction:

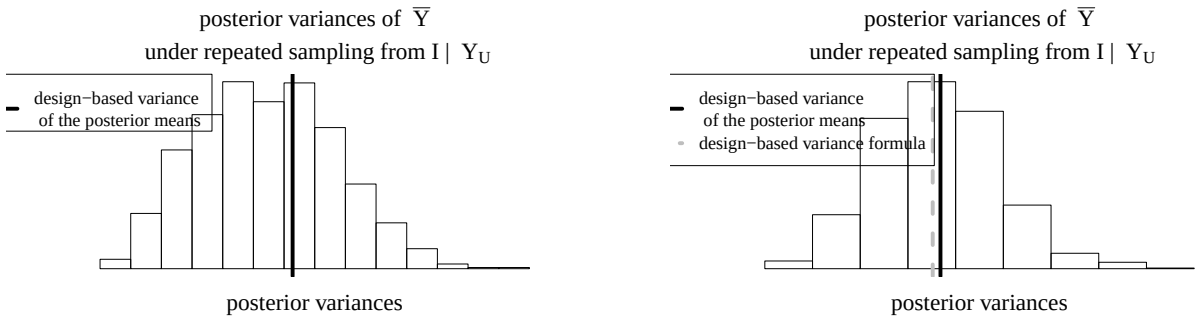
$$\begin{aligned} V_{SRS}(\widehat{\theta}_{SRS}; y) &= V_{SRS}(\overline{y}; y) \\ &= \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \end{aligned} \tag{5}$$

where $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{Y})^2$.

To assess our estimated deff in the model-based setting, we compare the posterior variance $V_{p\text{-ignorable}}(\theta|y_{obs})$ from fitting an ignorable model with respect to a sampling distribution p to the design-based sampling variance of the posterior means, $E_{p\text{-ignorable}}(\theta|y_{obs})$. The latter can be computed by simulation: we sample repeatedly from the full population using distribution p, fit the p-ignorable model, obtain a posterior mean of θ, and compute the variance of these across the samples from p. Fixing one finite population, in Figure 1a we create a histogram of posterior variances from fitting the p-ignorable model to each sample, and indicate with a vertical line the design-based variance of the posterior means, which is computed by simulation. We make the same comparison for p = a simple random sample (and its ignorable model with flat priors and no design variables), and include the closed-form design-based estimate (5) as a vertical line, in addition to the simulation-computed design-based estimate. See Figure 1b. We see that the posterior variances appear unbiased for the design-based variances.

7 Simulation results

Our results are displayed in Appendix B, where we see that neither the SRS nor PPS sampling of households is more efficient (i.e. has a lower design effect) in general.



(a) Sampling distribution $\mathfrak{p}$ is SRS sampling of households followed by an equal-probability systematic sample within households.

(b) Sampling distribution $\mathfrak{p}$ is SRS sampling of people.

Figure 1: Fixing one finite population, we show a histogram of posterior variances from fitting a $\mathfrak{p}$-ignorable model to each sample using sampling distribution $\mathfrak{p}$, and indicate in a vertical line the design-based variance of the posterior means, which is computed by simulation. When $\mathfrak{p}$ is simple random sample of people, we also include the closed-form design-based estimate (5).

We see that for modules with higher target sample sizes, SRS tends to be less efficient. For example, in the under-5 blood ($\mathfrak{n}_{\text{target}} = 300$) and adult ($\mathfrak{n}_{\text{target}} = 400$) modules the SRS scheme is less efficient. One explanation for this observation is the different numbers of people sampled per household in the SRS versus PPS schemes, which has efficiency implications due to the intra-house correlation. In the PPS scheme, the households sampled in the first stage are larger and therefore more likely to include people in the target demographics. In contrast, in the SRS scheme, the sample is often drawn from fewer households, with more people sampled per household. Moreover, the PPS scheme only samples one person per household draw (though this can result in more than one person being sampled per household due to the with-replacement sampling at the first stage).

For modules where the target sample size is low, there are fewer people sampled per household in the SRS sampling scheme, and the intra-house correlation does not substantially impact the design efficiency. Therefore, because SRS has near-equal individual-level probability of sampling (see the design-weights computed above), its design effect in the absence of household clustering should be close to one. In contrast, the PPS scheme does not have near-equal individual-level sampling probabilities because the measure of household size is not proportional to the target demographic (see the design-weights computed above).

The relative efficiency of SRS versus PPS is similar between design-based and model-based simulations. In the few cases where they differ, design-based results show that SRS has higher design effects than PPS, relative to model-based results. In general, our model-based simulations show more variability across simulations than the design-based simulations. Comparing systematic to stratified sampling at the second stage of the SRS schemes, we see few differences except that stratified sampling tends to have higher variance across simulations.

8 Final sampling plan

As described above, the PPS scheme requires a sampling frame that includes household sizes, whereas the SRS scheme only requires a list of households. Given our results, we cannot justify the additional resources required to collect the more detailed household list for the PPS scheme. Therefore, our sampling scheme will begin with an SRS sample of households. For the second stage of sampling for the adult, anthropometry, and blood modules, we prefer the control over sample size achieved by systematic sampling (as opposed to stratified sampling).

The household and nutrition modules follow a different sampling scheme. As mentioned above, the household module is administered to all household heads (or other knowledgeable household members) within the sampled households. The nutrition module consists of a food frequency questionnaire, which takes longer to administer than other modules. We suspect that the within-household correlation is very high for data on food frequency, because household members are likely to eat similar foods. (This intra-house correlation cannot be measured from project data, because the project has always limited this module to one member per household.) For these reasons, we limit the nutrition module to one adult (age 15 to 49 years) per household.

9 Software

For fitting multilevel models we use Stan in `R`, (Stan Development Team, 2013; R Development Core Team, 2014).

Acknowledgements

The authors would like to thank the following people for very valuable feedback and ideas: Jeffrey D. Sachs, Joseph K. Blitzstein, Qixuan Chen, Jennifer Hill, Macartan Humphreys, Michael Clemens, Alberto Abadie, Marc Levy, Linda Pistolesi, Keli Liu, Peng Ding, Natalie Exner, Abhishek Chakrabortty, Rachael Meager, and Natalie Bau.

All mistakes are our own.

A Properties of systematic sampling

Definition 1 (The fractional interval method of systematic sampling)**.** *Consider a population of size* N^* *consisting of people grouped into* n_I *households indexed by* h*, with* N_h *people within-household* h*. Let* $n_{target} < N^*$ *be the desired sample size. Set* $a = \frac{N^*}{n_{target}}$*. Order the households randomly, and randomly order the people in the target group within the households. Let* $k = 1, ..., N^*$ *label the people in this order:*

$$\underbrace{1.....N_1}_{household\ 1}\ \underbrace{(N_1+1).....(N_1+N_2)}_{household\ 2}\ .....\ \underbrace{.....N^*}_{household\ n_I}$$

Draw a random real number ξ *uniformly between 0 and* a*,* $\xi \sim U(0, a)$*, and sample all people with* k *such that*

$$k - 1 < \xi + (j-1)a \leqslant k \text{ for } j = 1, ..., n_{target}.$$

(Särndal et al., 1992, p.77)

Claim 1. *When performing the sampling scheme in Definition 1, the sample size will be* n_{target}*.*

Proof of Claim 1. Since $a \equiv \frac{N^*}{n_{target}}$ and $N^* > n_{target}$, $a > 1$. Since $k - 1 < x \leqslant k \Leftrightarrow \lceil x \rceil = k$, we can write $\lceil \xi + (j-1)a \rceil = k$. The ceiling function is monotone increasing and $\lceil x + 1 \rceil = \lceil x \rceil + 1$, so each time j increases by 1, we get a different value of k. Now we must show that the k's stay in the set $\{1, ..., N^*\}$, i.e. those from which we are sampling. The first k is such that $k - 1 < \xi \leqslant k$, where $\xi \in (0, a)$. Since $\xi > 0$, we must have $k \geqslant 1$. The last k is such that $k - 1 < \xi + (n_{target} - 1)a \leqslant k$, and we know $\xi + (n_{target} - 1)a \leqslant \xi + N^* - a < N^*$ because $\xi < a$. Then $k \leqslant N^*$. Thus, since each j maps to a unique k, we've proven we get a sample size of exactly n_{target}. □

Claim 2. *When performing the sampling scheme in Definition 1, the sample size within each household* h *is always the ceiling or the floor of the expected sample size in household* h *under simple random sampling:* $\frac{N_h}{a}$*.*

We first prove the following lemma which is used to prove the above claim:

Lemma 1. *Consider the set* $A(x) \equiv \{j \in \mathbb{Z}^+ | \xi + (j-1)a \leqslant x\}$*. The maximum of* $A(x)$ *is* $\frac{x}{a}$ *if* $\frac{x}{a} \in \mathbb{Z}$*,* $\lfloor \frac{x}{a} \rfloor$ *if* $\xi > da$*, or* $\lceil \frac{x}{a} \rceil$ *if* $\xi \leqslant da$*, where* $d \equiv \frac{x}{a} - \lfloor \frac{x}{a} \rfloor$*, the "decimal part."*

Proof of Lemma 1. If $\frac{x}{a} \in \mathbb{Z}$, let $j = \frac{x}{a}$, and we see that $\xi + \left(\frac{x}{a} - 1\right) a = \xi + x - a \leqslant x$ because $\xi < a$, so $\frac{x}{a} \in A(x)$. Increasing j by 1 increases the lefthand side of the inequality by a, and since $\xi > 0$, we see this $\frac{x}{a} + 1 \notin A(x)$. Therefore, $\frac{x}{a}$ is the maximum.

If $\frac{x}{a} \notin \mathbb{Z}$, let $d \equiv \frac{x}{a} - \lfloor \frac{x}{a} \rfloor$, the "decimal part." We see that $\xi + \left(\lfloor \frac{x}{a} \rfloor - 1\right) a \underset{+da}{<} \xi + \left(\frac{x}{a} - 1\right) a = \xi + x - a \underset{+a-\xi}{\leqslant} x$, so $\lfloor \frac{x}{a} \rfloor \in A(x)$. Increasing j by 1 (to $\lceil \frac{x}{a} \rceil$) increases the leftmost side of the inequality by a. If $a \leqslant da + (a - \xi)$, i.e. if $\xi \leqslant da$, then $\lceil \frac{x}{a} \rceil \in A(x)$. □

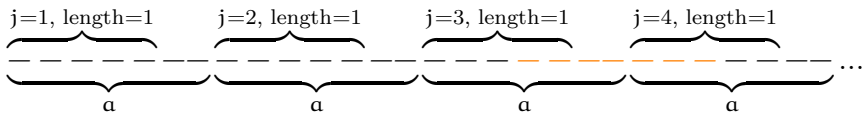
Proof of Claim 2. Consider household h, of size N_h. Let k^* be the last person before household h in the ordering used by the systematic sampling. Then $k^*+1, ..., k^*+N_h$ are the indices for all members of household h. In order to get the number sampled in household h, we consider the maximum of set $A(k^*+N_h)$ (the number of people sampled up through household h) and subtract from it the maximum of set $A(k^*)$ (the number of people sampled before household h). This gives, by Lemma 1, $\frac{k^*+N_h}{a} - \frac{k^*}{a} = \frac{N_h}{a}$. □

Claim 3. *The sampling scheme in definition 1 is self-weighted.*

Proof.

$$
\begin{aligned}
P(\text{person } k \text{ is sampled}) &= P\left(\exists j \in \{1, ..., n_{target}\} \text{ s.t. } k-1 < \xi + (j-1)a \leqslant k\right) \\
&= P\left(\cup_{j=1}^{n_{target}} \left\{k-1-(j-1)a < \xi \leqslant k-(j-1)a\right\}\right) \\
&\text{each interval is length } 1 = [k-(j-1)a] - [k-1-(j-1)a] \\
&\text{space between intervals j and j+1 is } a-1 = [k-1-j*a] - [k-(j-1)a] \\
&\text{so by the picture below, we see that } (0,a) \text{ has overlaps with the intervals} \\
&\text{of length totaling 1. So since } \xi \sim U(0,a), \\
&= \frac{1}{a}
\end{aligned}
$$

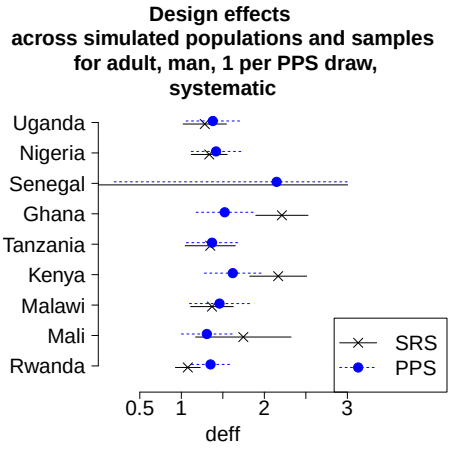
See below for a visual, in orange is the interval $(0, a)$, which can overlap at most 2 intervals of length 1 (shown as over-braces, with overlaps totaling a length 1:



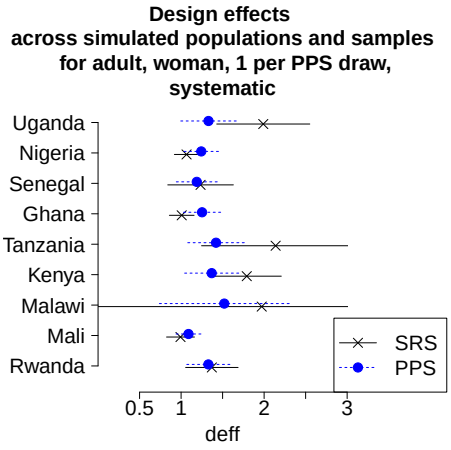
□

B Survey Sampling Simulation Results

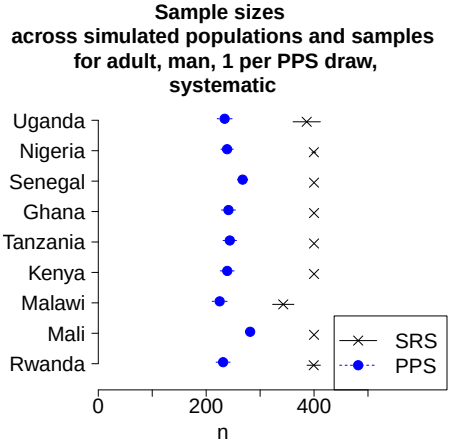
B.1 Design-based results



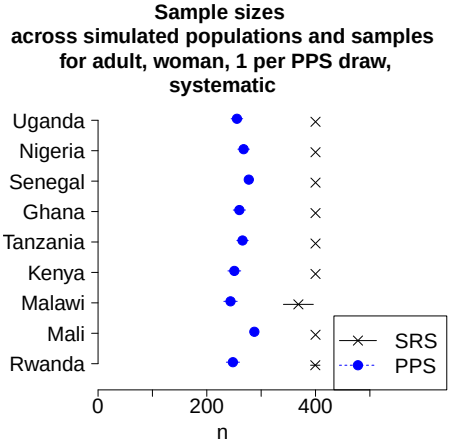
(a)



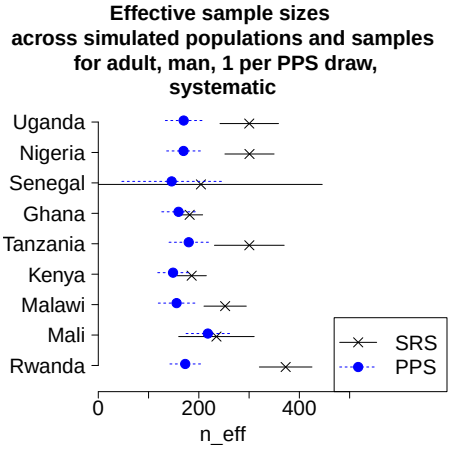
(b)



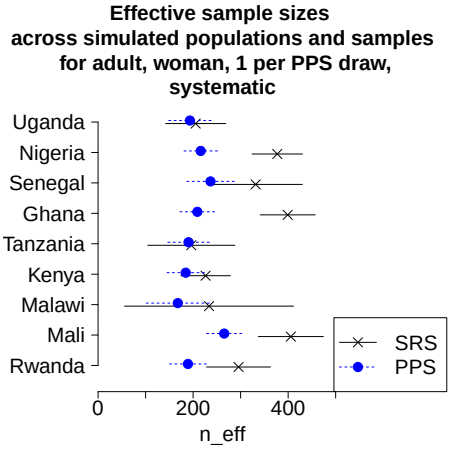
(c)



(d)

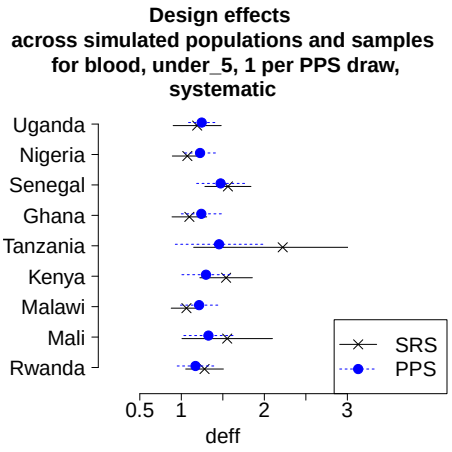


(e)

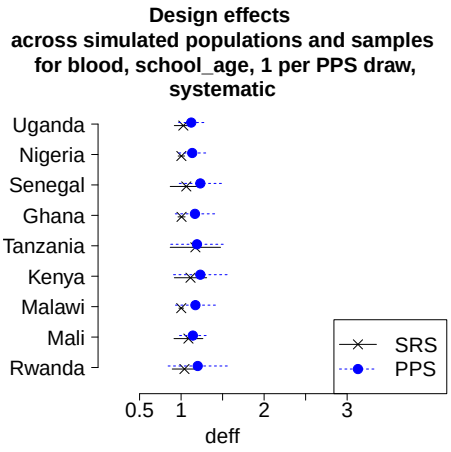


(f)

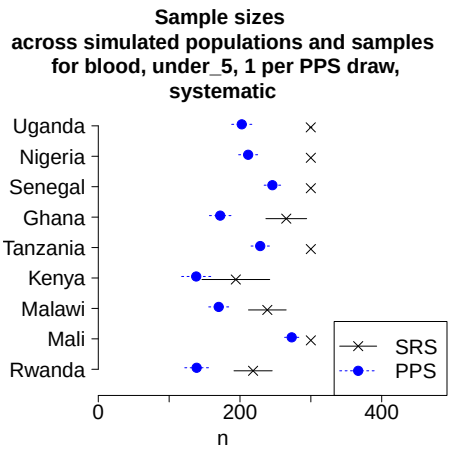
Figure 2: Design-based adult module results



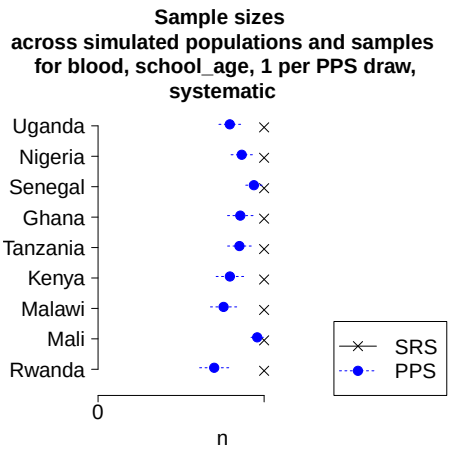
(a)



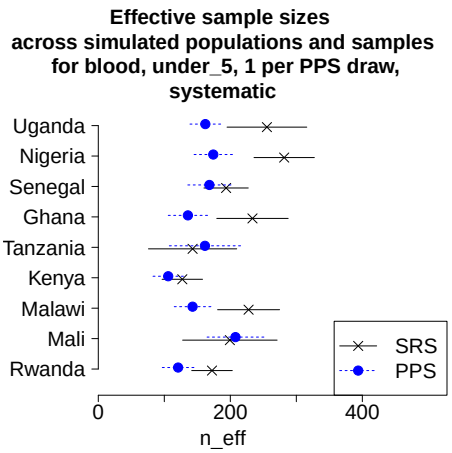
(b)



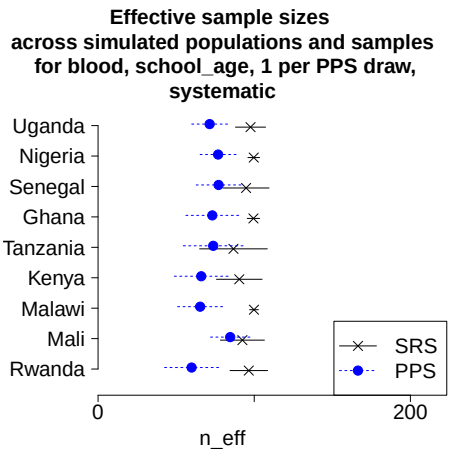
(c)



(d)

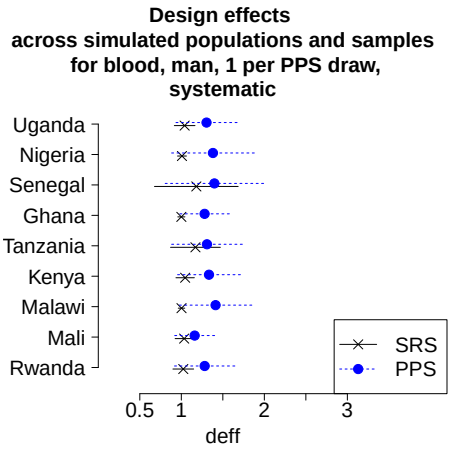


(e)

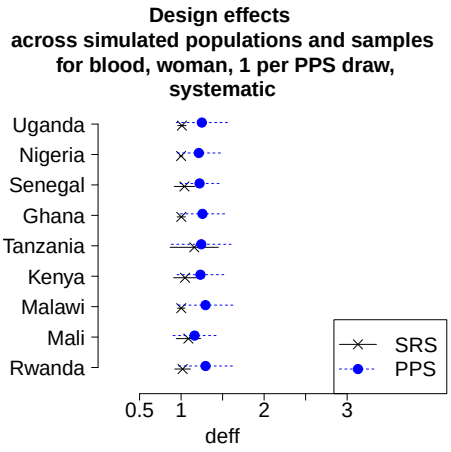


(f)

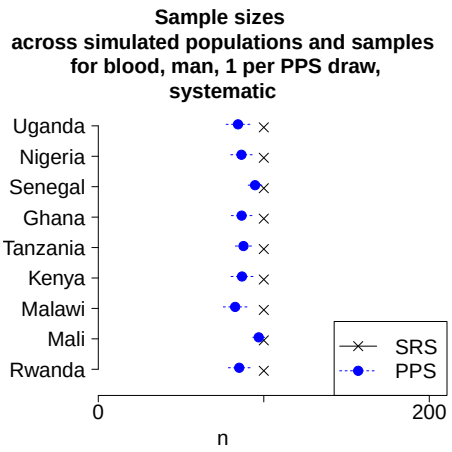
Figure 3: Design-based blood (malaria and anemia) module results: under 5 and school-age children



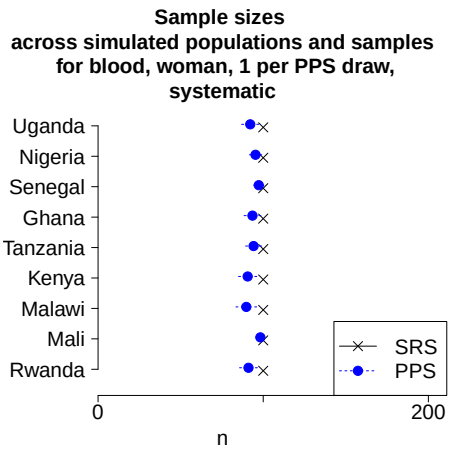
(a)



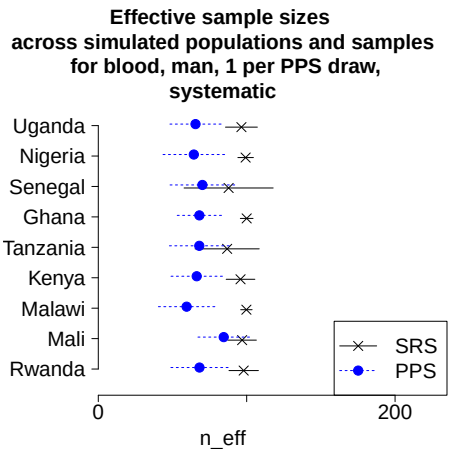
(b)



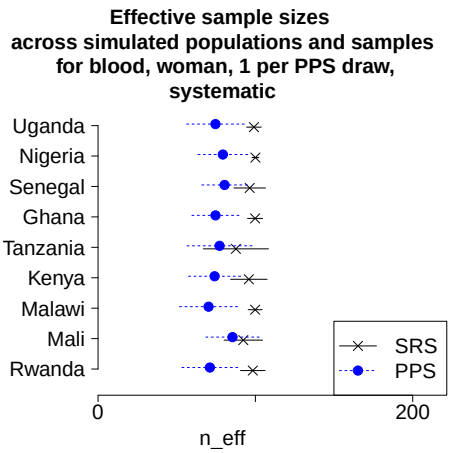
(c)



(d)

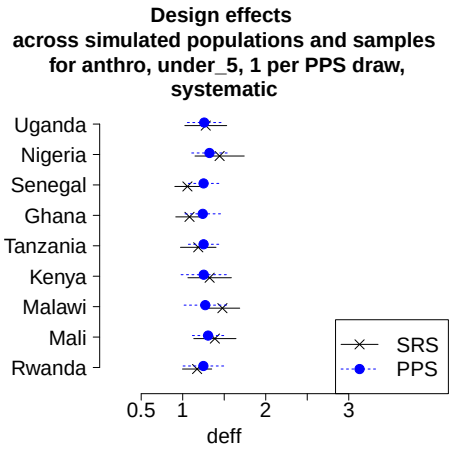


(e)

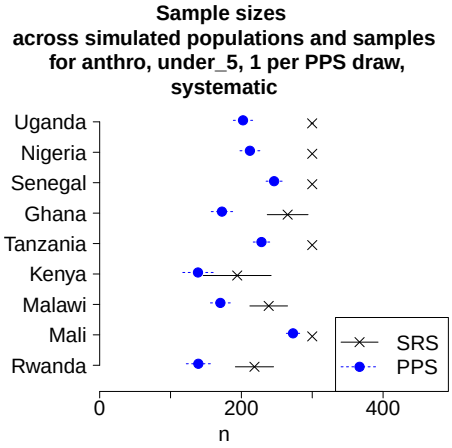


(f)

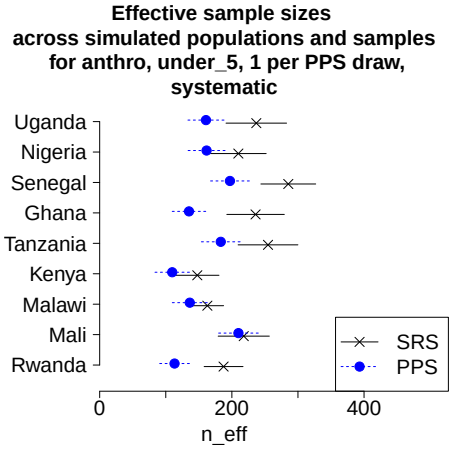
Figure 4: Design-based blood (malaria and anemia) module results: men and women



(a)

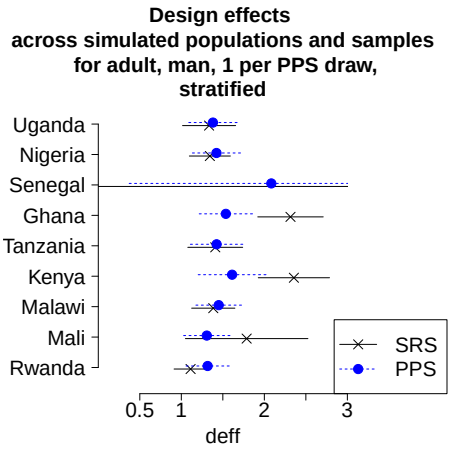


(b)

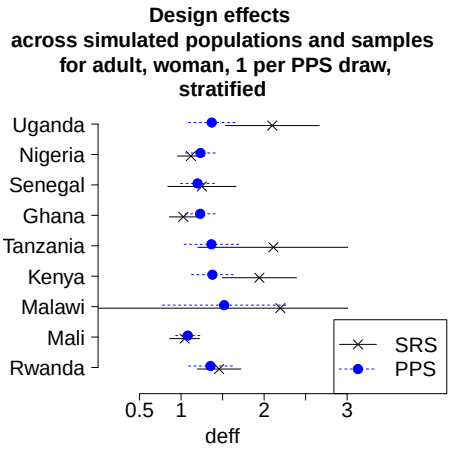


(c)

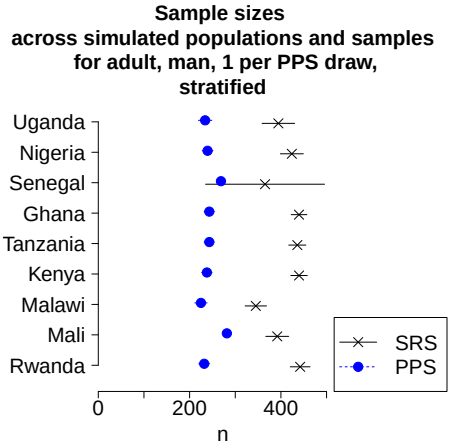
Figure 5: Design-based anthro module results



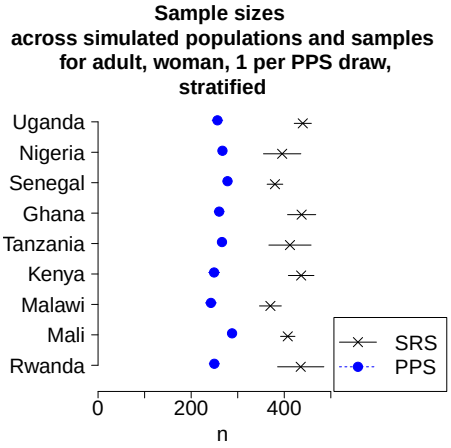
(a)



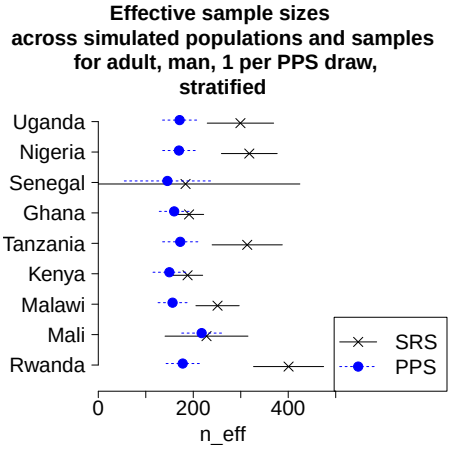
(b)



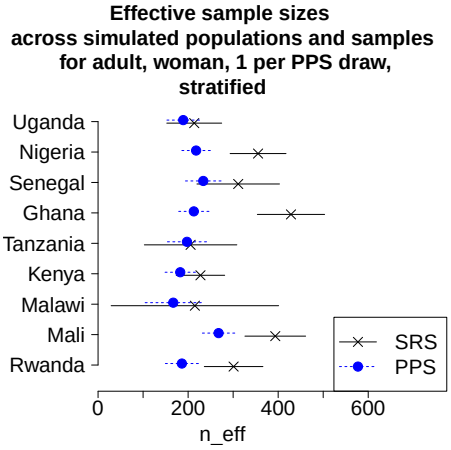
(c)



(d)

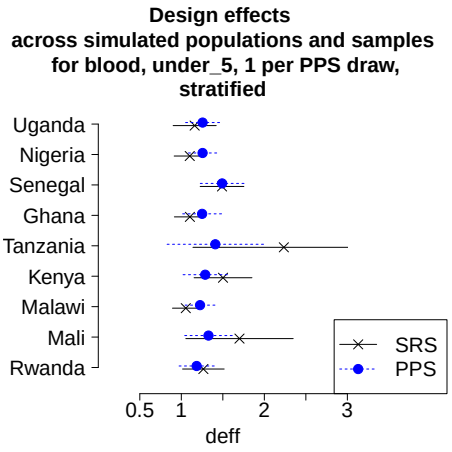


(e)

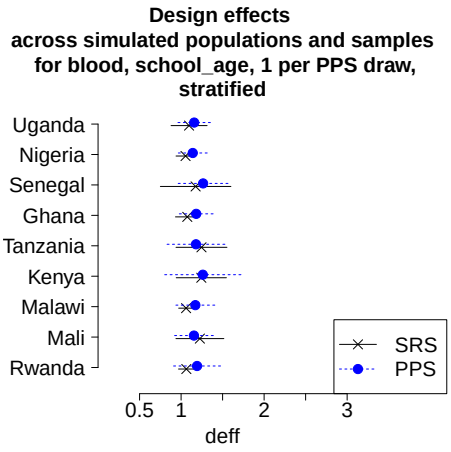


(f)

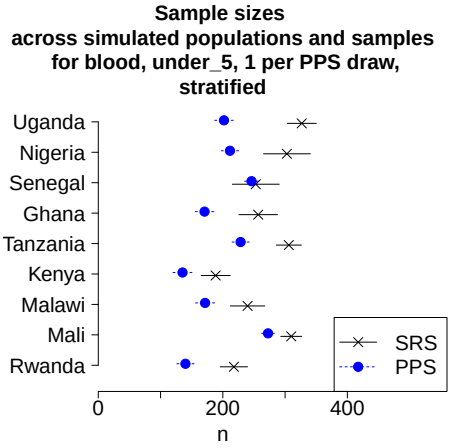
Figure 6: Design-based adult module results



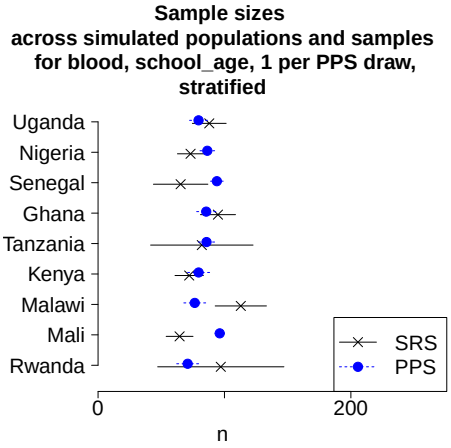
(a)



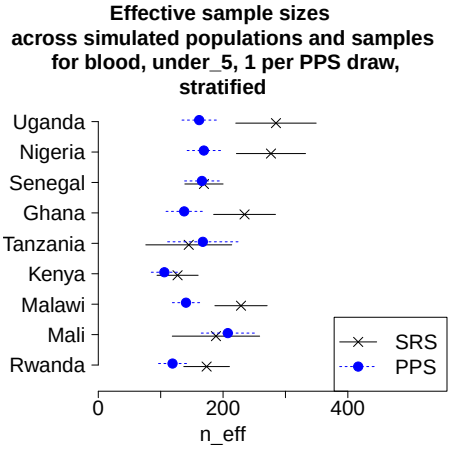
(b)



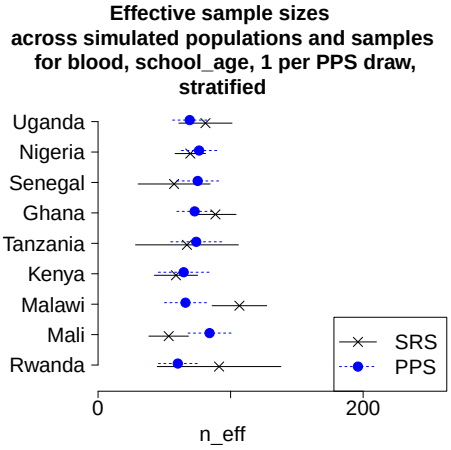
(c)



(d)

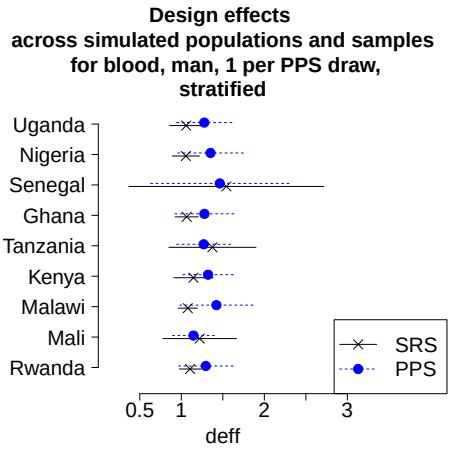


(e)

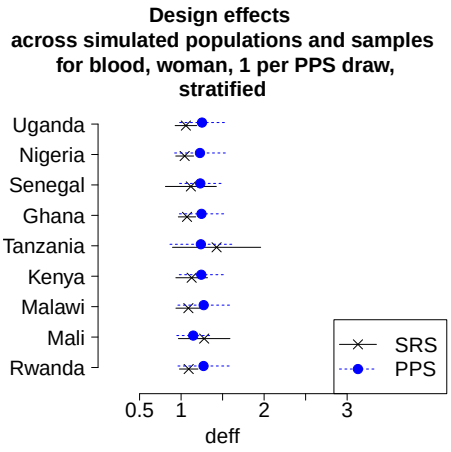


(f)

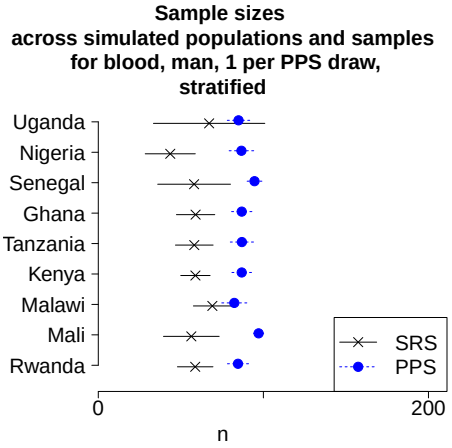
Figure 7: Design-based blood (malaria and anemia) module results: under 5 and school-age children



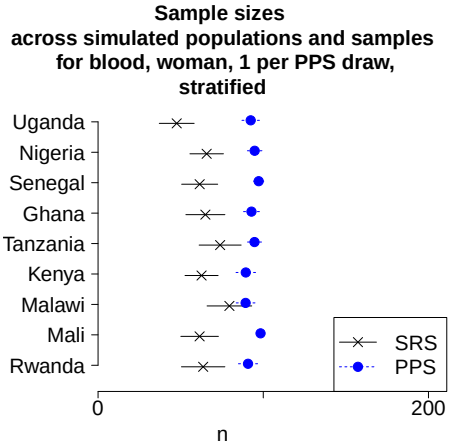
(a)



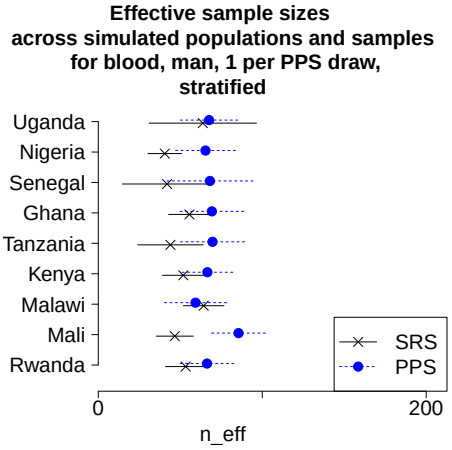
(b)



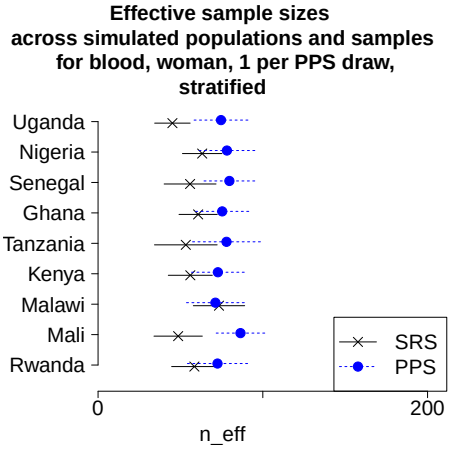
(c)



(d)

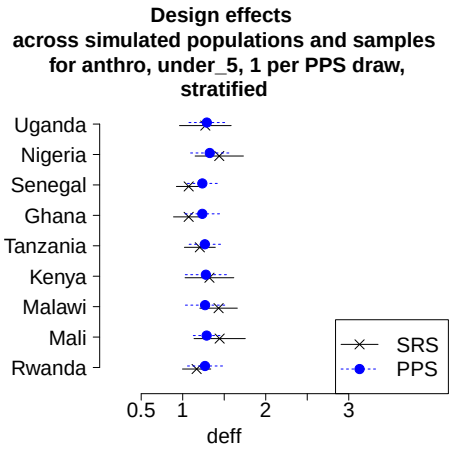


(e)

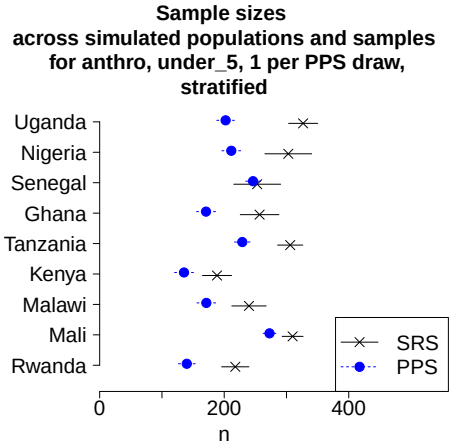


(f)

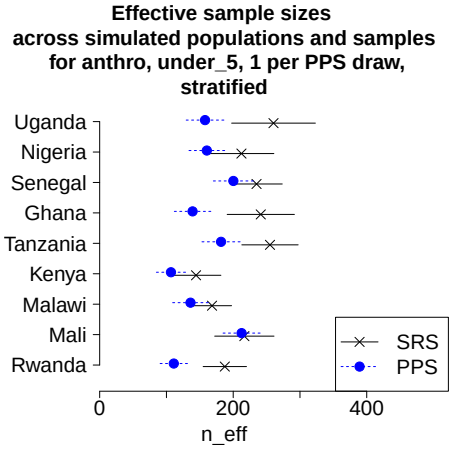
Figure 8: Design-based blood (malaria and anemia) module results: men and women



(a)



(b)



(c)

Figure 9: Design-based anthro module results

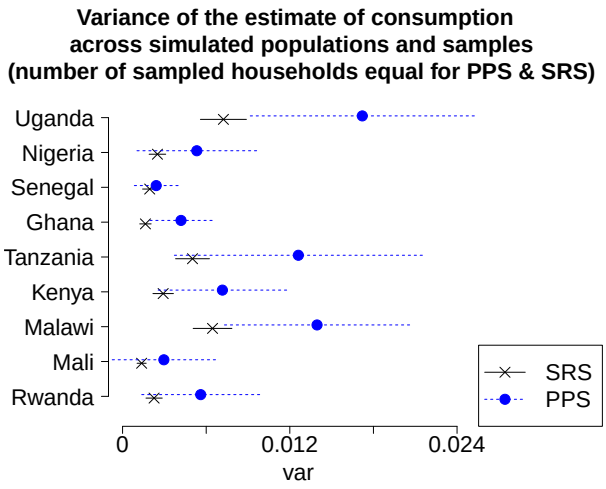
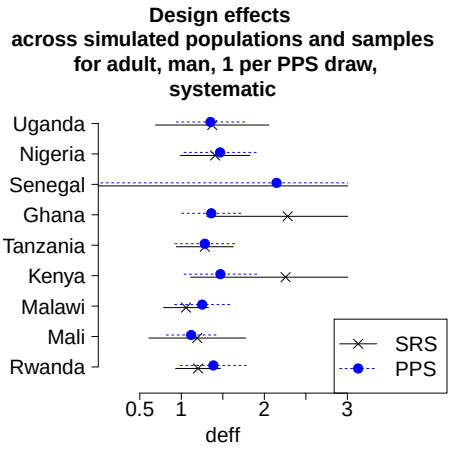
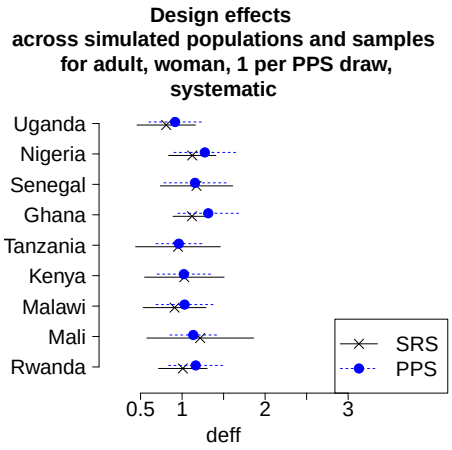


Figure 10: Design-based consumption module results

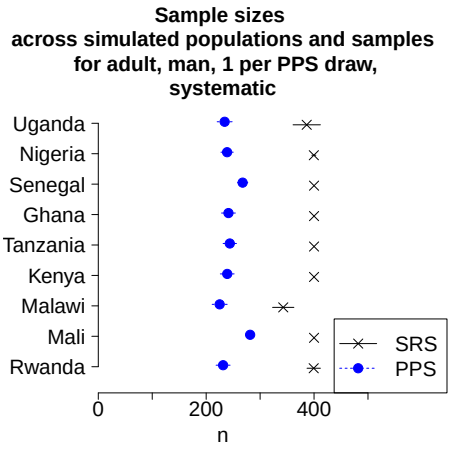
B.2 Model-based Results



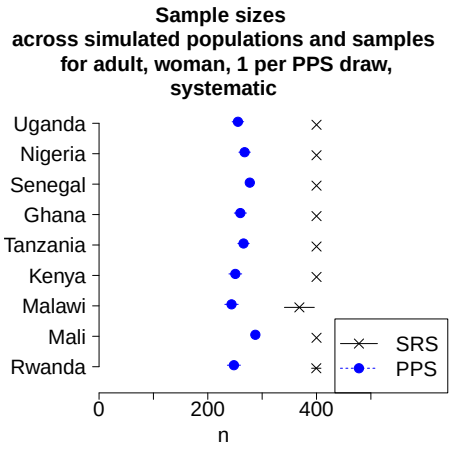
(a)



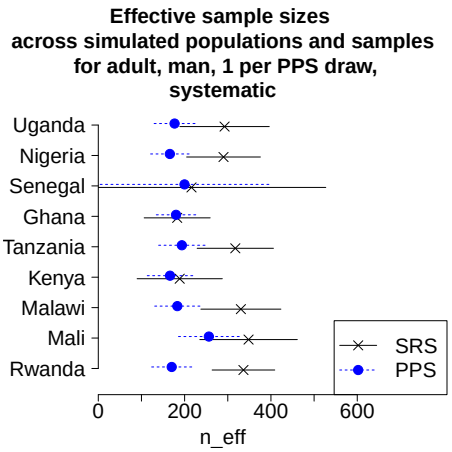
(b)



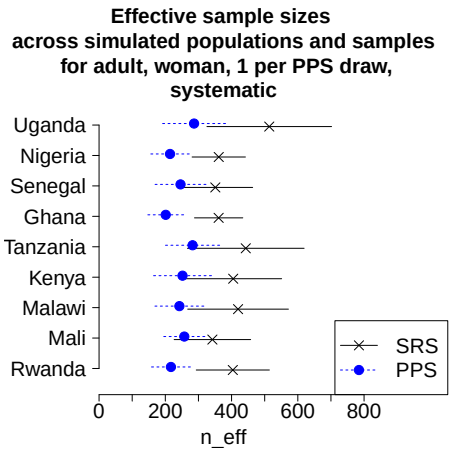
(c)



(d)

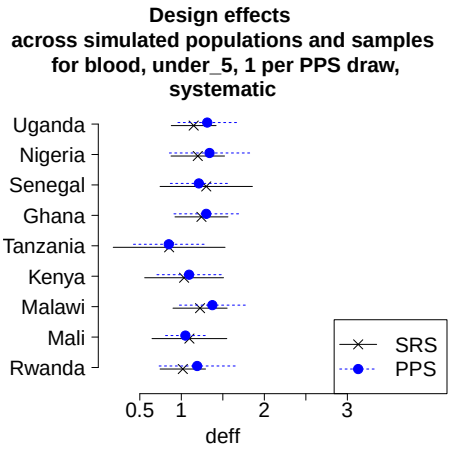


(e)

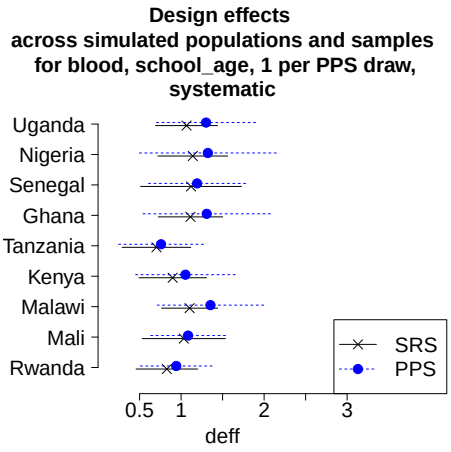


(f)

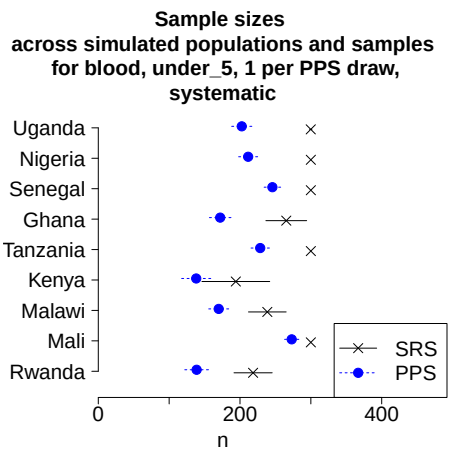
Figure 11: Model-based adult module results



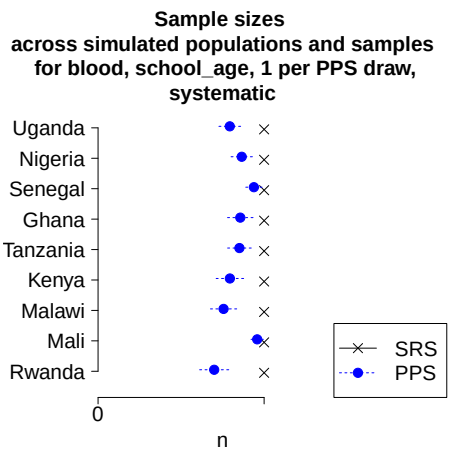
(a)



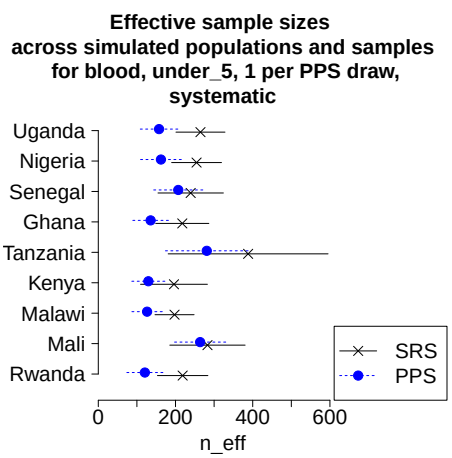
(b)



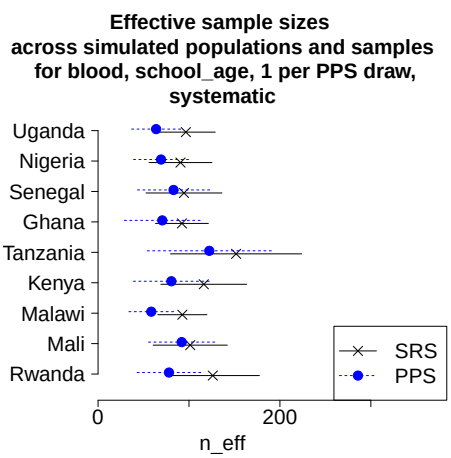
(c)



(d)

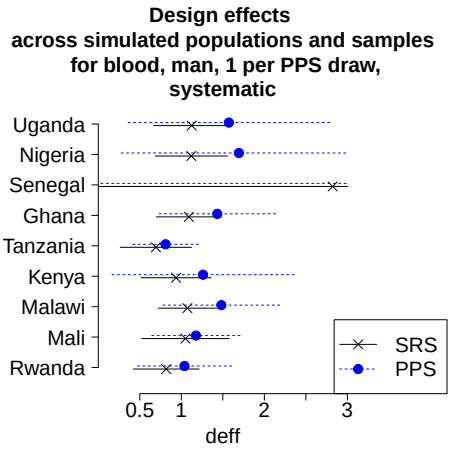


(e)

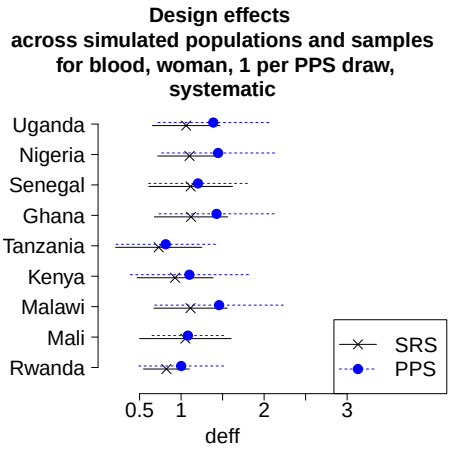


(f)

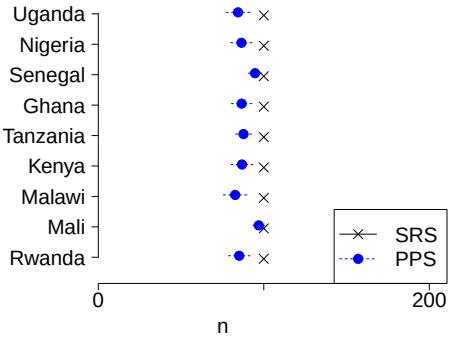
Figure 12: Model-based blood (malaria and anemia) module results: under 5 and school-age children



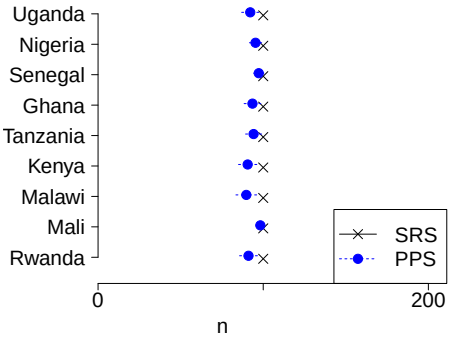
(a)



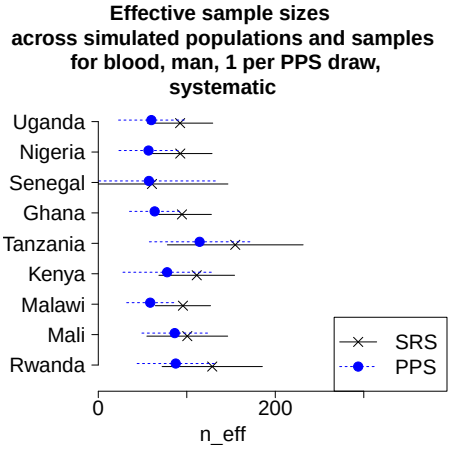
(b)



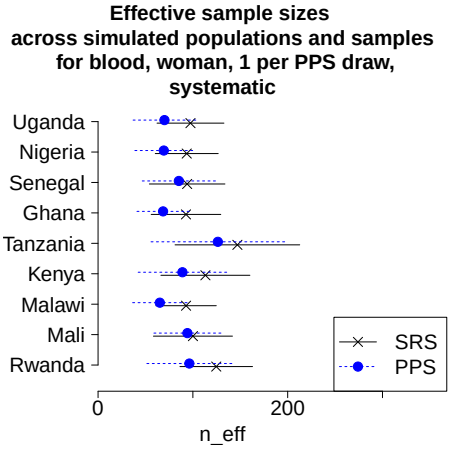
(c)



(d)

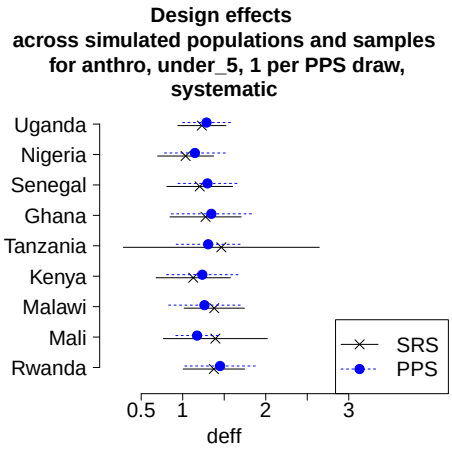


(e)

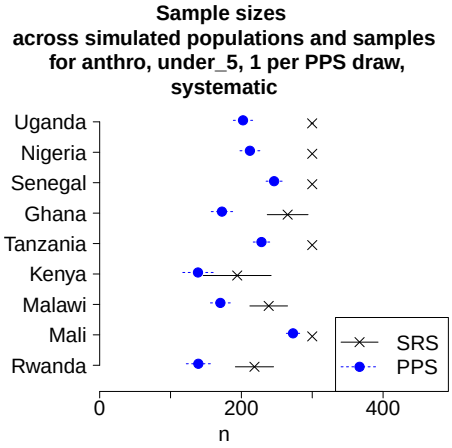


(f)

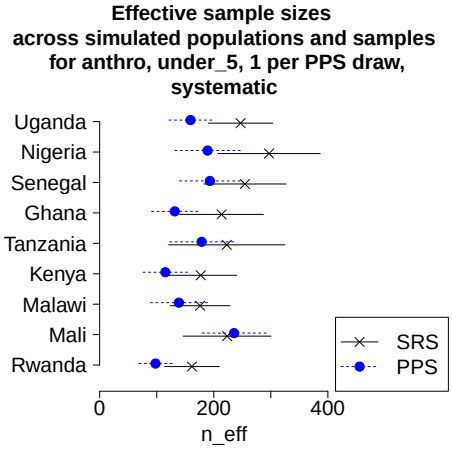
Figure 13: Model-based blood (malaria and anemia) module results: men and women



(a)

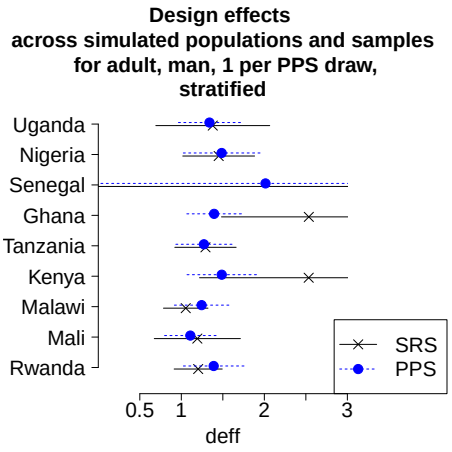


(b)

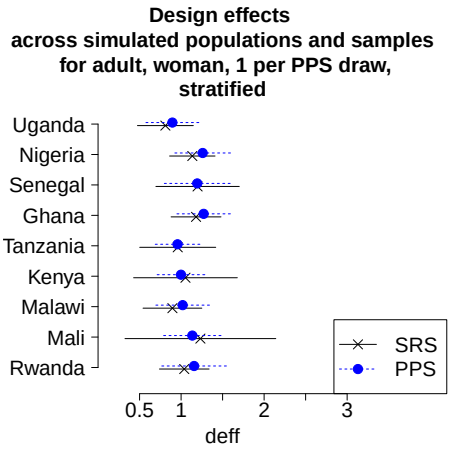


(c)

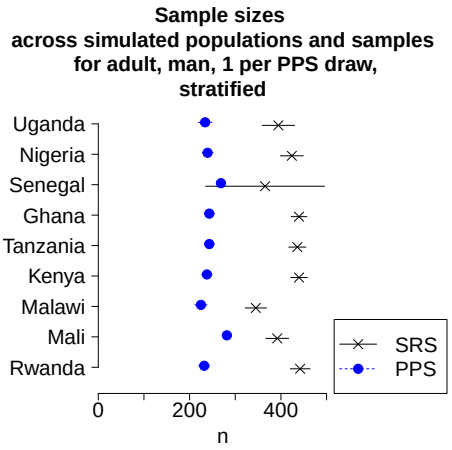
Figure 14: Model-based anthropometry module results



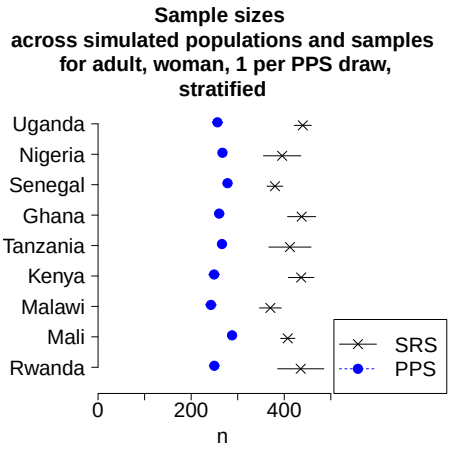
(a)



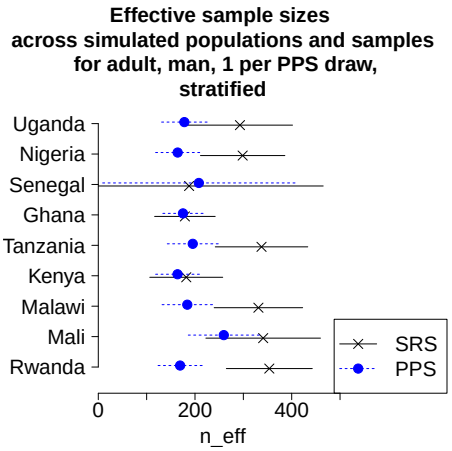
(b)



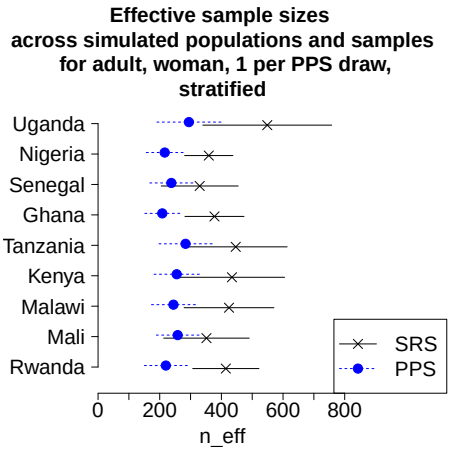
(c)



(d)

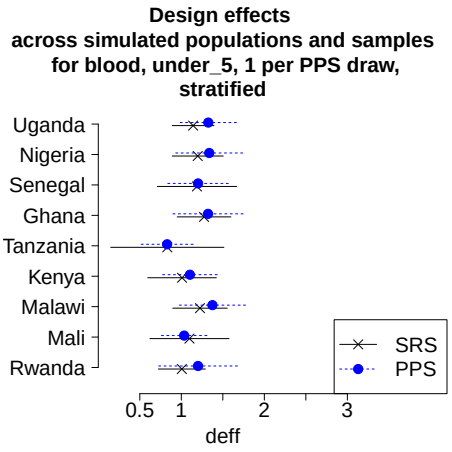


(e)

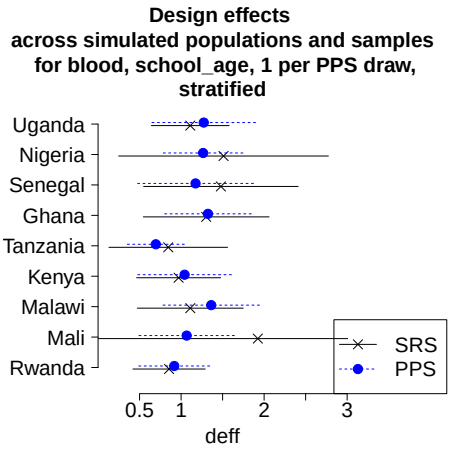


(f)

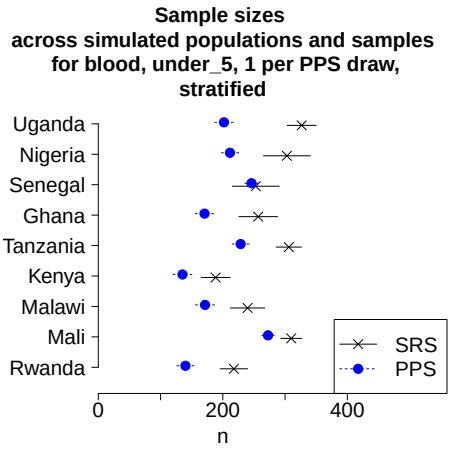
Figure 15: Model-based adult module results



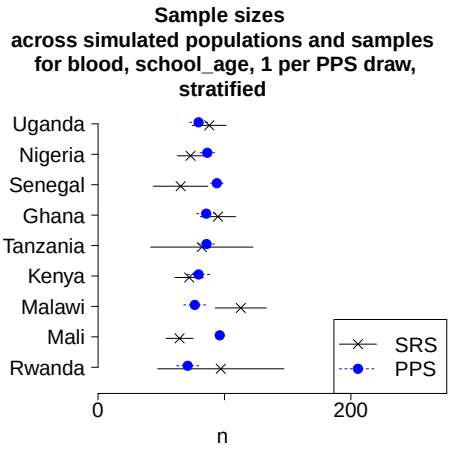
(a)



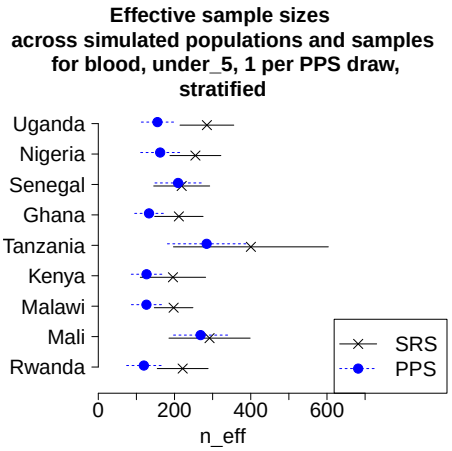
(b)



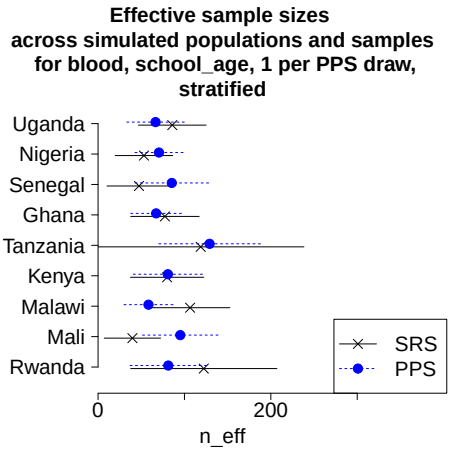
(c)



(d)

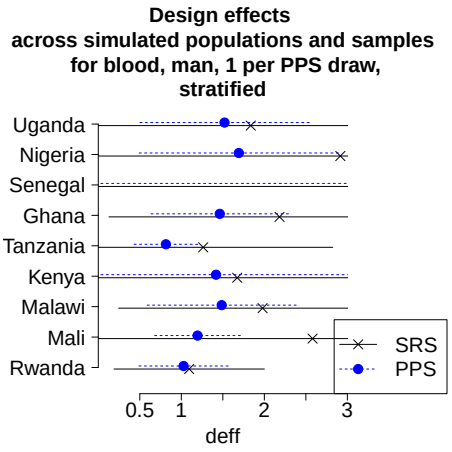


(e)

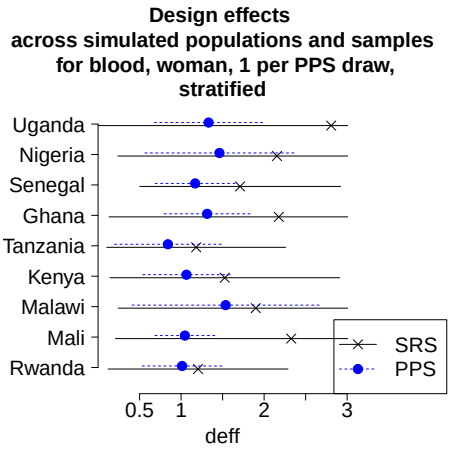


(f)

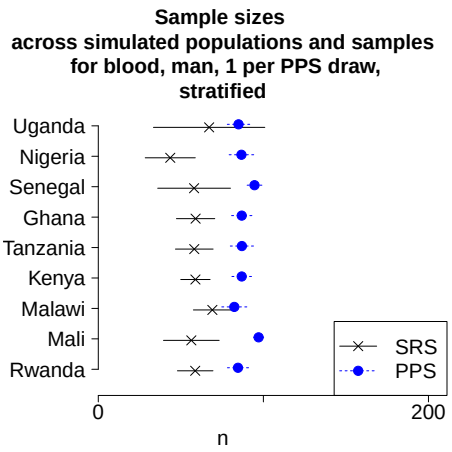
Figure 16: Model-based blood (malaria and anemia) module results: under 5 and school-age children



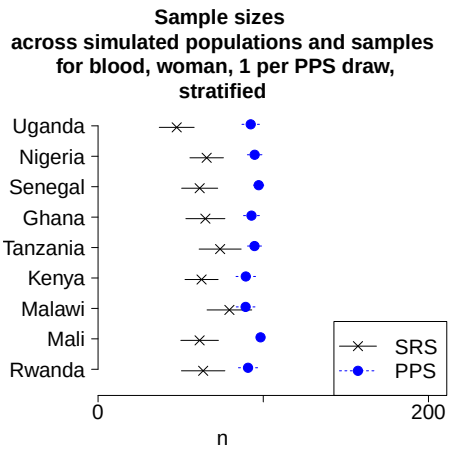
(a)



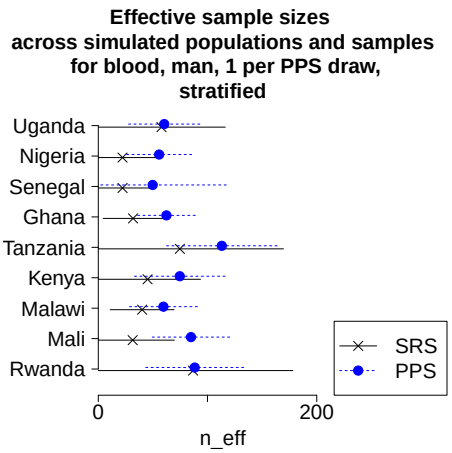
(b)



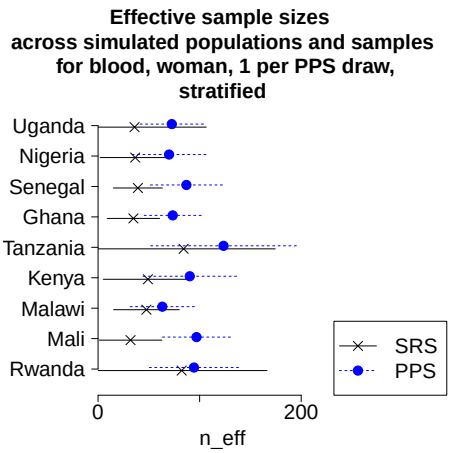
(c)



(d)

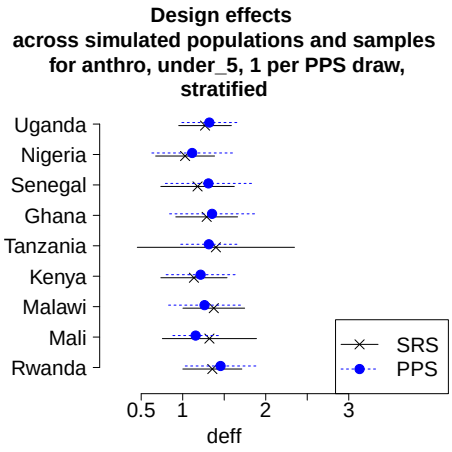


(e)

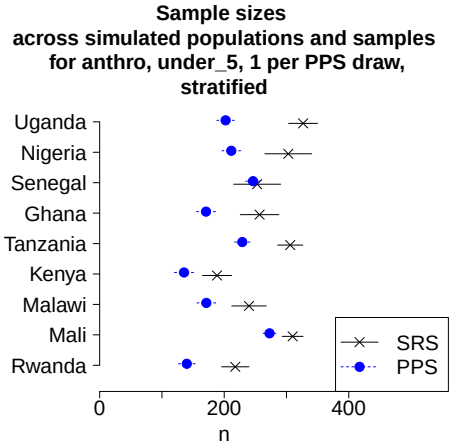


(f)

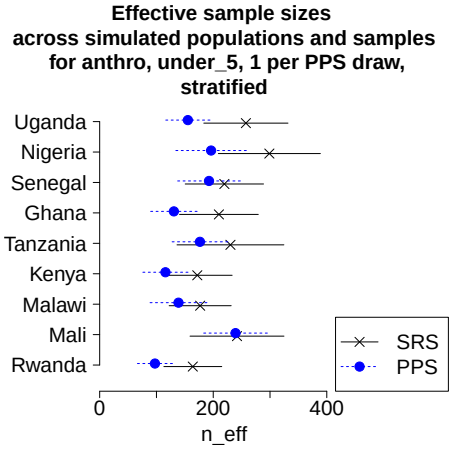
Figure 17: Model-based blood (malaria and anemia) module results: men and women



(a)



(b)



(c)

Figure 18: Model-based anthropometry module results

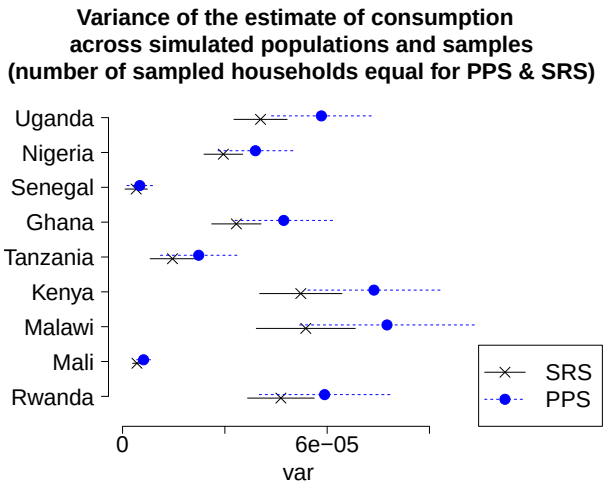


Figure 19: Model-based consumption module results

References

A Gelman, J B Carlin, H S Stern, D B Dunson, A Vehtari, and D B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC texts in statistical science, third edition, 2014.

L Kish. Weighting for unequal p_i. *Journal of Official Statistics*, 8(2):183–200, 1992.

S L Lohr. *Sampling: Design and Analysis*. Cengage Learning, 2 edition, 2010.

T Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8), April 2004.

S Mitchell, A Gelman, R Ross, U Kim Huynh, L McClellan, M Harris, S Bari, J Chen, S Ohemeng-Dapaah, P Namakula, S E Sachs C Palm, and J D Sachs. The Millennium Villages Project: A protocol for the final evaluation. *Protocol in The Lancet*, 2015a. URL `http://www.thelancet.com/doi/story/10.1016/html.2015.07.03.2167`.

S Mitchell, R Ross, S Makela, E A Stuart, A Feller, A M Zaslavsky, and A Gelman. Causal inference with small samples and incomplete baseline for the Millennium Villages Project. Working Paper, 2015b. URL `http://www.stat.columbia.edu/~gelman/research/unpublished/MVP_paper_technical_JRSSA_short.pdf`.

R Development Core Team. The R project for statistical computing, February 2014. URL `http://www.r-project.org/`.

J D Sachs and J W McArthur. The millennium project: a plan for meeting the millennium development goals. *The Lancet*, 365(9456):347–353, January 2005.

P Sanchez, C Palm, J D Sachs, G Denning, R Flor, R Harawa, B Jama, T Kiflemariam, B Konecky, R Kozar, E Lelerai, A Malik, P Mutuo, A Niang, H Okoth, F Place, S E Sachs, A Said, D Siriri, A Teklehaimanot, K Wang, J Wangila, and C Zamba. The african millennium villages. *Proceedings of the National Academy of Sciences*, 104(43):6775–80, 2007.

C E Särndal, B Swensson, and J Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

Stan Development Team. Stan: A c++ library for probability and sampling, version 1.3, February 2013. URL `http://mc-stan.org/`.